BUSINESS ANALYTICS

BY TANUJIT CHAKRABORTY
(RESEARCH SCHOLAR, ISI KOLKATA)
Mail : tanujitisi@gmail.com

BUSINESS ANALYTICS — TANUJIT CHAKRABORY.
RS, ISI KOLKATA

Statistical/ Machine Learning (is essentially non-parametric) technique for analysis of large data.

In regression, we don't find $y = f(x)$; rather we do $E(y|x) = f(x)$

[so, the x's are not random variables]

Technical → Statistical → Managerial

[Read: Categorical Data Analysis by Agresti]

Dependency Analysis :- When we try to predict 'y' for given 'x'.

In descriptive analysis, there is always a dependency analysis running in the background.

Descriptive Analysis :- Aims at establishing relationships. Essentially we are attempting to get ideas about $E(y|x)$ or $P(Y|X)$ for various subsets of X using raw data.

Descriptive analysis is the starting point of non-parametric analysis.

Analytics :- Two major types : — Supervised Analytics.
— Unsupervised Analytics.

Supervised analytics typically has a response and explanatory structures. (eg. Regression analysis)

Unsupervised analytics has no response variables. (eg: Segmentation (Dividing a data into parts such that the no. of parts is not known in advance)

Unsupervised :

eg :
- Scale development (such as finding out intelligence).
- Problem of grouping
  [eg: Medical fraud - where doctors and patients are involved in claims that are fraud. This may be difficult to find out and it requires grouping.]

[Where as in ATM frauds, there is immediate complaint from the card holders, so requires grouping]

Analytics Problems:- Typical types of analytics problems:

→ Value estimation (where you want to estimate the value of a random variable on the basis of values of other variables, eg: Regression and forecasting).

→ Problem of classification [Response variable is categorical] Difference between segmentation and classification is that in classification the no. of classes are known in advance.

→ Grouping and Segmentation (eg: Cluster Analysis) [No. of groups and segments are known in advance]

→ scale Development [Principal Component & Factor Analysis]

→ Scenario Analysis (simulations)

Explanatory Vs Predictive Analytics (Supervised Learning) :-

In explanatory analytics, we try to estimate the impact of the explanatory variables on the response.

In predictive analytics, we try to estimate accurately the value of the response variable in a given situation,

Parametric and Non-parametric methods:-

In parametric method, a model form (and possibly some distn) are assumed. In these models, interpretation is generally easier. However, these models are not flexible (the form is fixed) and hence prediction may not be good.

Non-parametric models assume that the data distribution can't be defined in terms of a finite set of parameters

Books:-

- Introduction to Statistical Learning by Tibshirani.
- Elements of Statistical Learning by Tibshirani.

Regression Analysis :-

Steps : 1. Variable Selection
2. Fitting the model and Validation
3. Interpret and use the model.

[Download : Boston Housing Data and analyse]

1. Variable Selection : — Choose certain variables out of a large list.
— If possible, decide about the form.

(a) Access whether X_i (one at a time) and Y are related.
[Correlation is not defined for Categorical Variables]

(b) Visual Representations : — Scatter plot
— Dot plot [Where X_i are categorical, almost same as Scatter plot]
— Mean functions
— Stratified box plot [For categorical variables]

Relationship between X and Y :

(c) Using Contingency Tables :

Y

X	1	2	3	......	c	Marginals
						N_1
1						N_2
2						⋮
⋮						
n						
	N_1	N_2	....			N

$P(Y=1 \mid X=1)$
$P(Y=1 \mid X=2)$
} The ratio of this two is called odds ratio.

If R^2 value is large, it means that X is significant.

Output of a model : R^2 Adj R^2 F-test
Estimated coeff. SE t-value p value

1. Check whether the basic assumptions are satisfied or not.
[Residual plots]

2. Check the existence of outliers / influential observation.

Scanned by CamScanner

'Basic assumptions' to be checked:

Suppose $Y_i = \sum \beta_i X_i + \epsilon_i$

Check i) $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

ii) $E(\epsilon_i) = 0 \;\forall\; i = 1, 2, \ldots$

iii) $cov(\epsilon_i, \epsilon_j) = 0$, i.e., ϵ_i and ϵ_j are uncorrelated for all i and j.

[Read: Exploratory Data Analysis by John Tukey]

Read and apply: — Stratified Boxplot
— Stem and leaf plot
— Matrix Plot
— Mean function plot.

Note: The stratified boxplots may show the following patterns:

— The mean/median of the response may not change as the explanatory variables change.

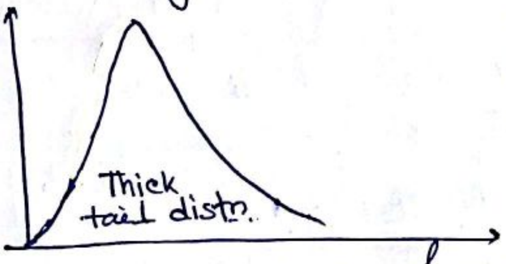
— Some high/low percentile of the response may change as the explanatory variables.

In such a case the mode to be fitted: Quantile Regression Model.

Preliminary Analysis:-

1. Univariate Analysis — Histogram, boxplot, Understand the levels of skewness and kurtosis, identify obvious groups if it exists (bimodal/multimodal distn.s).
Consider transformations when response is highly skewed. (J-shaped)
(Do log transformation, Box-cox transformation)

Cauchy Distribution:



Here the prob. of getting an observation at the extreme is not as low as normal. The kurtosis is very high and tails converge much slowly than a normal distribution.

The variability of mean even with a large sample will be large. So, there exists no expectation, } $\Rightarrow$ In such cases regression is not valid.

2. Relationship Analysis :-

(a) Construct scatter plots, mean functions plot, matrix plots, stratified box plots. Estimate and report correlations. Identify variables that may impact the response. Theorize the form.

(b) Construct two-way tables linking the response and explanatory variables.

— Odds ratio, Relative risk, phi-coefficients.

Note:- 1. If the conditional prob. of Y doesn't depend on X, then X and Y are independent.

2. Relative risk is valid only for prospective samples.

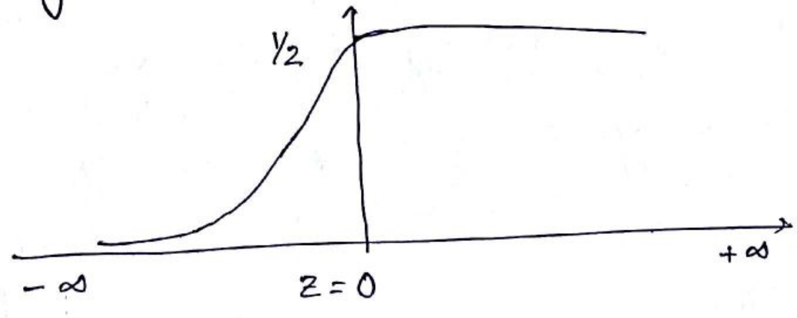
(c) Compute different odds ratio, relative risks, χ^2 values and phi-coefficients to assess impact of X on Y.

[The occurance probability of an avoidable event is called risk.]

Logistic Regression:- (Binary)

Essentially used for classification and risk analysis.
Classification means when response is categorical.
Explanatory → Numeric and Categorical.

Sigmoid function:



Logistic function: $f(z) = \dfrac{1}{1+e^{-z}}$; $-\infty < z < \infty$.

z = Exposure variable. So as z changes, the risk changes.
[eg: Blindly crossing a road. If it is a highway, the exposure is more there than if it's a village road.]

z = Typically a linear combination of the explanatory variables. This variable attemps to quantify risk (Risk event, expenses, probability of risk event).

Bayes Optimality Criteria:- If we have a response variable y which can take values $v_1, v_2, \dots;$

$$P\left(v_j \mid X_1 = x_1, \dots\dots, X_p = x_p\right), \quad j = 1, 2, \dots, K \text{ (classes)}$$

Then the classification is best if we put it in the class where probability is maximum.

eg:

$$Y = \begin{cases} 0 & \text{if transaction is genuine} \\ 1 & \text{if fraudulant} \end{cases}$$

$$\begin{cases} P(Y=1 \mid \underset{\sim}{X}) = p_1 \\ P(Y=0 \mid \underset{\sim}{X}) = p_0 \end{cases}$$

$$Z = \beta_0 + \beta_1 X_1 + \dots\dots + \beta_p X_p$$

$$\text{Odds of } Y=1 \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p = \frac{P(Y=1 \mid \underset{\sim}{X})}{1 - P(Y=1 \mid \underset{\sim}{X})}$$

$$\therefore \ln\left(\text{odds}\left(Y=1 \mid \underset{\sim}{X}\right)\right) = \beta_0 + \sum \beta_i X_i \qquad \left[P(Y=1 \mid \underset{\sim}{X}) = \frac{1}{1 + e^{-(\beta_0 + \sum \beta_i x_i)}} \right]$$

Logistic regression model:-

$$\ln\left(\text{odds}\left(Y=1 \mid \underset{\sim}{X}\right)\right) = \beta_0 + \sum \beta_i X_i$$

Likelihood function:- $L = P(Y=y_1)\, P(Y=y_2) \dots\dots P(Y=y_n)$.

Data:

SL. No	Y	X_1	X_2	.....	X_p
1	y_1	–	–		–
2	y_2	–			
⋮	⋮	⋮	⋮		⋮

Assumption:- Odds ratio remains contant.

[So this assumption is dangerous; we are assuming the risk to be contant in the entire range]

Receiver Operating Characteristic (ROC) Curve :-

This is a standard technique for summarizing classifier performance over a range of trade-offs between true positive (TP) and false positive (FP) error rates.

ROC curve is more informative than the classification table.

For logistic regression you can create a 2X2 classification table of predicted values from your model for your response if $\hat{y} = 0$ or 1 Vs. the true value of $y = 0$ or 1.

$$s_1 = \text{Sensitivity} = P(\hat{y} = 1 \mid y = 1) = \text{Prob. of } (T \mid D)$$

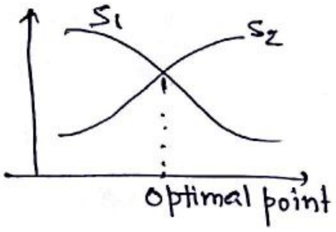
$$s_2 = \text{Specificity} = P(\hat{y} = 0 \mid y = 0) = \text{Prob. of } (\bar{T} \mid \bar{D})$$

↓ Probability of testing negative in case the subject doesn't have desease.

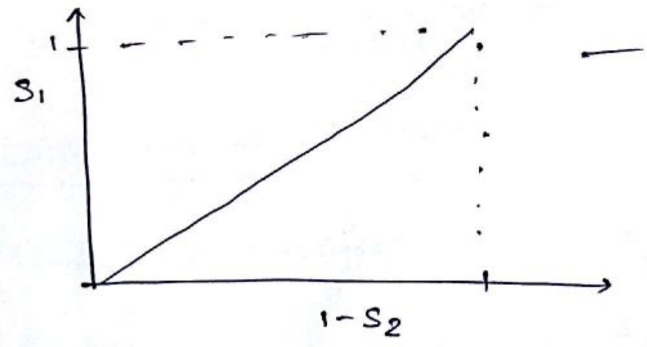
[T = test positive
D = Disease exists
$\bar{T}$ = Test negative
$\bar{D}$ = Disease doesn't exist]

$1 - S_2 = P(T \mid \bar{D})$ = False positive.

S_1 = True positive.



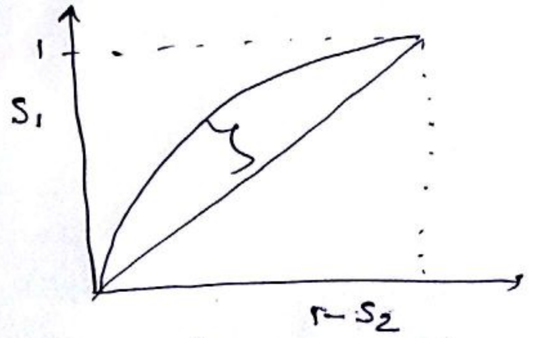
ROC Curve :



— Linear means that for all cutoff points will have same value of prob. of true positive and false positive.

(This is meaningless, because it's just like tossing a coin)

(True positive should always have higher prob. than FP)



— Higher the departure, higher is the power of discrimination.

— Area under the curve is an accepted traditional performance metric for a ROC curve. The higher the AUC, the better prediction power the model has.

[Usually it should be > 0.7 and < 0.9]

- In a goodness of fit test, rather than using the classification table, we should use area under ROC curve.
- Usually it should be greater than 0.7.
- And in most cases it should be less than 0.9.
- If ROC is > 0.9 then there is a high chance of quasi-complete separation.
- If ROC is low then (< 0.6) classification power is low.
- If ROC is around 0.9 and report of quasi-complete seperation is not clear, then it is advisable to refit the model using subsamples. You will notice that coefficients of parameters become unstable.

Classification Methodologies:-

Parametric

- Logistic Regression (essentially Binary)
- Linear Discremenant Analysis (LDA)
- Quadratic Discriminant Analysis (PDA)
  - ↳ is generally not preferred as the number of the parameters to be estimated increases manifold.

Non-parametric

- Naive Bayes
- Decision Trees (Bagging, Boosting CART, Random forest)
- K-NN
- Regression splines
- Neural Networking

Discriminant Analysis :-

1. LDA
2. QDA

Problem in classification is : Estimate $P(Y=K|\underset{\sim}{X})$ where Y is categorical response variable. Allocate to class K where $P(Y=K|\underset{\sim}{X})$ is maximum.

Logistic Regression Vs Discriminant Analysis :-

- In case the response variables (classes of Y) are well separated (w.r.t. X) logistic regression becomes unstable, then LDA is preffered.
- Discriminant analysis assumes normality for $\underset{\sim}{X}$ for different classes of Y. Large departure from normality leads to poor classification (logistic is preffered).
  Nominal or Categorical variables(X) may invalidate classification using discriminant analysis.
- When explanatory variables are measured in natural/ratio scale and has approximately normal distribution, then discriminant analysis performs well even for small data.
- Prospective data are not required for classification.
- Logistic becomes complex for $K > 2$.

Linear Discriminant Analysis approach:-

Let Y be a categorical response with k classes (assuming normal response).

We try to estimate $p_K = P(Y = K \mid \underset{\sim}{X} = \underset{\sim}{x})$

given the prior probability and the inverse probabilities

$$\pi_K = Pr(Y = k) \longrightarrow \text{unconditional preposition.}$$

$$f_K = P(\underset{\sim}{X} = \underset{\sim}{x} \mid Y = k) \rightarrow \text{inverse probability}$$

f_K is assumed to be normal with means μ_K and constant variance Σ, since estimation of f_K is difficult.

Questions:-

1. How do you check whether $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$? (Hints: Q-Q Plots, Skewness Kurtosis, A-D test, S-W test)

2. How do you check whether $\Sigma_1, \Sigma_2, \ldots, \Sigma_K$ are same or not ?

LDA with one predictor:-

$$p_K = P(Y=k \mid X=x)$$

$$= \frac{P(X=x \mid Y=k)\; P(Y=k)}{\sum_{l \geq 1}^{n} P(X=x \mid y=l)\, P(y=l)}$$

$$= \frac{\pi_K f_K}{\sum_{l=1}^{n} \pi_l f_l}$$

Take $q_K = \pi_K \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_K)^2}{2\sigma^2}}$.

$$\Rightarrow \ln q_K = \ln(\pi_K) + \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \left(\frac{x-\mu_K}{\sqrt{2}\,\sigma}\right)^2$$

Estimation of parameters:-

We need to estimate μ_K, π_K, σ.

$\hat{\pi}_K = \frac{f_K}{N}$ where f_K = No. of observations with $Y=k$

N = total sample size.

$\hat{\mu}_K = E\left(\widehat{\underset{\sim}{X} \mid Y=k}\right) = \frac{\sum x_j}{f_K}$; where x_j is taken when $Y=k$.

$\hat{\sigma}$ = pooled standard deviation from k-classes.

K – Nearest neighbour Rule :- An algorithm where we look at the K (≥ 2, ≥ 3 for classification) $\underset{\sim}{x}$ vectors nearest to the observed $\underset{\sim}{x}_0$. In case of value estimation, an average or median of the observed y values corresponding to the nearest $\underset{\sim}{x}$ values is considered.

In case of classification, majority vote is taken.
We use cross validation technique to choose the value of K.

Cross Validation :- A method to

- Check model accuracy (possibly model comparison)
- Assess the correct degree of flexibility.

Approaches :

- Hotel out sample (for test data)
- Leave one out cross validation (LOOCV)
- K-fold cross validation.

Boosting Algorithm :-

1. (Initialize) Set $\hat{f}(x) \leftarrow 0$
   $r_i \longleftarrow y_i$ for $i = 1, 2, \ldots, n$ sample size n;
   $r_i \rightarrow$ residuals
   $y_i \rightarrow$ response
2. (Computation/Fitting) Repeat the following steps:
   (a) fit a small tree with d splits, say $\hat{f}_b(\underset{\sim}{x})$ on the training data $(\underset{\sim}{X}, Y)$.
   (b) Update the tree $\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}_b(\underset{\sim}{x})$ where λ is the shrinkege paramet
   (c) Update the resideal $r_i \longleftarrow r_i - \lambda \hat{f}_b(y)$
3. Output: $\hat{f}_b(\underset{\sim}{x}) = \sum \lambda \hat{f}_b(\underset{\sim}{y})$

Parameters: $\lambda = 0.01$
$d =$ split size (1 or 2)

Regression Splines:- (Variants of regression models that can take a wide variety of smooth shape)

Basis function :- $E(Y|X)$

$$E(Y|X) = \beta_0 + \beta_1 l_1(X) + \beta_2 l_2(X) + \cdots + \beta_K l_K(x) + \epsilon$$

- The function $l_i(x)$ are called basis functions.

- Typically the basis function in splines would be restricted to polynomials $[1/x, \sqrt{x}, x^2, x^3, \ln x]$

Approach :- In splines, we divide the entire range of x into a set of subranges. Different models are fitted at different subranges of x.

Spline :- Fitting different models (often referred to as piecewise approach) to the entire range of X divided into a set of subranges.

Step functions :- Examples : Indicator function:

$$I_0(x) = \begin{cases} 1 & \text{when } x < x_1 \\ 0 & \text{ow} \end{cases}$$

$$I_1(x) = \begin{cases} 1 & \text{when } x_1 \le x < x_2 \\ 0 & \text{ow} \end{cases}$$

$$\vdots$$

$$I_K(x) = \begin{cases} 1 & x > x_K \\ 0 & \text{ow} \end{cases}$$

Note :- When applying step function on an explanatory variable measured in ratio scale we should be careful. Particularly we should verify whether conversion of ratio scale measurement to ordinal has a significant impact on the model imple-cation.

Cubic Splines :- The most commonly used spline where polynomials of degree 3 are fitted on each subrange.

$$E(Y/X) = \begin{cases} \beta_{10} + \beta_{11} x + \beta_{12} x^2 + \beta_{13} x^3 + \epsilon & \text{where } x < K \\ \beta_{20} + \beta_{21} x + \beta_{22} x^2 + \beta_{23} x^3 + c & \text{where } x \geq K \end{cases}$$

Definition :- The boundary points of the subranges are called knots.

Constraints imposed on Splines:- The fitted spline needs to be continuous at knots and also need to be smooth. In order to meet these constraints, the spline software impose a number of constraints like this, one such is 1st and 2nd derivatives also need to be continuous.

Definition:- The regions below the smallest knot and above the highest knot are called boundaries.

Fitting Constrained piecewise Polynomial :-

In the case of cubic polynomial the constrained function can be written as a basis function representation with K+4 degree of freedom.

$$y = \beta_0 + \beta_1 l_1(x) + \beta_2 l_3(x) + \cdots + \beta_{K+3} l_{K+3}(x) + \epsilon$$

We arrive this representation using a truncated power basis function

$$h(x) = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{ow} \end{cases}$$

Fitting cubic Polynomial:- In order to fit a piecewise polynomial of degree 3, we use 3 basis functions x, x^2, x^3 and K truncated power basis functions; where K is the no. of knots.

— Parameters can be estimated here using least square.

Note : Cubic splines are found to work better than high order polynomial models. However these models are unstable sometimes at the boundaries.

Natural Splines: When cubic spline becomes unstable at boundaries, we use natural splines (linearity constraint is imposed at the boundary).

Deciding about number of knots :-

Choices are : If 3 knots : at 25th, 50th, 75th percentile.
If 4 knots : at 20th, 40th, 60th, 80th percentile.

Ways of fitting :-

- Decide about a few alternative number of knots (at predefined cut points at specified percentile of x).
- Fit natural splines and cubic splines for each model.
- Use K-fold and LOOCV Cross validation to choose.

Alternative Approach :-

Spline Smoothing :- We use the concept of cost and complexity. We find an estimator $g(\cdot)$ such that

$$\sum (y_i - g(x_i))^2 + \lambda \int (g''(t))^2 dt \quad \text{is minimized.}$$

λ is called tuning parameter. (non-negative).

Note : The 2nd derivative gives the change of slope of g and hence measures the flexibility in same sense. As $\lambda \to \infty$, $g \approx$ linear function.

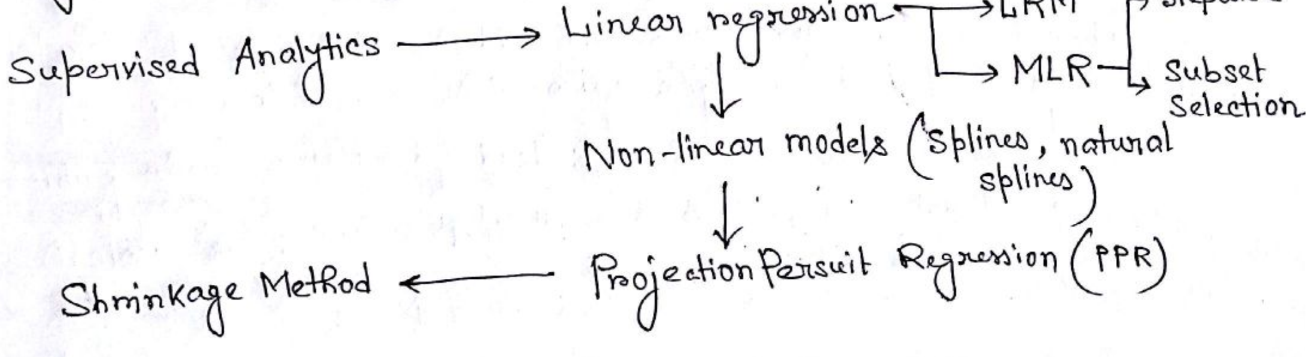
— We need to choose λ. The method of cross validation is used for different alternative values of λ.

Ridge Regression :- (Shrinkage Method)

Methods like Stepwise Regression and Best subset selection selects a subset of variables (either selected or not). In shrinkage methods, i.e., Ridge Regression, we try to reduce the individual coefficients : Minimize $\sum_1^n \left(y_i - \beta_0 + \sum_1^n \beta_j x_{ij}\right)^2 + \lambda \sum_1^n \beta_j^2$

We get, $\hat{\beta} = (X'X + \lambda I)^{-1} X'Y$

Way Out :-



Scale Invariance :- Least square solutions are scale invariant. Consequently, $\beta_j X_j$ is independent of scale.

Ridge solutions are not scale independent because of the pentalty term.

Transformation of X_j :- We carry out the following scale tranmformation on X_j :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

Interpretation of β_0 :- Average of y where all x_j are 0.

We can centre the explanatory variables, i.e., choose $x_j' = x_j - \bar{x}_j$ to ensure that β_0 has the form $\frac{1}{n}\sum_i y_i$.

Note : In order to carry out Ridge regression, we will carry out two different transformation : centerity of x_j and scale transformation.

- We need to choose the optimal value of λ. Use k-fold cross validation to choose the model.

Projection Pursuit Regression (PPR) :- Let $\underset{\sim}{X}$ be the input feature. (p-dimensional)

Let $U_m = \omega_m' \underset{\sim}{x}$ be the projection of $\underset{\sim}{x}$ onto a different hyperplane. (ω_m is a unit vector, i.e., $||\omega_m|| = 1$)

Let $g_m(U_m)$ be any function. We define $f(x) = \sum_{m=1}^{M} g_m(U_m)$.

- When M is large, this formulation may be used to approximate a very large number of situations and is called the universal approximation.
- We take an additive model defined on a projection of the input $\underset{\sim}{X}$,

$$f(\underset{\sim}{X}) = \sum_{j=1}^{M} g_j(\omega_j' \underset{\sim}{X}).$$

Estimation of Parameters :-

1. Estimation of g :- We estimate g for a given ω using any smoothing technique (typically smoothing spline).
2. Estimation of ω :- We start with an initial value ω_0 and estimate ω using iterative Gauss-newton method.
3. Update $y_i \leftarrow y_i - g(\omega' \underset{\sim}{x})$ and repeat step 1 and 2.

- M is predefined parameter and is decided on the basis of cross validation.

Neural Network:- Essentially a multistage regression model with a number of "hidden" layers built in a fashion of PPR.

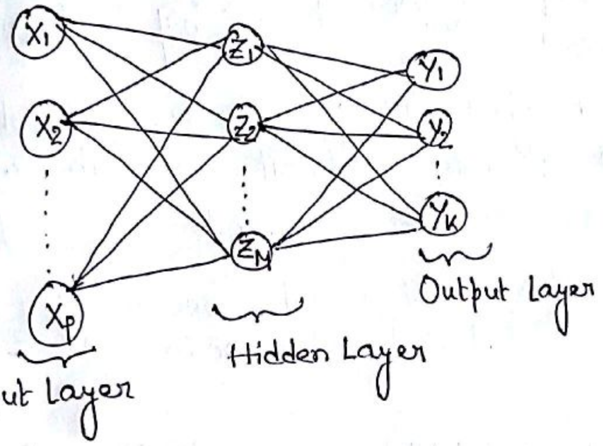
Suppose we have a K-class classification model:

$$Y_1 \quad Y_2 \cdots\cdots Y_K$$
$$X_1 \quad X_2 \cdots\cdots\cdots X_p$$

Here we estimate $P(Y=j|\underset{\sim}{X})$; $j=1,2,\dots,K$ as output function

$$g_j(\underset{\sim}{X}) = \frac{e^{t_j}}{\sum e^{t_j}} \text{ ; where } t_j = \beta_j'\underset{\sim}{x} .$$

We propose a network structure as follows:



Model Structure:-

Hidden layer:

$$Z_m = \sigma\left(\alpha_{0m} + \alpha_m'\underset{\sim}{X}\right) ; \quad m=1,2,\dots,M$$

$$T_K = \beta_{0K} + \sum_j \beta_{jK} Z_j ; \quad K=1,2,\dots K.$$

$$\boxed{f_K(\underset{\sim}{X}) = g_K(T_K)} \rightarrow \text{o/p funct}$$

Activation function (σ):- Considered to be a step function that fires only when the input crosses a threshold.

Currently the activation function is taken as Sigmoid function

$$\sigma(v) = \frac{1}{1+e^{-v}} .$$

— For value estimation case, we take output function as identity function.

— In case we have one hidden layer, Neural network becomes equivalent to PPR.

Estimation of model parameters:- Let $p_{j|\underset{\sim}{x}}$ be the probability that the response takes the value $j/\underset{\sim}{x}$.

Classification Set up:- Estimate parameters by minimizing cross entropy

$$R = -\sum y_i \ln p_i$$

Fitting NN Models:-

Z_m → Hidden units (function of projection of $\underset{\sim}{x}$)

T → Linear function of Z.

g_K → Output function.

Use the NN model for the purpose of prediction only.

Inventary Problem:- We need to estimate the service level (prob. of being able to provide material and corresponding capital costs (level of inventary).

- Monte Carlo Model,

Model Validation & Recalibration:- Look at validation of model using simulation.

Model fitting in NN:- Suppose the set of parameters is given by Q.

In Value estimation, we try to minimize the sum of squared error

$$R(Q) = \sum_{1}^{K} \sum_{1}^{N} \left(y_{IK} - f_K(\underset{\sim}{x_i}) \right)^2 \quad ; \quad \begin{array}{l} N \text{ observations \&} \\ K \text{ response variable.} \\ (\text{Usually } K=1) \end{array}$$

Gives us solution through least square using iterative approach.

Classification Problem:- We have K classes.

Defining variables Y_{ij} ; $i = 1, \ldots, N$
$j = 1, \ldots, K$

$$y_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ obsn. is in class } j \\ 0 & \text{OW} \end{cases}$$

$$\text{cross entropy} = -\sum p_{jK} \log \hat{p}_{jK}$$

In classification set up, we minimize deviance, in regression tree, we use sum of squared error (SSE).

Naive Bayes Classification:-

- Prospective study / Follow-up Study is difficult in business as it deals with treatment effects over period of time. When a treatment is applied to a set of people (and not applied on another subset) and the outcome is noted later.
- Retrospective study is where we identify the people who buy a product and then we move backward to find their characteristics.

In business analytics, we normally deal with observational study.

	Lung Cancer		
Smoking	Yes	No	Total
Yes	153	73	226
No	47	127	174
Total	200	200	400

$$P(\text{Lung Cancer} \mid \text{Smoking}) = \frac{153}{200}.$$

$$P(\text{Lung Cancer} \mid \text{Non-smoker}) = \frac{47}{200}.$$

Conditional probability can be computed only for prospective study. This is a retrospective study. This is a retrospective study and thus conditional prob. doesn't hold good.

Let Y_i be a categorical response variable with K classes,
Let $A_1, A_2, \ldots, A_p$ be different conditions that impact

$$P(Y=j \mid A_1, A_2, \ldots, A_p) \; ; \; j = 1, 2, \ldots, K.$$

If $P(Y=j \mid A_1, A_2, \ldots, A_p)$ are estimable, we allocate the subject to class j where this probability is maximum. (Recall Bayes Optimality Criterion)

Note:- In most practical situations retrospective studies would be conducted and hence $P(Y=j \mid A_1, \ldots, A_p)$ is not estimable.
However, $P(A_i \mid Y=j)$; $i = 1, 2, \ldots, K$ can be estimated from a case control study.

By Bayes theorem:-

$$P(Y=j \mid A_1, A_2, \dots, A_p) = \frac{P(A_1, A_2, \dots, A_p \mid Y=j)\, P(Y=j)}{P(A_1 \cap A_2 \cap \dots \cap A_p)}$$

$$\Leftrightarrow \text{Maximizing } P(A_1 A_2 \cdots A_p \mid Y=j)\, P(Y=j)$$

<u>Rule:</u> Allocate that subject to that j where $P(A_1 A_2 \cdots A_p \mid Y=j)$ is maximum.

<u>Naive Bayes Assumption:-</u> $A_1, A_2, \dots, A_p$ are conditionally independent given $Y=j$.

Under this assumption $P(A_1 A_2 \cdots A_p \mid Y=j) = \prod_{i=1}^{p} P(A_i \mid Y=j)$

(It has been noted that $P(A_i \mid Y=j)$ are estimable under a case-control set up).

Decision Trees [Non-parametric Method]:-

Used both for value estimation and classification.

Suppose the region covered by the explanatory variable R is partitioned into $R_1, R_2, \dots, R_M$.

[Let the explanatory variables be such that $X_{0i} \le x_i \le X_{1i}$; $i=1,2,\dots,K$
There $R = \{(x_1, x_2, \dots, x_K) \mid x_i \in X_1\}$]

Let $E(Y \mid \underset{\sim}{X} \in R_j \,;\, j=1,2,\dots,M)$ be the conditional expectation of Y given that $\underset{\sim}{X} \in R_j$.

$$E(Y \mid \underset{\sim}{X}) = \sum I_j(R_j)\, E(Y \mid X \in R_j).$$

Difference & Similiarity between Regression Model and Decision Tree:

Similiarity : In Both cases we find $E(Y \mid X)$.

Difference : In Decision trees, no linearity assumption is required where as for regression the linearity is necessary.

<u>Input Data:</u>

X_1	X_2	$\cdots$	X_K	Y
x_{11}	x_{12}	$\cdots$	x_{1K}	y_1
x_{21}	x_{22}	$\cdots$	x_{2K}	y_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_{N1}	x_{M2}	$\cdots$	x_{NK}	y_N

Greedy Algorithm :- Whichever value at a particular point of time maximizes or minimizes the objective function. No backtracking.

[eg: Suppose someone is waiting for a bus and boards the first bus he gets. May be he would have got a better bus if he would have waited.]

Arriving at the partition :-

1. **Objective :** To arrive at a partition such that the mean squared error $\sum(y_i-\hat{y}_i)^2$ is minimum.

2. We use a greedy algorithm as follows :-

Step1: Compute baseline SSE = $\sum(y_i-\bar{y})^2$.

Step 2: For each X_i choose different cut point and divide the input space into R_1 and R_2. Choose that partition where the decrease of MSE is maximum.

$$\left[\sum_{X\in R_1}(y_i-\hat{y}_1)^2+\sum_{X\in R_2}(y_i-\hat{y}_2)^2\right]$$

Step 3: Continue till all variables are exhausted.

MLR

$$E(Y|X)=\beta_0+\sum_{j=1}^{p}\beta_j x_j$$

Decision Trees

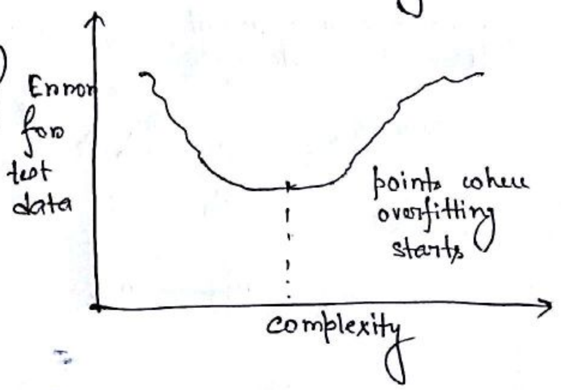
$$E(Y|X)=\sum_{j=1}^{M} I_j(R_j)\,E(Y|\underset{\sim}{X}\in R_j).$$

Note: A tree grows very fast. Consequently trees may result in "over fitting" (saturated model).

Non-parametric → More flexible, less interpretable
Parametric → Less flexible, more interpretable
[But Decision tree is an exception]

Overfitting :- A fitted model that fits the training data very well but doesn't fit the test data well in likely to be overfitted.

A model whose accuracy of fit (test data) decreases (from a given complexity) as complexity increases, is said to be overfitted.



Training Data :- The data used to fit the model. Use 70% of the available data are used for training.

Test Data :- A portion of the data (separate from the training data) that is used to check the accuracy of the fitted model.

Validation Data :- A portion of the data (other that training & test data) that is used to compare competing models.
[Usually used in competitions for comparing models, never available to competitors.]

Cross Validation :-

→ Bootstrap :- Introduced by Effron. In this technique we attempt to assess the sampling fluctuations by taking samples repeatedly from the observed data.
[Concept of Bagging is almost similar but objective is different for two cases]

→ Jackknife :- (Leave one out) Use the entire data bearing one point and predict the same using the other observation.

→ Leave one out Cross Validation :- Suppose there are n observations. The prediction in this case is repeated n times.

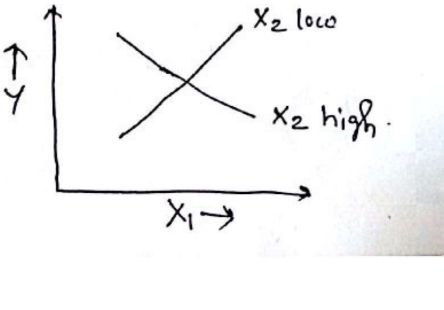
→ K-fold cross validation :- Divide the data into K subgroups. Use K-1 subgroups to build the model (tree) and 1 subgroup to test it. Apply it for all the cases.

Building Pruned Trees :- (i) Build a large tree using recursive binary split.
(ii) Prune the tree using tuning parameter α.
(iii) (Main step) : Use K-fold cross validation to choose the 'right' α.

(a) for different values of α, estimate the test errors, for each α as a function of α.
(b) compute the average test error
(c) choose α with minimum K-fold cross validation error.

Question :- Can you use the tree for interactions between explanatory variables ?

→ By drawing diagrams for finding interactions.



Each split divides the data into two parts. Suppose the split is done on salary. If salary $< x$, then it is dependent on age whereas if salary $\geq x$, then it depends on level of education. Thus this is an example of interaction in a tree.

Question:- When can we use decision trees satisfy for the purpose of explanation?

→ Use the underlying interaction structure. In case the structure keeps changing on different subsamples to a large extent explanation is risky.

Classification Tree:- The response variable is categorical. The aim may be to predict the outcome (predictive analytics) or understand why it happens (explanatory analytics).

Measure of Partition effectiveness:-

(a) Rate of Misclassification:- Suppose the response has K levels. Suppose the proportion of occurance of these levels in the subset R_j are

p_{R_jK} ; $k = 1, 2, \ldots, p$.

Let $p_{R_jl} \geq p_{R_jK}$ $\forall K$

i.e., p_{R_jl} is maximum.

Then rate of misclassification $= 1 - p_{R_jl}$.

(b) Gini Index:- Sum of → $\sum_{K=1}^{p} p_{R_jK}\left(1-p_{R_jK}\right)$, p classes.

Measures node impurity. If Gini index is higher when node is impure.

(c) Cross Entropy:- $\sum -p_{R_jK} \ln p_{R_jK}$

In classification tree we use node purity and in regression tree we use sum of square of error.

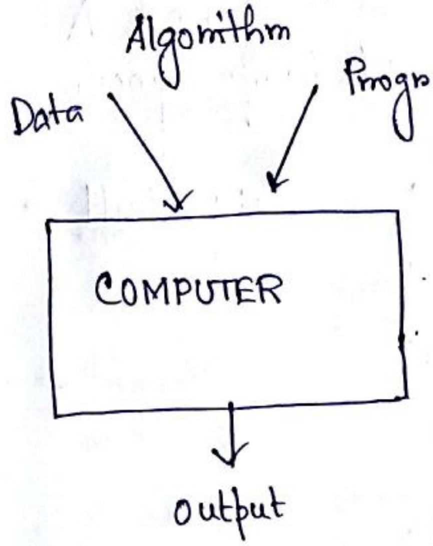
MACHINE LEARNING

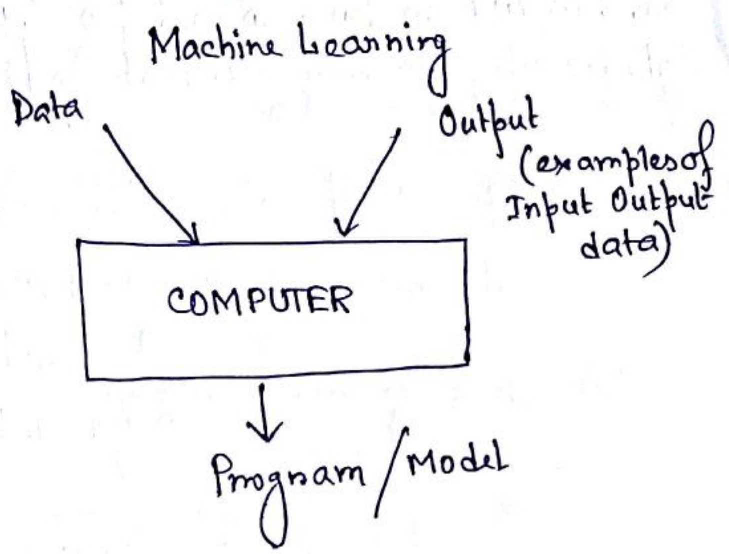
A machine that is intellectually capable as much as humans has always fired the imagination of learners and computer scientists.

- History :-
  - 1950s — Samuel's checker-playing program
  - 1957 — Neural network : Rosenbiatt's perceptron
  - 1960s — Pattern Recognition
  - 1969 — Minsky and Papert prove limitations of Perceptron.
  - 1970s —
    - symbolic concept induction
    - Expert systems and Knowledge acquisition bottleneck.
    - Quinlan's ID3
    - Natural Language processing (symbolic)
  - 1980s —
    - Advanced decision tree & rule mining
    - Resurgence of neural networking
    - Valiant's PAC learning theory
  - 90's ML & Statistics :
    - Support Vector Machines
    - Data Mining
    - Text learning
    - Bayes Net Learning
  - 2000s onwards :
    - Neural networks (software)
    - Deep learning
    - Big data
    - Google's self driving car

Ref. Books :- 1. Machine Learning : Tom Mitchell (1997)

2. Introduction to Machine learning by Ethem Alpaydin.





Learning : The ability to improve behaviour based on experience.

Machine Learning : explores algorithms
— learn / build models from data
— model used for prediction, decision making or solving tasks.

Tom Mitchell's definition of Machine Learning :-

A computer program is said to learn from experience E w.r.t. some class of tasks T and performance measure P if its performance on tasks in T as measured by P improves with experience E.

Applications :- — Medicine :

- Diagnose a disease : Input : Symptoms, lab measurements, test results DNA tests, ...
  Output : one of set of possible diseases, or, none of the above.
- Data mine historical medical records to learn which future patients will respond best to which treatments.

— Robot Control :

- Design autonomous mobile robots that learn to navigate from their own experience.

— Natural Language Processing, Image Processing, Speech recognition.

— Sentiment Analysis, Machine Translation.

— Financial :

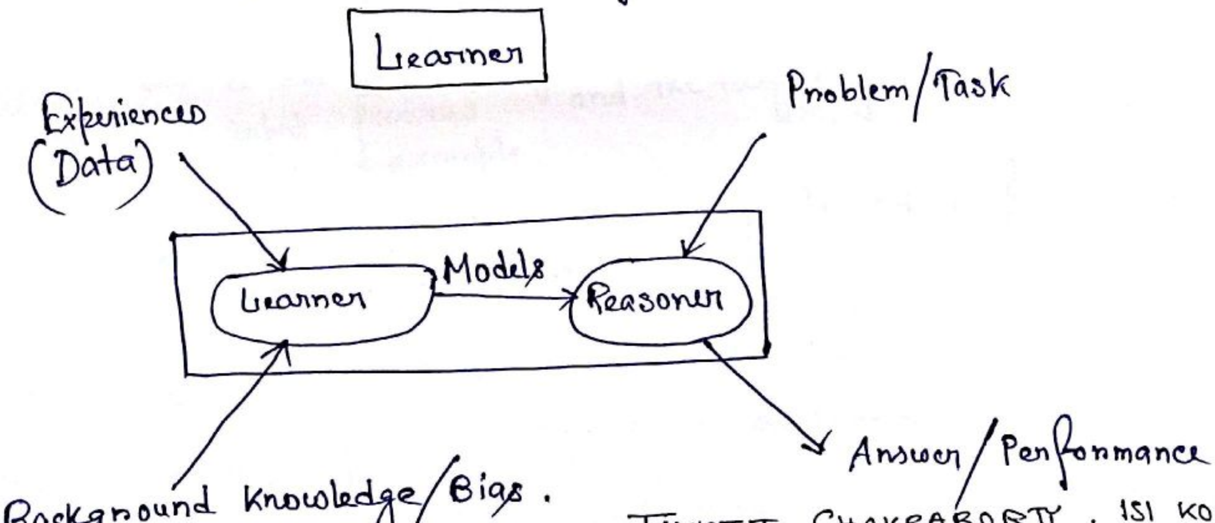
- Predict if a stock will rise or fall in the next few milliseconds.
- Predict if a user will click on an ad or not in order to decide which ad to show.

— Business Intelligence

- Robustly forecasting product sales quantities taking seasonality and trend into account.
- Identifying cross selling promotion opportunities for consumer goods.
- Identify the price sensitivity of a consumer product and identify the optimum price point that maximizes net profit.
- Optimizing product location at a super market retail outlet.

— Other Applications

- Fraud detection : Credit card providers
- Determine whether or not someone will default on a home mortgage.
- Understand consumer sentiment based off of unstructured text data.
- Forecasting women's conviction rates based off external macroeconomic factors.



— TANUJIT CHAKRABORTY , ISI KOLKATA
MAIL: tanujitisi@gmail.com
MOB: 8420253573.

—

Notes On

Business

Analytics

Logistic Regression

15.8.2016

<u>Introduction</u> : The general problem addressed by logistic regression is that of establishing relationship between certain explanatory variables — both numeric and categorical with a categorical response variable.

Logistic Regression addresses the problem of classification.

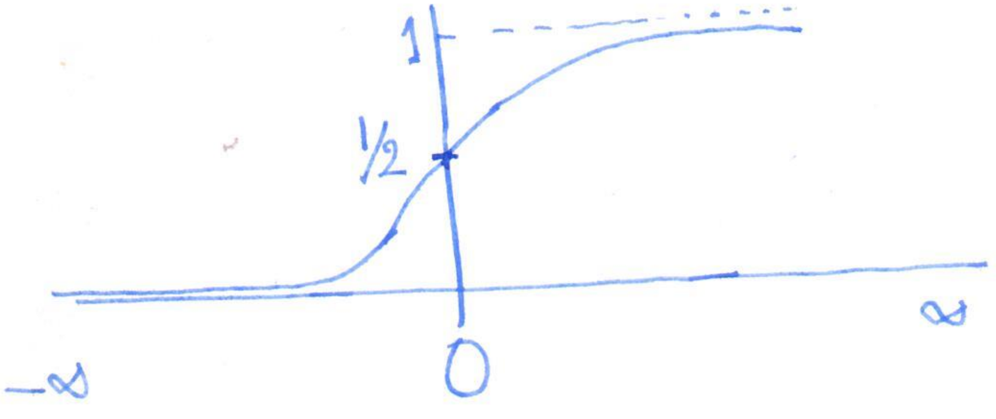
Logistic Regression is also used to estimate / assess risk.

<u>Concept of logistic function</u> :

The function $f(z) = \dfrac{1}{1+e^{-z}}$, $-\infty < z < \infty$

is called the logistic function.

Note that the logistic function has the following graph



Note that $0 \leq f(z) \leq 1$

Note further that $f(z)$ has an S-shaped curve (often referred to as the sigmoid curve)

ogistic Regression   15. 8. 2016

xamples : Usage of the sigmoid curve.

1 The dosage of insecticide has an impact of killing insects. The probability is low when dosage is very small. From a threshold, the probability increases fast.

2 The probability of a customer returning a loan may depend on factors like value of the loan and level of disposable income. In this case the variable z may be considered to be a linear combination of these variables.

Note :

Logistic Model : In general the logistic model may be considered to be the following fn.

$$z = \beta_0 + \sum_{i=1}^{p} \beta_i X_i ,$$ where $X_1, X_2 \cdots X_p$ are the explanatory variables

In essence then z is an index that combines the explanatory variables.

Logistic ~~Model~~ Regression   15. 8. 2016

~~n a log~~.
Consider a ~~class~~ binary classification problem with the explanatory variables as $X_1, X_2 \cdots X_p$ and Y being the response variable. ~~Suf~~

Suppose Y takes values 0 and 1.

Then $P(Y=1/X_1, X_2, \cdots, X_p) = \dfrac{1}{1+e^{-(\beta_0+\Sigma\beta_i X_i)}}$

The coefficients $\beta_0, \beta_1, \beta_2 \cdots \beta_p$ are the ~~pa~~ unknown parameters.

Concepts of Odds Ratios & Relative Risks

Logistic Regression

15. 8. 2016

Data Collection :

1 ~~Risk~~ Framework 1 : In certain data collection ~~risk~~ frameworks, the explanatory variables related to a subject are observed at a point of time and the outcomes are observed later. In such a case the subjects ~~may~~ being studied may have to be followed up over a period of time. Such studies are called follow-up studies.

Example 1 : We observe a set of people with certain lifestyle habits over a period of time. We then observe how many of these people have developed a particular disease.

Example 2 : We observe a set of people who have been recruited. We note their characteristics and follow them up for a period of time to see how ~~many~~ long they stay with the company (or how many of them leave within a given time frame)

2 Framework 2 : In other data collection formats we observe the outcomes of certain subjects. We then find the value of the explanatory variables pertaining to the subject.

Logistic Regression

15. 8. 2016

Logit Transformation : The transformation

~~$\text{logit}\,P(\underset{\sim}{X}) = \ln\Big(P$~~

$$\text{logit}\,(P(\underset{\sim}{X})) = \ln\left(\frac{P(Y=1/\underset{\sim}{X})}{1-P(Y=1/\underset{\sim}{X})}\right)$$

Note that $P(Y=1/\underset{\sim}{X}) = \dfrac{1}{1+e^{-(\beta_0+\Sigma\beta_i x_i)}}$

$$\Rightarrow \quad 1-P(Y=1/\underset{\sim}{X}) = \frac{e^{-(\beta_0+\Sigma\beta_i x_i)}}{1+e^{-(\beta_0+\Sigma\beta_i x_i)}}$$

$$\Rightarrow \ln\left(\frac{P(Y=1/\underset{\sim}{X})}{1-P(Y=1/\underset{\sim}{X})}\right) = \beta_0 + \Sigma\beta_i x_i$$

Note further that $\dfrac{P(Y=1/\underset{\sim}{X})}{P(Y=0/\underset{\sim}{X})}$ ~~is the~~ gives the odds of $P(Y=1)$ vs. $P(Y=0)$ for a given explamtory set up.

Baseline Odds : Note that β_0 gives the baseline odds. This refers to the odds that would result for a logistic model without any odds at all.

Logistic Regression 16. 8. 2016

nterpretation of β_j : Suppose X_j is a variable measured in the ratio scale. Then

$$\ln\left(Odds\left(Y=1/X_1=x_1, X_2=x_2 \ldots X_j=x_j, \ldots X_p=x_p\right)\right)$$

$$= \beta_0 + \Sigma \beta_i X_i$$

$$\ln\left(Odds\left(Y=1/X_1=x_1, X_2=x_2, \ldots X_j=x_j+1, \ldots X_p=x_p\right)\right.$$

$$= \beta_0 + \sum_{i=1}^{j-1} \beta_i x_i + \beta_j (x_j+1) + \sum_{i=j+1}^{p} \beta_i x_i$$

$$\Rightarrow \ln\left(Odds(Y=1/X_j=x_j+1\right) - \ln\left(Odds\left(Y=1/X_j=x_j\right) = \beta_j\right.$$

$$\Rightarrow \frac{Odds(Y=1/X_j=x_j+1)}{Odds(Y=1/X_j=x_j)} = e^{\beta_j}$$

Thus logistic regression model is one of 'Constant Odds Ratio'.

Logistic Regression

16.8.2016

Maximum Likelihood Estimates

Note that $\pi(\underset{\sim}{x}_i) = P(Y=1/X_1 = x_{i1}, X_2 = x_{i2} \ldots, X_p = x_{ip})$

$$= \frac{1}{1-e^{-(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})}}$$

gives the probability that the response takes the value 1 ~~when the~~ for a given setting of explanatory variables.

Likelihood fn. $l(\hat{\beta}) = \prod_{i=1}^{n} \pi(\underset{\sim}{x}_i)^{y_i}(1-\pi(x_i))^{1-y_i}$

follows directly from the Bernoulli pmf.

Two important likelihood functions

Likelihood of the null model : $L_0 = \hat{p}^{\sum y_i}(1-\hat{p})^{\sum(1-y_i)}$

— where $\hat{p}$ is the estimated proportion of the response variable taking value 1.

Saturated Model : $L_S = \prod y_i^{y_i}(1-y_i)^{(1-y_i)} = 1$

Deviance :- $D = -2\ln\left[\frac{\text{Likelihood of the fitted model}}{\text{Likelihood of the saturated model}}\right]$

Likelihood Ratio : $LR = -2\ln\left[\frac{\text{Likelihood of the fitted model}}{\text{Likelihood of the null model}}\right]$

Logistic Regression

16. 8. 2016

Logit transformation : The transformation

$$g(\underset{\sim}{x}) = \ln\left(\frac{\pi(\underset{\sim}{x})}{1-\pi(\underset{\sim}{x})}\right) \quad \text{where } \pi(\underset{\sim}{x}) = P(Y=1/\underset{\sim}{X}=\underset{\sim}{x})$$

has many desirable properties. The properties are given below:

a) The logit $g(\underset{\sim}{x}) = \beta_0 + \sum \beta_i x_i$ are linear in its parameters

b) The logit $g(\underset{\sim}{x})$ is a continuous function

c) $-\infty < g(\underset{\sim}{x}) < \infty$

Error in logistic ~~to~~ regression (binary) :

We estimate Y by $\pi(\underset{\sim}{x}) = P(Y=1/\underset{\sim}{x})$

If $Y=1$ then $\varepsilon = 1-\pi(\underset{\sim}{x})$ with probability $\pi(\underset{\sim}{x})$

If $Y=0$ then $\varepsilon = -\pi(\underset{\sim}{x})$ with probability $(1-\pi(\underset{\sim}{x})$

Thus $E(\varepsilon) = \pi(\underset{\sim}{x})(1-\pi(\underset{\sim}{x})) - \pi(\underset{\sim}{x})(1-\pi(\underset{\sim}{x})) = 0$

Note that ~~$\varepsilon \sim$ Bin~~ each ε_i may be considered to be a Bernoulli trial. The variance is not constant.

[If $X \sim Ber(p)$ $P(X=1) = p$, $P(X=0) = 1-p$]

$\Rightarrow E(X) = p$, $V(X) = E(p^2) - p^2 = p - p^2 = p(1-p)$

016 (1) Logistic Regression

valuation of a screening test

Let B = Risk event

B^c = Risk event does not happen

Let T = Test result is positive

T^c = Test result is negative

Prob(T/B) is called sensitivity. This is the probability of the test showing positive result given that the risk event turns out to be true.

Examples:

1 Suppose on the basis of a logistic regression model, a transaction is classified to be ~~fau~~ fradulant. Sensitivity is the probability that the model identifies a transaction to be fraudulant when it actually is fraudulant.

2 Similar logic is applicable when a model is used to ~~cla~~ classify a loan application.

Prob$(T/\bar{B})$ is called specificity. This is the probability of a false alarm, i.e. the model identifies a transaction to be fraudulant when in reality it is not.

Logistic Regression

22.8.2016

oodness Of Fit : Basic criteria for goodness-of-fit is ~~that~~ that the distances between the observed & estimated values be unsystematic & within the variation of the model. This criteria is not satisfied in classification matrix.

~~Sensitivity & Specificity from the classification matrix~~

~~Sensitivity and specificity~~

Drawbacks of classification table

a) Classification is sensitive to the relative size of the component groups ~~an~~ and always favours classification into the larger group (i.e. probability of correctly classifying when a subject belongs to the larger group is high)

b) The classification matrix converts a probability – an outcome measured on a continuum into a ~~dico~~ dichotomous variable leading to substantial loss of information.

c) The sensitivity and specificity measured from a 2 X 2 ~~specifica~~ classification table depends entirely on the distribution of the subjects rather than superiority of a model.

ogistic Regression 23.8.2016

oodness Of Fit (Classification Tables)

Impact of Distribution of Subjects on Sensitivity and Specificity :

Consider the following hypothatical case (Hosmer and Lameshow, page 157)

Classification through model	Observed Values		Total
	~~0~~ 1	0	
1	16	11	27
0	131	417	548
Total	147	428	575

Sensitivity = Prob(~~Correct Classification~~ Predicted disease / Disease)
(Let disease = 1) $= 16/147 = 10.9\%$

Specificity = Prob(Predicted disease free / No disease)
$= 417/428 = 97.4\%$

Overall correct classification $= \frac{16+417}{575} = 0.753$

Logistic Regression

23. 8. 2016

Classification tables :

Notice that in the above table the distribution of the subjects with disease probability $> 0·50$ actually had about 40% of the subjects without disease. This implies that the estimated probabilities were $> 0·50$ but sufficiently close to 0·5.

[Note: Suppose among n subjects, the probability of disease is a constant, say $\hat{\pi}$. Then $n\hat{\pi}$ subjects are expected to actually have the disease and $n(1-\hat{\pi})$ would not develop the disease. Thus, when $\hat{\pi} > 0·50$, $n(1-\hat{\pi})$ subjects are expected to be misclassified]

Suppose in the same table given above, the probability of having ~~the~~ or not having the disease are as follows:

If $\hat{\pi} < 0·50$ then $\hat{\pi} = 0·05$

and if $\hat{\pi} \geq 0·50$ then $\hat{\pi} = 0·95$

Assuming that the classification rule remains same, the table becomes as follows:

Classification	Observation		Total
	1	0	
1	26	1	27
0	27	521	548
Total	53	522	575

$\text{Sensitivity} = \frac{26}{53} = 0·491$

$\text{Specificity} = \frac{521}{522} = 0·99$

ogistic Regression

23. 8. 2016

Classification Table

Note: The above computations were carried out under the assumption that the model is correct, i.e. the estimated probability of disease is correct.

Thus the sensitivity & specificity depend heavily on the subject mix.

Area under Receiver Operating Characteristic Curve:

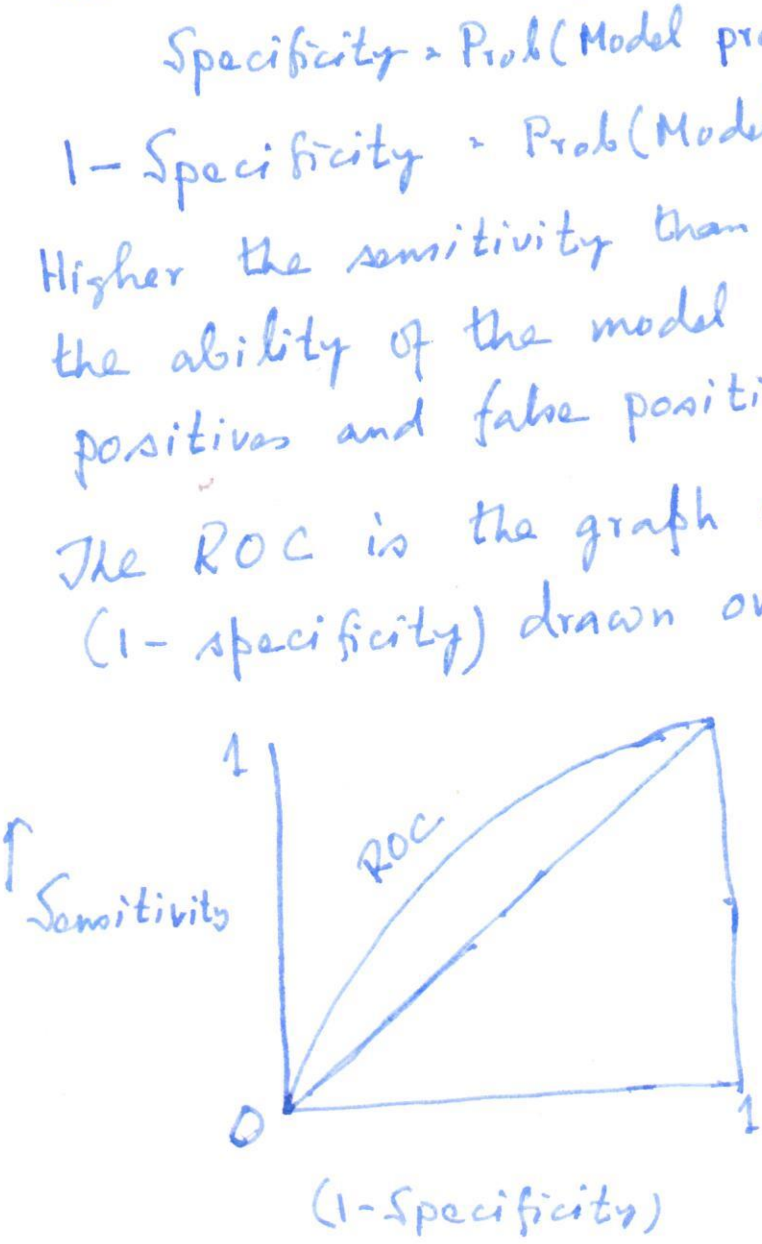
Note that Sensitivity = Prob(Model predicts disease/disease)

Specificity = Prob(Model predicts no disease/no disease)

1 − Specificity = Prob(Model predicts disease/no disease)

Higher the sensitivity than (1-specificity) better is the ability of the model to discriminate true ~~pos~~ positives and false positives.

The ROC is the graph of ~~sensivit~~ sensitivity vs. (1-specificity) drawn over all possible cut points.



When the ROC is on the diagonal line (area = 0·5) there is no discrimination.

gistic Probability

$$P(Y=1/\underset{\sim}{X}) = \frac{1}{1+e^{-(\beta_0+\Sigma\beta_i x_i)}} = \frac{1}{1+\frac{1}{e^{\beta_0+\Sigma\beta_i x_i}}}$$

$$= \frac{e^{\beta_0+\Sigma\beta_i x_i}}{1+e^{\beta_0+\Sigma\beta_i x_i}}$$

Covariate pattern

Fully parameterized model

Saturated model

Deviance

$$P(Y=j/A_1, A_2, \cdots A_p) = \boxed{P(A_1 A_2 \cdots A_p/Y=j)}\ \frac{P(Y=j)}{P(A_1 A_2 \cdots A_p)}$$

~~P(~~

Assume conditional independence

$$P(A_1 A_2 \cdots A_p/Y=j) = P\cancel{(A_1/Y=j)} \prod_{i=1}^{p} P(A_i/Y=j)$$

$$\Rightarrow P(Y=j/A_1, A_2 \cdots A_p) = \frac{P(Y=j).\prod_{i=1}^{p} P(A_i/Y=j)}{\boxed{P(A_1 A_2 \cdots A_p)}}$$

$$\propto P(Y=j).\prod_{i=1}^{p} P(A_i/Y=j)$$

Use Bayes' optimality criteria.

Maximum Likelihood Estimators

16. 8. 2016

Likelihood function : The probability (likelihood) of the observed sample given the parameter. The likelihood function is a function of the parameter.

Suppose θ is the unknown parameter. We write the likelihood function as $L(\theta/x_1, x_2 \cdots x_n)$

Note: Likelihood function is ~~no~~ not probability. If we sum (or integrate) $L(\theta/x_1, x_2 \cdots, x_n)$ over all possible values of θ, it will not become 1.

Maximum Likelihood Principle : Choose as your estimates those values of the parameter that maximizes likelihood of the observed data.

Log likelihood : The natural logarithm of the likelihood function. It is often preferable to work with the log likelihood for both practical & theoratical reason.

Note: Likelihood fn. $L(\theta/x_1, x_2 \cdots x_n) = \prod_{i=1}^{n} p(x_i, \theta)$

The log likelihood converts the product into sum & is hence easier to handle.

Secondly all theoretical results concerning maximum likelihood are based on log likelihood.

Maximum Likelihood Estimators

16.8.2016

Advantages of log likelihood : Log likelihoods increase the numerical stability of the estimates. Likelihood functions are products of marginal probabilities and tend to become very small for large samples. Log likelihoods are large negative numbers and hence their usage improves ~~stability~~ numerical stability.

Kernel likelihood & full likelihood : The likelihood function can be written as:

$$L(\theta, \underset{\sim}{x}) = k(\underset{\sim}{x}) . p(\underset{\sim}{x}, \theta) \propto p(x, \theta)$$

— $k(\underset{\sim}{x})$ is merely a function of the observed data and does not involve the parameter to be estimated.

Example : Let $x_1, x_2, \cdots x_n$ ~~to~~ be independent random observations from the same Poisson population with unknown parameter λ.

Then $L(\lambda / x_1, x_2 \cdots x_n) = \prod_{i=1}^{n} p(x_i, \lambda)$

$$= \prod_{i=1}^{n} \frac{e^{-\lambda} . \lambda^{x_i}}{x_i!}$$

$$= k(\underset{\sim}{x})\, p(\underset{\sim}{x} / \cancel{\theta}\, \lambda)$$

— Where $K(x_1, x_2 \cdots x_n) = \cancel{\frac{1}{tt}} \prod_{i=1}^{n} 1/x_i!$

$p(\theta / x_1, x_2 \cdots x_n) = \cancel{\prod_{i=1}^{n} \binom{n}{x_i} e^{-\lambda}} . e^{-n\lambda} \lambda^{\Sigma x_i}$

1 —

Classification

26.9.2016

Linear Discriminant Analysis: In logistic regression we attempted to model $P(Y=K / \underset{\sim}{X} = \underset{\sim}{x})$.

The LDA provides an alternative ~~of~~ approach. In this approach we model the distribution of the predictor variables $\underset{\sim}{X}$ separately for each response class and then use Bayes' theorem to get $P(Y=k / \underset{\sim}{X} = \underset{\sim}{x})$

(In discriminant analysis we get the inverse probability)

Why use LDA instead of logistic regression?

- When classes are well separated, logistic regression is very unstable. (Remember cases of complete and quasi complete separation)
- If n is small and the distribution of ~~the~~ $\underset{\sim}{X}$ is approximately normal, LDA perform well
- Logistic regression tends to become complex for multiple response classes (>2)

Using Bayes' thm for classification:

Let $(y_i, \underset{\sim}{X}_i)$ $i=1,2,\ldots n$ be the observations

Y is a categorical variable with $k (\geq 2)$ classes.

Thus y_i can take values $1, 2, \ldots K$

Let $\pi_K = P(Y=K)$ i.e. π_K gives the prior probabilities of the different classes. (This is the unconditional probability)

Classification

26.9.2016

LDA continued ...

Let $f_K(\underset{\sim}{x}) = P(X = \underset{\sim}{x} / y = k)$

Then $P(Y = K / X = \underset{\sim}{x}) = \dfrac{f_K(\underset{\sim}{x}) \cdot \pi_K}{\sum \pi_l f_l(\underset{\sim}{x})}$

Let $p_K(x) = P(Y = K / \underset{\sim}{X} = \underset{\sim}{x})$

We want to estimate $p_K(\underset{\sim}{x})$

Note that estimating π_K is easy.

Thus we only need to find $f_K(x) = P(\underset{\sim}{X} = \underset{\sim}{x} / Y = K)$

LDA for one predictor

Assume that $f_K(x) = \dfrac{1}{\sigma_K \sqrt{2\pi}} e^{-\frac{(x-\mu_K)^2}{2\sigma_K^2}}$

$$\text{Now, } p_K(x) = \frac{\pi_K \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_K)^2}{2\sigma^2}}}{\prod_{i=1}^{p} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\cancel{\Sigma}(x-\mu_i)^2}}$$

Assumig $\sigma_1 = \sigma_2 = \dots = \sigma_p$

Bayes' ~~optimait~~ optimality criteria leads us to allocate to class $j \ni p_j(x)$ is highest.

[We take a simple Gaussian case. In class k,

$X \sim N(\mu_K, \sigma)$]

LDA Classification 26.9.2016

Ignoring the denominator (constant) & taking logarithms we get

$$\delta_K(x) = \ln(\pi_K) + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x-\mu_K)^2}{2\sigma^2}$$

$$= \ln(\pi_K) + C - \frac{x^2 + \mu_K^2 - 2x\mu_K}{2\sigma^2}$$

Ignoring C we may rewrite $\delta_K(x)$ as

$$\delta_K(x) = \frac{x.\mu_K}{\cancel{2}\sigma^2} - \left(\frac{x^2}{2\sigma^2}\right) - \frac{\mu_K^2}{\cancel{2\pi}2\sigma^2} + \ln(\pi_K)$$

$$\Longrightarrow \delta_K(x) = \frac{x\mu_K}{\cancel{2}\sigma^2} - \frac{\mu_K^2}{2\sigma^2} + \ln(\pi_K)$$

For a 2 class problem with $\pi_1 = \pi_2$

$$\delta_1 - \delta_2 = \frac{x\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} - \frac{x\mu_2}{\sigma^2} + \frac{\mu_2^2}{2\sigma^2}$$

$$= \frac{x}{\sigma^2}(\mu_1 - \mu_2) - \frac{1}{2\sigma^2}(\mu_1^2 - \mu_2^2)$$

Note: Even when we are reasonably sure that $x_i \sim N(\mu_i, \sigma)$; we still have to estimate $\mu_1, \mu_2, \cdots \mu_p$; ~~&~~ $\pi_1, \pi_2, \cdots \pi_p$; and σ.

LDA **Classification** 27.9.2016

Assume that $\underset{\sim}{X} = (X_1, X_2, \cdots X_p) \sim N(\underset{\sim}{\mu}_{p\times 1}, \Sigma_{p\times p})$

Then $f(\underset{\sim}{x}) = \dfrac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})}$

It can be shown that

$$\delta_K(x) = \underset{\sim}{x}'_{1\times p}\,\Sigma^{-1}_{p\times p}\,\mu_{K\,p\times 1} - \frac{1}{2}\mu_K'\Sigma^{-1}\mu_K + \ln(\pi_K)$$

We allocate to class K when $\delta_K(x)$ is max.

Note: Observe how the area under the ROC ~~curve~~ curve would be applicable in this context.

Quadratic Discriminant Analysis (QDA):

LDA assumes that observations in the different classes have class specific mean vectors $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \cdots, \underset{\sim}{\mu}_K$, but a common variance-covariance matrix Σ.

In QDA ~~we~~ we assume that observations $\underset{\sim}{X}$ from the k-th class are such that $\underset{\sim}{X} \sim N(\underset{\sim}{\mu}_K, \Sigma_K)$

Note that $p_K(\underset{\sim}{X} = x / Y = K) = \dfrac{\pi_K f_K(x)}{\sum \pi_l f_l(x)}$

QDA Classification 27.9.2016

Note further that

$$f_K(\underset{\sim}{x}) = \frac{1}{(\sqrt{2\pi})^{p/2}|\Sigma_K^{-1}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\mu_K)'\Sigma_K^{-1}(\underset{\sim}{x}-\mu_K)}$$

$$\Rightarrow p_K(\underset{\sim}{X} = \underset{\sim}{x} / Y = K) = \frac{\pi_K \cdot \frac{1}{(\sqrt{2\pi})^{p/2}|\Sigma_K^{-1}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_K)'\Sigma_K^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_K)}}{\prod_{l=1}^{K} (\cdot\ ---)}$$

$$\delta_K(\underset{\sim}{x}) = \ln(\pi_K) - \frac{1}{2}\ln|\Sigma_K| - \frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_K)'\Sigma^{-1}(\underset{\sim}{x}-\mu_K)$$

~~No~~ Comparison between LDA and QDA

- No. of parameters : Estimation of a variance-covariance matrix requires estimation of $p(p+1)/2$ parameters. Thus QDA requires estimation of many more parameters compared to LDA.
- LDA is much less flexible and hence has substantially lower variance
- If equality of variance assumption is badly off the mark, LDA tends to perform much worse compared to QDA

Resampling Methods

26.9.2016

Bootstrap : Typically estimates the expected prediction error quite well.

Let $\tilde{Z} = (\tilde{z}_1, \tilde{z}_2, \cdots \tilde{z}_N)$ be the training data

where $\tilde{z}_i = (\tilde{x}_i, y_i)$, $i = 1, 2 \cdots N$

The basic idea is to randomly draw data sets B times from the training data

We refit the model to each of the bootstrap data sets and examine the behaviour of the fits over the B ~~rep~~ replications.

Let S(Z) be any quantity computed from the training data, Z.

$$\text{Then} \quad V(\widehat{S(Z)}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(S(Z_b^*) - \bar{S}^*\right)^2$$

$$\text{Where} \quad \bar{S}^* = \frac{1}{B}\sum_{b=1}^{B} S(Z_b^*)$$

Note : $V(\widehat{S(Z)})$ can be thought of as a Monte Carlo estimate of $V(S(Z))$

[**Note** - Cross validation explicitly uses non-overlapping data. In case this condition is violated, overfitted samples may look very attractive]

Resampling Methods

26.9.2016

Bootstrap : Can be applied in a wide range of statistical learning methods to compute measures of variability (eg. SE) or other statistics otherwise difficult to obtain and are not automatically reported by statistical software.

$$V(\alpha X + (1-\alpha)Y) = \alpha^2 V(X) + (1-\alpha)^2 V(Y) + 2\alpha(1-\alpha)\,\mathrm{Cov}(X,Y)$$

$$= \alpha^2\sigma_X^2 + \sigma_Y^2 + \alpha^2\sigma_Y^2 - 2\alpha\sigma_Y^2$$

$$+ 2\alpha\sigma_{XY} - 2\alpha^2\sigma_{XY}$$

$$\frac{\partial V(\alpha X + (1-\alpha)Y)}{\partial \alpha} = 0$$

$$\Rightarrow \quad 2\alpha\sigma_X^2 + 2\alpha\sigma_Y^2 - 2\sigma_Y^2 + 2\sigma_{XY} - 4\alpha\sigma_{XY} = 0$$

$$\Rightarrow \quad \alpha(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) = \sigma_Y^2 - 2\sigma_{XY}$$

$$\Rightarrow \quad \hat{\alpha} = \frac{\sigma_Y^2 - 2\sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Bootstrap Sampling : Consists of generating distinct data sets by sampling repeatedly from the original data set. The sampling is carried out with replacement.

~~Regression Splines~~

~~17.10~~ 18.10.2016

Non-linear Regression Models

We cover

- Polynomial Regressions
- Step functions
- Splines
- Local Regression
- Generalized Additive Models

Polynomial Regression :

Linear : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Polynomial : $Y_i = \beta_0 + \beta_1 X_i + \beta_i X_i^2 + \cdots + \beta_K X_i^K + \varepsilon_i$

With large K polynomial models can be highly non-linear.

$K \leq 4$ in most cases

Note that we can use polynomials for logistic regression as well.

[Variance of a least square fit :

Let $\hat{C}$ be the estimated variance - covariance matrix of the coefficients $\hat{\beta}_j$.

Let $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$

Let $l_0' = (1, x_0, x_0^2, x_0^3, x_0^4)$

Then $\hat{V}(\hat{f}(x_0)) = l_0' C l_0$]

Non-linear Regression Models 18.10.2016

olynomial Regression (Continued ...)

[To study – Estimation of the var-cov matrix C of $\hat{\beta}$]

Step Functions

Using polynomial functions impose a global structure. In a step function we partition X ~~into a~~ into a set Of contiguous bins & fit a constant for each bin.

In practice we create cutpoints ~~in the~~ ~~$c_1, c_2, \dots$~~ $c_1 < c_2 < \dots < c_K$ in the range of X and then construct $(K+1)$ new variables as follows:

$$f_0 \cancel{f_0}(x) = I(X < c_1)$$

$$f_1(x) = I(c_1 \leq X < c_2)$$

$$\dots$$

$$f_{K-1}(x) = I(c_{K-1} \leq x \leq c_K)$$

$$f_K(x) = I(X \geq c_K)$$

Where $I(\cdot)$ is an indicator function.

We fit the model

~~$Y_i = \beta_0 + \beta_1 c_1 \beta_1 f_1(x)$~~

$$Y_i = \beta_0 + \sum_{i=1}^{K} \beta_i f_i(x)$$

Note: We exclude f_0 to ensure that the functions remain independent

Nonlinear Regression Models 18.10. 2016

Basis Function:

We ~~take~~ use a set of functions $b_1(x), b_2(x)$ $\cdots b_K(x)$ and fit the model

$$y_i = \beta_0 + \sum \beta_i b_i(x) + \varepsilon_i$$

Note that polynomials & step functions are special cases of this general model.

The functions $b_j(\cdot)$ are called the basis functions.

Regression Splines:

Essentially an extension of polynomial regression and piecewise constant regression approaches.

Piecewise Polynomial:

Instead of fitting a high degree polynomial over the entire range of X, we fit separate low degree polynomials (typically polynomials of degree 3) over different regions of X

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

The coefficients $\beta_0, \beta_1, \beta_2$ and β_3 differ in different parts of the range of X

Non-linear Regressions 18.10.2016

The points where the coefficients change are called knots.

A piecewise cubic polynomial with a single knot at the point c takes the form:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \overset{\beta_{31} x_i^3}{\cancel{\beta_{22} x_i^3}} + \varepsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{if } x_i \geqslant c \end{cases}$$

Using more knots leads to a more flexible piecewise polynomial.

Constraints and splines :

We need to add a few constraints.

First, the fitted curves must be continuous everywhere.

Second, both the first and second derivatives must be continuous.

These constraints are imposed to ensure that the fitted polynomial is both ~~conf~~ continuous & smooth.

Non-linear Regression 16.10.2016

Spline Basis Representation

A cubic spline with K ~~#~~ knots can be modeled as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \cancel{\beta_2}\ \beta_2 b_2(x_i) + \dots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

A direct way is to use a basis for cubic polynomial x, x^2, x_3 — and then add a truncated power basis for each knots.

Truncated power fn. is defined as:

$$h(x, \xi) = \begin{cases} (x - \xi)^3 & \text{if } x > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Boundary Constraints :

Splines typically have high variance towards the boundary. Boundary constraints (often linearity constraints) are often imposed to take care of this situation.

A cubic spline with additional boundary constraints (linearity) is referred to as 'natural spline'.

Regression Splines

Piecewise Polynomial : Instead of fitting one high-degree polynomial over the entire range of X, piecewise polynomial regression involves fitting separate low degree polynomials over different ranges of X.

We fit the model $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$ over different regions of X.

Thus, the coefficients $\beta_0, \beta_1, \beta_2$ and β_3 differ in different parts of the range of X. The points where the ~~coe~~ coefficients change are called the knots.

A piecewise cubic polynomial with no knots is just a standard cubic polynomial.

A piecewise cubic polynomial with a single knot at point c takes the form

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{if } x_i \geqslant c \end{cases}$$

Using more knots leads to a more flexible piecewise polynomial.

In general, if we have k knots, we need to fit (k+1) different cubic polynomials.

Regression Splines 24.10.2016

onstraints and Splines : When we fit splines without any constraint, the resulting function is likely to be discontinuous at the knots.

In order to obtain ~~a~~ 'smooth' splines, the following ~~con~~ constraints are added

i) The spline is continuous everywhere (particularly at the knots)

ii) ~~Bo~~ Let $g(x)$ be the spline. Both $g'(x)$ and $g''(x)$ are continuous.

Note: In general a cubic spline ~~uses~~ with K knots uses 4+K degrees of freedom.

In general a degree-d spline requires continuity in derivatives upto degree (d-1)

Spline Basis Representation : Fitting a piecewise polynomial of degree-d appears to be complex in view of imposing the continuity constraints.

However, a cubic spline with K knots may be modelled as:

$$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \cdots + \beta_{K+3} b_{K+3}(X) + \varepsilon$$

Thus a cubic spline with K knots can be modelled in terms of a basis function representation.

Regression Splines 25.10.2016

The most direct way to represent a ~~bas~~ cubic spline with K knots as a basis function is

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=4}^{K+3} \beta_j b_j(X) + \varepsilon$$

Where $b_j(X)$ are truncated power basis function defined as

$$~~b_j(x)~~ \; b_j(x) = (x - \xi_{j-3})^3_+ = \begin{cases} (x-\xi_{j-3})^3 & \text{if } \ldots \\ 0 & \text{otherwise} \end{cases}$$

Note that we may call each $b_j(x)$ as ~~h(x)~~ $h(x, \xi_{j-3})$

where $h(x, \xi_j) = \begin{cases} ~~h~~ (x - \xi_1)^3 \\ 0 \end{cases}$

$\xi_1, \xi_2, \cdots \xi_K$ are the K knots.

It can be shown that in this representation the piecewise polynomial will have a ~~dis c~~ discontinuity only in the third derivative.

This representation simplifies the cubic spline substantially and allows us to fit the model using least squares with an intercept and 3+K predictors of the form $X, X^2, X^3, h(X, \xi_1), h(X, \xi_2) \ldots, h(X, \xi_K)$ where $\xi_1, \xi_2 \cdots \xi_K$ are the ~~K~~ knots.

Regression Splines 25.10.2016

Natural Splines : Whiles splines are flexible and often provides good prediction, they are likely to be unstable at the boundary.

A natural spline is a regression spline with additional boundary constraints. Usually the function is ~~requ~~ required to be linear at the boundary. This additional ~~constraints~~ constraint generally produces more stable estimates.

How many knots and where : More knots increase the model flexibility. ~~&~~ We may wish to place more knots where the response is likely to have more variation with respect to the explanatory varible.

However, in practice knots are placed at fixed ~~percentiles~~ percentiles - may be 25th, 50th and 75th.

(Read again about cubic splines and natural splines and their degrees of freedom)

The 'best' number of knots may be determined using the technique of cross-validation.

Regression Splines

25.10.2016

Smoothing Splines

[Concepts of RSS, MSE, likelihood and ~~their~~ their usage in inference / analytics]

An alternative to knots is the usage of a tuning parameter. In this approach we find the function g such that

$$\sum_{i=1}^{n}\left(y_i - g(x_i)\right)^2 + \lambda\int\left(g''(t)\right)^2 dt$$ is minimized.

$\lambda\ (\geq 0)$ is called a 'tuning parameter'.

Note 1: The first derivative measures the slope ~~and~~ and the second derivative measures the change of slope. Hence, roughly speaking, the 2nd derivative ~~is~~ of a function is measure of its roughness.

Note 2: The function $g(x)$ obtained through the 'loss + penalty' approach can be shown to be a piecewise continuous polynomial with knots at the unique values of $x_1, x_2 \cdots x_n$. In addition, it is linear in the region outside the extreme knots. Then, it is a natural spline but not the same spline obtained though the piecewise linear approach.

Regression Splines

25.10.2016

[Concepts of degrees of freedom, shrinking and tuning parameters, bias-variance tradeoff, parametric and non-parametric, sufficiency]

Choosing λ : Let $\hat{g}_\lambda = S_\lambda y$ [Similar to $X\beta = X(X'X)^{-1}y$]

~~The~~ $df_\lambda = tr(S_\lambda)$

It is easy to see that as λ increases from 0 to ∞ df_λ decreases from n to 2.

Note: In the spline smoothing approach we minimize:

$$\sum (y_i - g(x_i))^2 + \lambda \int (g''(t))^2 dt$$

In a spline basis (cubic spline) representation, we use

~~$y_i = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \sum_{j=4}^{K+3} b_j$~~

~~y_i~~ $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=4}^{K+3} \beta_j b_j(x) + \varepsilon$

where $b_1(x) = x$, $b_2(x) = x^2$, $b_3(x) = x^3$

$$b_j(x) = \begin{cases} (x - \xi_{j-3})^3 & \text{for } j = 4, 5, \dots K+3 \\ 0 & \text{otherwise} \end{cases}$$

Regression Splines 31.10.2016

Suppose we have fitted a smoothing spline

Let $\hat{g}_\lambda = S_\lambda \cdot Y$

Then the effective degrees of freedom is defined as

$$df_\lambda = \sum_{i=1}^{n} S_{\lambda ii} \rightarrow \text{Trace of matrix } S_\lambda$$

Analytics → Shrinkage Mathods

31.10.2016

Ridge Regression :

Suppose we are fitting the model $Y = \beta_0 + \sum \beta_i x_i + \varepsilon$

We may estimate the coefficients using OLS that minimizes

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

In the ridge regression we take the "loss + penalty" approach and minimize

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \text{—— Ⓐ}$$

– where $\lambda \geq 0$ is the tuning parameter (shrinkage ~~param~~ parameter)

An equivalent way to write the ridge problem is

$$\left. \begin{array}{l} \min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{Subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \end{array} \right\} \text{—— Ⓑ}$$

There is a one-to-one correspondence between the tuning parameter λ defined in Ⓐ and t in Ⓑ.

Formulation Ⓑ is sometimes preferred as it makes the size constraint explicit.

Analytics - Shrinkage Methods 31. 10. 2016

sage of Ridge Regression :

Ridge is often used when there are many correlated variables. When the explanatory variables are highly correlated, the coefficients can become poorly determined and exhibit high variance. A pair of correlated variables often have large +ve and -ve coefficients, cancelling each other. The size constraint / tuning parameter alleviates this problem.

Notes :

a) The ridge solutions are not equivariant under scaling of the inputs. Thus x_{ij}'s are usually standardized before solviry the ridge equation.

b) The intercept has not been tuned. If it is tuned then the procedure would depend on the origin of Y. That is adding a constant to Y would not result in simply adding the same constant to the predictions.

c) When we center the inputs, i.e. use $x_{ij} - \bar{x}_j$ instead of x_{ij}. We estimate β_0 by $\bar{y} = \frac{1}{n} \Sigma y_i$. The remaining coefficients get estimated by a ridge regression without intercept.

Analytics — Shrinkaye Methods 31. 10. 2016

$$RSS(\lambda) = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$ (X has p columns)

$$\frac{\partial R(\beta)}{\partial\beta} = \frac{\partial}{\partial\beta}\left((Y' - \beta'X')(Y - X\beta) + \lambda\beta'\beta\right)$$

$$= \frac{\partial}{\partial\beta}\left(Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta + \lambda\beta'\beta\right)$$

$$= -X'Y - \overset{X'Y}{\cancel{X'X}} + 2\beta X'X + 2\lambda\beta$$

$$\Rightarrow 2X'Y = 2(X'X + \lambda I)\beta$$

$$\Rightarrow \hat{\beta} = (X'X + \lambda I)^{-1}X'Y$$ ——— ⓒ

Traditional ~~definition of~~ description of ridge regression start with ⓒ.

Note : The choice of quadratic penalty adds a +ve constant to the diagonal terms of X'X. This makes the problem nonsingular even if X'X is not of full rank (why?)

Note further that the ridge solution ~~to~~ remains a linear function of y.

Analytics – Shrinkage Methods 31.10.2016

Singular Value Decomposition (SVD):

The SVD of the centered input matrix X gives us some additional insights into ridge regression.

$$\underset{N\times p}{X} = \underset{N\times p}{U}\,\underset{p\times p}{D}\,\underset{p\times p}{V'}$$ — U & V are ~~otho~~ orthogonal matrices

Note that the ~~SVD may b~~ least square solution may be writen as:

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\Rightarrow X\hat{\beta} = X(X'X)^{-1}X'y$$

Using SVD, ~~w~~ we get

$$X\hat{\beta} = \underset{N\times p}{U}\,\underset{p\times N}{U'}\,\underset{N\times 1}{y}$$

Note that U'y projects y into a p dimensional space

$X = UDV'$

$$\underset{N\times p}{U}$$

$$\underset{p\times N}{U'}\ \underset{N\times p}{U} \rightarrow U'U$$

$$X(X'X)^{-1}X'y$$

$$= UDV'(VD\underset{\sim}{U'}UDV')^{-1}\ VDU'\ y$$

$$= UDV'(VDDV')^{-1}\ VDU'\ y$$

$$= (D^2)^{-1}\ UDV'VDU'\ y$$

$$= UU'y$$

——x——

<u>Ridge solutions</u>

$$X\hat{\beta}_R = X(X'X + \lambda I)^{-1}\ X'y$$

Analytics - Shrinkage Methods 1.11.2016

dge Regression :

Note: The standard least square coefficients are scale independent, i.e. multiplying x_j by c leads to scaling of β_j by $1/c$. Thus $x_j \hat{\beta}_j$ remains the same, no matter what unit is used to measure x_j

In contrast ridge regression coefficient estimates change substantially due to change of scale. This is due to the sum of square of coeffs. constraint. Thus we apply standardization of the predictors using

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

By virtue of the standardization, each predictor will have unit standard deviation.

Rationale behind improvement :

Primarily due to bias-variance trade off. As λ increases, the flexibility decreases leading to increasing bias but decreasing variance. We look at the MSE of the test data to choose the 'right' value of λ.

Analytics – Shrinkage Methods 1.11.2016

Lasso :

A significant difficulty with Ridge is its inability to select a subset of variables. The penalty $\lambda \Sigma \beta_j^2$ or the constraint $\Sigma \beta_j^2 \leq t$ shrinks all coefficients to 0 but does not set any one of them to 0 unless $\lambda = \infty$

In order to select a subset of variables we ~~use the~~ minimize

$$Q = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Thus lasso uses an L_1 penalty rather than L_2 penalty.

The lasso may be alternatively formulated as

$$\text{Minimize} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \right.$$

$$\text{Subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

1 — Shrinkage Methods   Analytics Notes   6.11.2016

Introduction : Selection of subsets of variables in a regression context is a widely used technique. This usually produces a more interpretable model that possibly has a lower prediction error. However, this is a discrete process — variables are either retained or discarded. This process tends to have a high variance. Best subset may lead to different subsets on cross validation. In contrast shrinkage methods are more ~~continous~~ continuous and have lower variance.

Ridge Regression : Attempts to shrink the coefficients by imposing a penalty on their size. Two alternative formulations of ridge regression are:

$$\hat{\beta} = \min\left[\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right] \text{ —— } (A)$$

Here, λ is the shrinkage parameter, $\lambda \geqslant 0$

When $\lambda = 0$, ridge regression reduces to OLS

As $\lambda\uparrow$, the shrinkage becomes greater. The model becomes a null model when $\lambda \to \infty$

An equivalent formulation of ridge regression is

$$\left.\begin{aligned}&\text{Min}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\\ &\text{subject to } \sum\beta_j^2 \leq t\end{aligned}\right\} \text{ —— } (B)$$

The 2nd formulation makes the size constraints on the coefficients directly visible.
The two formulations are equivalent to each other.

Non-equivariance : The ridge solutions are not equivariant under change of scales.

~~Note~~ Choice of λ : In ridge regression the tuning (shrinkage) parameter plays a crucial role ~~a~~ as stated earlier.
~~When~~ Note that unlike least square regression, ridge regression will produce a different set of coefficient estimates for each value of λ.

Impact of scale : The least square coefficient estimates are scale ~~invarian~~ invariant. Thus multiplying X_j by c simply leads to multiplying β_j by $1/c$. Hence $\hat{\beta}_j X_j$ remains the same irrespective of scale
However, in ridge regression, change of scale would impact the estimated coefficient of the predictor and may even impact other predictors due to the sum of square constraint.
Thus, in ridge regression, the predictors are standardized using the formula :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\Sigma(x_{ij}-\bar{x}_j)^2}}$$

Shrinkage Methods Analytics Notes 7.11.2016

Centering of variables : ~~Not~~ The shrinkage penalty is applied ~~in~~ to the coefficients $\beta_1, \beta_2, \cdots, \beta_p$ but not to intercept β_0. This is because β_0 is simply a measure of the mean value of the response when ~~the~~ the predictors are 0.

Thus when the predictors are centered to have mean 0, i.e. when the predictors are transformed as $x_{ij}' = x_{ij} - \bar{x}_j$, the estimated intercept takes the form $\beta_0 = \bar{y} = \frac{1}{n}\Sigma y_i$

Parameter Estimate in Ridge :

~~x~~ Minimize $\Sigma(y_i - \beta_0 - \Sigma \beta_j x_{ij})^2 + \lambda \Sigma \beta_j^2$

$\Rightarrow$ ~~$RSS(\beta) = (Y - X\beta')$~~

$$\Rightarrow RSS(\beta) = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

$$= (Y' - \beta'X')(Y - X\beta) + \lambda\beta'\beta$$

$$= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta + \lambda\beta'\beta$$

$$\Rightarrow \frac{\partial R(\beta)}{\partial\beta} = -2X'Y + 2\beta X'X + 2\lambda\beta$$

$$\Rightarrow (X'X + \lambda I)\beta = X'Y$$

$$\Rightarrow \cancel{\hat{\beta}}\ \hat{\beta} = (X'X + \lambda I)^{-1}X'Y \text{ ——————— } \textcircled{c}$$

Note that the traditional definition of ridge regression starts with ⓒ.

It is easy to note that the ridge coefficients can be estimated using the least squares methodology.

Notice further that the choice of quadratic penalty adds a +ve constant to the diagonal elements of $X'X$. This forces a solution in all cases. Further, the ridge solution is a linear ~~f~~ function of y.

Why does Ridge Regression improve over least square?

Essentially due to bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases. At the same time, as λ increases, the shrinkage of a ridge coefficient leads to substantial reduction of variance at the cost of a small increase in bias.

Decision Trees

26. 9. 2017

Content

1 Introduction – What is a decision tree? How is it used?

2 Introduction to basic terminologies – nodes, branches, terminal (leaf) and non-terminal nodes

~~3 How does a tree work~~?

3 Phases of tree construction

4 Estimation and classification rules

5 Growing the tree

6 Concept of node purity and its measurement

7 Parameters controlling the splitting process

8 Pruning of tree. Concepts of cost-complexity pruning

9 Variable Importance Measures

10 Concepts of bagging, boosting and random forests

11 Advantage and disadvantage of trees

Decision Trees

26.9.2017

Introduction : Decision tree is a non-parametric supervised learning technique that can be used for both value estimation as well as classification problem. In this technique the form of the estimator is not pre-specified and consequently it is called a non-parametric technique.

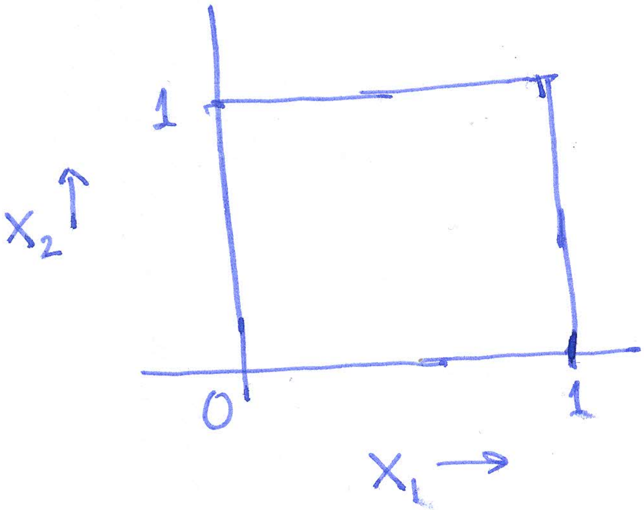
As the decision tree is a supervised learning technique, the data are collected in the following format

Y	X_1	X_2	$\cdots$	X_p
y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdot$
y_N	x_{N1}	x_{N2}	$\cdots$	x_{Np}

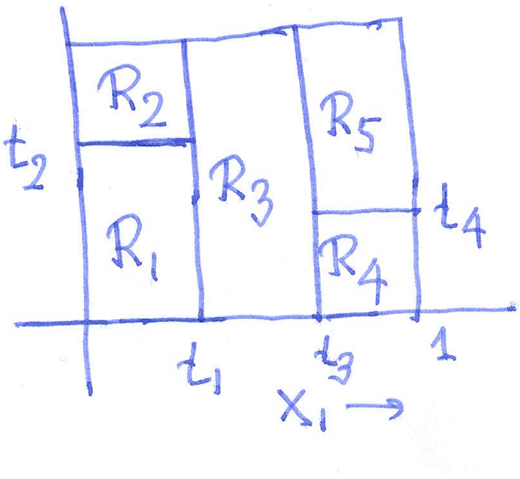
Here Y is the response variable and $X_1, X_2, \cdots X_p$ are the explanatory variables. Each row of the matrix except the header gives the observed values of Y and $X_1, X_2 \cdots X_p$. We assume that N observations were collected.

The decision tree algorithm divides the feature space (i.e. the theoretical space covered by the explanatory variables) into a number of mutually exclusive and non-overlapping regions. Such a region is called a partition of the feature space. The decision tree gives the same estimate for each of the regions.

Example 1 Suppose we are trying to estimate the value of a response variable Y given two explanatory variables $0 \leq X_1, X_2 \leq 1$. Notice that the feature space may be drawn as follows:



We may divide the feature space into 5 regions as follows:

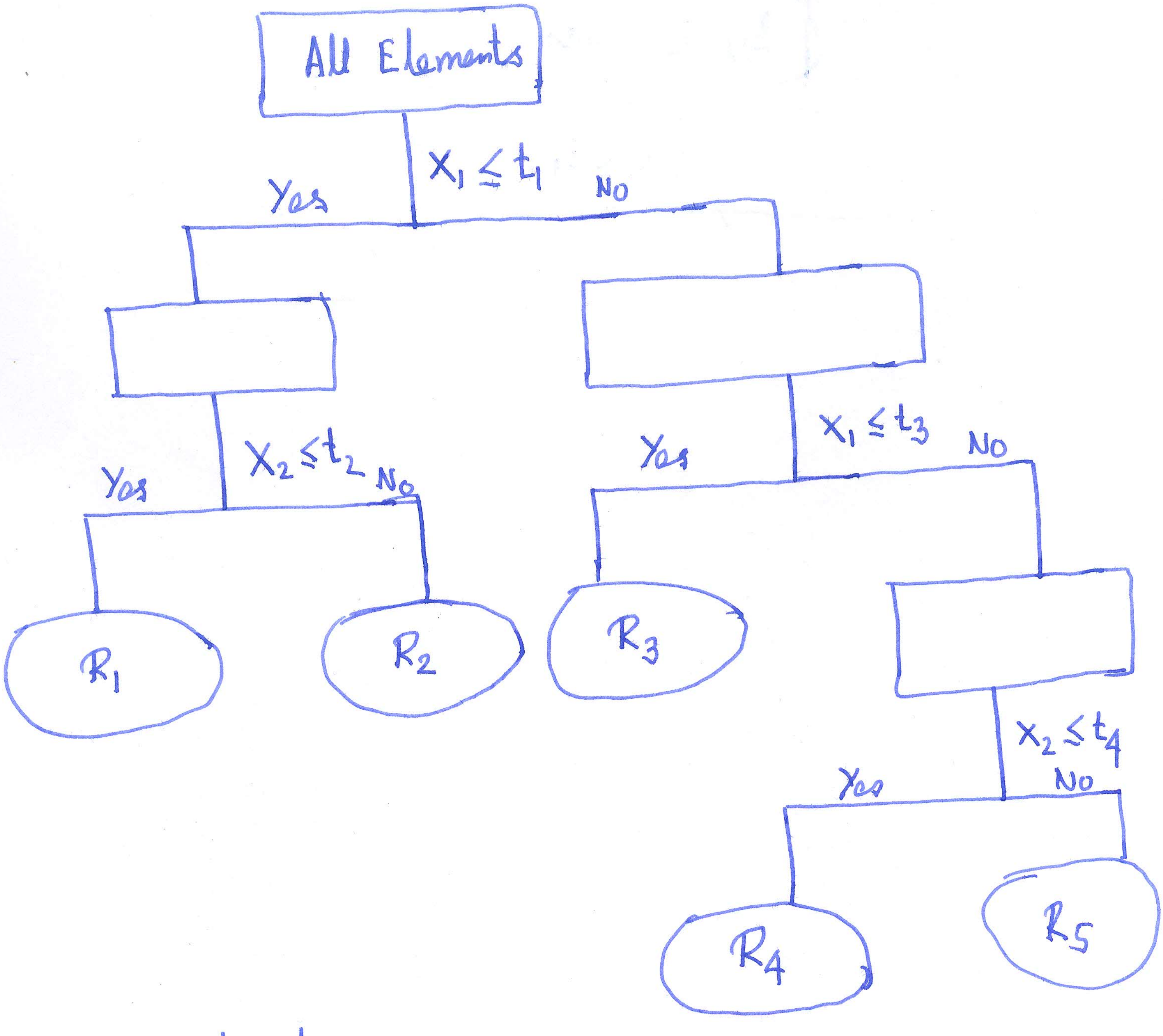


Note that $\{R_1, R_2, R_3, R_4, R_5\}$ constitute a partition of the feature space ~~$S=\{(x_1, x_2$~~

$S=\{(x_1, x_2) / 0 \leq x_1, x_2 \leq 1\}$

Decision Trees

The partition may be arrived at as follows:

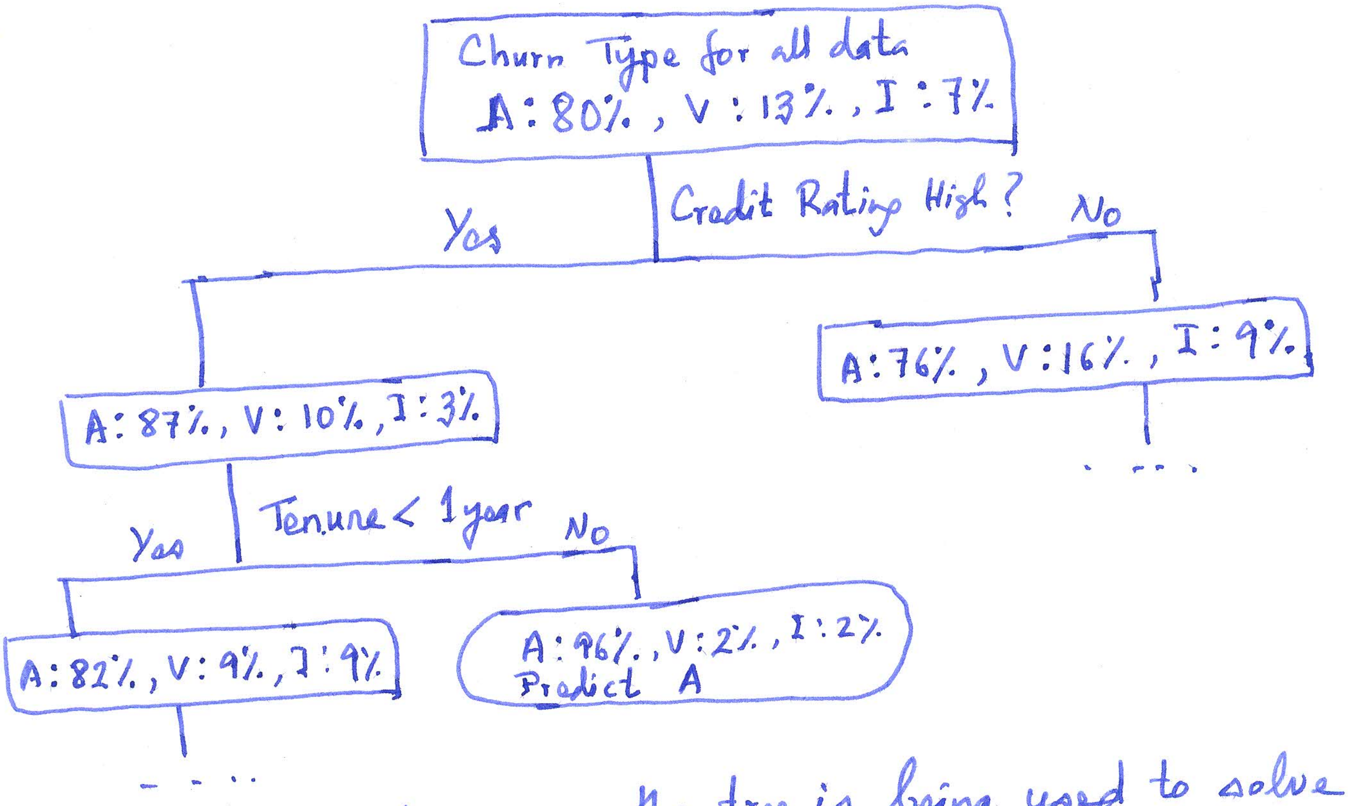


~~Notice that at~~

Notice that at each step the feature space is ~~being~~ being divided into two groups. This method of splitting the entire data set into a number of smaller sets, each time dividing the larger set into two is called binary splitting. The resulting structure is called ~~a~~

a binary tree.

Example 2 : A mobile telephone service provider wants to develop a model to predict behaviour of postpaid subscribers. The behaviour has three possible values → the customer remains active, the customer churns ~~volunta~~ voluntarily (by informing the service provider) and the customer churns involuntarily (stopped paying without information). The target (response) variable, therefore, takes three nominal values denoted by A (active), V (voluntary churn) and I (involuntary churn). A decision tree for predicting behaviour may be constructed as follows:


Churn Type for all data
A : 80% , V : 13% , I : 7%
Credit Rating High?
Yes
No
A : 87% , V : 10% , I : 3%
A : 76% , V : 16% , I : 9%
. . .
Tenure < 1 year
Yes
No
A : 82% , V : 9% , I : 9%
A : 96% , V : 2% , I : 2%
Predict A
. . .

Notice that in this case the tree is being used to solve a classification problem.

Note : The tree given above is partial and illustrative.

Decision Trees

26. 9. 2017

Notations and basic definitions :

The tree consists of nodes and branches. The rectangles (▭) and ovals (⬭) are the nodes. The nodes are connected by branches. The first node containing all elements is called the root ~~nodes~~ node. The nodes not split further are called the terminal or leaf nodes. The leaf nodes contain the decision. We will discuss two types of trees. The trees used for value estimation problems are called 'Regression Trees' and trees used for classification problems are called ~~'Ca~~ 'Classification Trees'. Together the trees are called Classification And Regression Trees (CART).

Three phases of tree construction

Three steps taken are :

a) Selection of splits in the non-terminal nodes

b) Deciding whether to make a node a terminal node or not (i.e. whether a node is to be split further)

c) Assigning rules for estimation or classification at the ~~terminal~~ terminal node

Decision Trees

26. 9. 2017

Note: The process of breaking a node into two subnodes is called splitting. Notice that every node consists of a set of points $(y, x_1, x_2 \cdots x_p)$. Whenever a node is split, this set is broken into two subsets according to some rule. The rules related to the construction of subsets use one variable at a time.

Thus at any stage the splitting rule may ~~be x~~ be $x_j \leq t$ or $x_j \geq t$ in case x_j is measured in ordinal, interval or ratio scale. When x_j is measured in nominal scale ~~the~~ with K values (say 1, 2, … k), the splitting rule would be of the form $x_j = 1$ or 2 … ~~(i.e. x_j will either take any one of~~ (i.e. x_j will either belong to a subset of $\{1, 2, \cdots k\}$ or not).

NOTE THAT THE SPLITTING RULE MUST DIVIDE THE PARENT NODE (THE NODE BEING SPLIT) INTO EXACTLY TWO SUBSETS. RULES LIKE $a \leq x_j \leq b$ THAT LEADS TO A THREE-WAY SPLIT IS NOT ALLOWED.

~~Estimation and Classification Rules:~~

8

Decision Trees

26.9.2017

Estimation and Classification Rules :

Suppose the feature space has been divided into J mutually exclusive and non-overlapping region.

Let $R_1, R_2, \cdots, R_J$ be the identified regions.

For any region R_j, $j \in \{1, 2, \cdots, J\}$ the ~~value~~ estimated value is $\hat{c}_j = \text{Avg}(y_i / \underset{\sim}{x}_i \in R_j)$. Thus for each region we estimate y through the average y in that region.

For classification problem with the response having K different values 1, 2, ⋯ K with frequencies as $f_{j1}, f_{j2} \cdots f_{jK}$ for the jth class, we allocate the response to class l such that $f_{jl} = \max\{f_{j1}, f_{j2} \cdots f_{jK}\}$.

Thus classification is carried out by identifying the particular value with maximum frequency in every node.

Estimation and Classification Model :

Let $\{R_1, R_2, \cdots R_J\}$ be a partition of the feature space.

Suppose we have p predictor variables $X_1, X_2, \cdots X_p$

Let $\hat{C}_m = \text{Avg}\{y_i / \underset{\sim}{x}_i \in R_m\}$, $m = 1, 2, \cdots J$

Then $\hat{f}(x) = \sum_{m=1}^{J} \hat{C}_m . I((x_1, x_2, \cdots, x_p) \in R_m)$ is used for the ~~purpose~~ purpose of value estimation.

Decision Trees

26.9.2017

Note that $I((x_1, x_2, \dots, x_p) \in R_m)$ is the indicator ~~function. t~~ function defined as follows:

$$I((x_1, x_2 \cdots x_p) \in R_m) = \begin{cases} 1 \text{ if } (x_1, x_2 \cdots x_p) \in R_m \\ 0 \quad \text{otherwise} \end{cases}$$

As $\{R_1, R_2, \dots, R_J\}$ is a partition of the feature space, $\hat{f}(\underset{\sim}{x})$ is essentially the average y for the region R_m such that $\underset{\sim}{x} \in R_m$.

~~Exercise~~

We can write the model similarly for a classification problem.

Let the response variable Y have ~~p~~ K classes $1, 2, \dots K$.

Suppose we are deciding about the classification rule for region m.

Let $f_{m1}, f_{m2}, \cdots f_{mK}$ be the ~~values~~ frequencies of the values $1, 2, \dots K$.

Let $\hat{C}_m = \{ p \,/\, f_{mp} = \max\{f_{m1}, f_{m2}, \cdots f_{mK}\}, p = 1, 2 \cdots K \}$

Then $\hat{f}(\underset{\sim}{x}) = \sum_{m=1}^{J} \hat{C}_m \cdot I((x_1, x_2 \cdots x_p) \in R_m)$

<u>Note</u>: It is interesting to note that $\hat{f}(\underset{\sim}{x})$ is a ~~con~~ conditional average or a conditional proportion. We find the conditional average or proportion for each region.

2.10.2017

Growing the tree :

The tree is grown starting from the root node that initially contains the entire training data. A top-down greedy approach known as recursive binary splitting is used to grow the tree. The approach is top down because it begins at the top of the tree with all observations belong to a single region represented by the root node. Subsequently the predictor space is subsequently split into subregions.

During every split one node ~~is split~~ containing a subset of the training observations is split into two ~~subnodes~~ subnodes provided some conditions are met. The ~~st~~ process of splitting stops when none of the terminal nodes satisfy the criteria for further splitting. At that point the tree is said to be 'fully grown'.

The approach is said to be greedy because at each step of the tree building process the best split ~~is~~ with respect to the splitting criteria for that step is chosen rather than looking ahead and choosing a split that will lead to a better tree in a future step.

3.10. 2017

Splitting Criteria : Different criteria are used for regression and classification trees. The criteria are given below:

Splitting Criteria for Regression Trees : In regression trees attempts are made to minimize RSS. The criteria may be explained as follows:

Suppose we are attempting to split the region R that is a subset of the feature space.

Let $\bar{y}_R$ be the observed average of y_i when $\underset{\sim}{x}_i \in R$

Then $RSS_R = \sum_{i=1}^{|R|} (y_i - \bar{y}_R)^2$ – where $|R|$, called the cardinality of R, gives the number of elements in the set R.

Let $RSS_R = D_0$ (The RSS is also referred to as deviance - a term we shall be using).

~~Note : The RSS measures the node impurity. If the y_i's in a given node are close to each other, RSS would be low, the smallest possible value being 0. This is often referred to as regression deviance as well.~~

Note 1 : The RSS is also referred to as the regression regression deviance. Notice that the regression deviance measures the node impurity (disorder) and assesses the homogeneity of the responses within the node. When the y_i's in a node are close to each other, RSS is low with the smallest possible value being 0.

3.10.2017

Note 2 : The deviance of a regression tree T is obtained by adding the deviances of its leaf nodes. Thus $D_T = \sum_{i=1}^{J} D_{\text{Leaf Node } i}$ – where J gives the number of leaf nodes.

✗ We now discuss the splitting method.

Let $X_1, X_2 \cdots X_p$ be the predictor variables.

Note that the region R we are trying to split is a subset of the predictor space (feature space).

Note further that the deviance of the region R is given by D_0.

In order to split R we select a predictor $X_j, j=1,2\cdots p$ and a cutpoint s. The region R is split into two subregions

$$R_1 = \{ (x_1, x_2 \cdots x_p) \in R \,/\, x_j \le s \}$$ and

$$R_2 = \{ (x_1, x_2 \cdots x_p) \in R \,/\, x_j > s \}$$

We select the predictor X_j and the cutpoint s such that

$$D(j,s) = \sum_{i:\, x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2$$

is minimum.

3.10.2017

Note that $D(j, s) \leq D_0$ and $(D_0 - D(j,s))$ gives the improvement measured in terms of reduction of RSS or equivalently regression deviance.

Note: When X_j is a ratio, interval or ordinal scale variable, we define two subsets as ~~$X_j \leq s$~~ $X_j \leq s$ ~~x~~ and $X_j > s$. However, this method of binary split does not work when X_j is a nominal variable.

Suppose X_j is a nominal variable with three values, say A, B and C. Notice that there are three splits namely $\{A\}$ vs. $\{B, C\}$, $\{B\}$ vs. $\{A, C\}$ and $\{C\}$ vs $\{A, B\}$. In case X_j has 4 values, namely A, B, C and D, there are 7 splits $\{A\}$ vs $\{B, C, D\}$; $\{B\}$ vs $\{A, C, D\}$; $\{C\}$ vs $\{A, B, D\}$; $\{D\}$ vs $\{A, B, C\}$; $\{A, B\}$ vs. $\{C, D\}$; $\{A, C\}$ vs. $\{B, D\}$ and $\{A, D\}$ vs. $\{B, C\}$. In general for nominal variables with K possible values one has to evaluate $2^{K-1} - 1$ ~~of~~ binary splits.

Notice that the above splitting logic may be applied to ordinal variables as well. However, that should be avoided as ordinal variables have an implicit ordering.

3.10. 2017

Node Splitting in Classification Setting :

In the classification setting the ~~class~~ proportions of the different values of the response variable within a given node are used for the purpose of prediction. In the classification setting different methods may be used to split the nodes. The simplest and intuitively appealing method is the classification error rate.

Classification Error Rate : Suppose we are trying to split the m-th region R_m.

Suppose the response variable Y takes p different values and $\hat{p}_{mk}$, $k=1, 2, \cdots p$ be the estimated proportions of the values of the response variable Y.

Then the classification error rate of the ~~response~~ region m, denoted by E_m would be :

$$E_m = 1 - \max_{1 \le k \le p} \hat{p}_{mk}$$

Clearly for $\underset{\sim}{x} \in R_m$ is classification is made to the largest occurring value.

Note : The classification error rate is not sufficiently sensitive for tree growing and in practice two other measures are preferable.

3.10.2017

Concept of node purity : A node is perfectly pure if it has only one value. Notice that in the regression setting the deviance is 0 for a perfectly pure node. In the classification setting, a perfectly pure node leads to an error rate of 0.

Both regression deviance and classification error rate are measures of node purity. However, as noted in the previous section, classification error rate is not sufficiently sensitive to change of level of node purity.

Cross Entropy : This is a measure of node purity.

Cross entropy of ~~a~~ node m is given by

$$D = -\sum_{k=1}^{p} \hat{p}_{mk} \log \hat{p}_{mk}$$

It is easily observed that the cross-entropy takes a small value when $\hat{p}_{mk}$'s are near zero or one.

Gini Index : This is another ~~meas~~ measure of node purity and is defined as

$$G = \sum_{k=1}^{p} \hat{p}_{mk}(1-\hat{p}_{mk})$$

3.10.2017

Notice that Gini Index and Cross Entropy are similar measures and result in very similar ~~clas~~ classification trees.

Note: The approach to the growing of trees is as follows:

Step 1: We start with the ~~x~~ root node containing all elements. We compute the node purity D_0. This is computed using RSS / Regression Deviance approach for regression (value estimation) setting. For classification setting we use either Gini Index or Cross Entropy. Suppose the initial value of node purity is D_0. This step may be called the initialization step.

Step 2 (Processing stage) Look at the leaf nodes one by one (initially there is only one leaf node). Suppose the deviance of the chosen leaf node is D_j. Find the best split and suppose the corresponding deviance (or cross entropy) is D_{j+1}. Decide whether the node with deviance D_j needs to be split.

Step 3 (Termination stage) Stop when none of the terminal nodes can be split any further. At this stage the tree is fully grown.

3.10.2017

Parameters that define splitting of feature space:

Three parameters are used to ~~define~~ decide whether a node needs to be split further. These are:

a) Requirement of minimum leaf size (i.e. number of observations in a leaf)

b) Minimum improvement in node impurity (i.e. if D_j is the current level of impurity and the best split yields an ~~impu~~ impurity of D_{j+1}, then $D_j - D_{j+1}$ must exceed a predetermined threshold)

~~c) Maximum depth of the tree (nu i.e. number of~~

c) Maximum depth of the tree (i.e. the number of steps required to traverse from the root node to the leaf node along the longest path)

Pruning of tree: A fully grown tree restricted only by the three parameters mentioned above is likely to overfit the data. A smaller tree with fewer splits might lead to lower variance and better interpretation at the cost of a little bias.

3.10.2017

One possible strategy of getting such a tree is to grow a very large tree T_0 and then prune it back in order to get a subtree. A method of pruning a very large tree to a smaller tree is the ~~cost complex~~ 'cost-complexity pruning.

Cost Complexity Pruning: Suppose the fully grown tree is T_0.

We find a subtree $T \subset T_0$ such that

$$D = \sum_{m=1}^{|T|} D_m + \alpha |T| \text{ is minimum.}$$

Here $|T|$ gives the number of terminal nodes of the tree T.

The parameter α (≥ 0) is the tuning parameter and it controls the trade-off between accuracy and complexity. It is easy to note that the tuning parameter is actually a penalty for increased complexity.

Notice that $|T|$ gives the complexity of a tree.

Trees are pruned from the leaf node. As a leaf node is pruned, D decreases by α.

However, every time a leaf node is pruned, $\sum D_m$ increases as the impurity of the parent

3.10.2017

node is necessarily higher than the child node. Thus pruning a leaf node is carried out only when the parent to child difference of impurity exceeds α.

When $\alpha = 0$, there is no pruning at all.

Determination of α : The value of α is determined using cross-validation. The error rate needs to be estimated for different values of α using K-fold ~~cro~~ cross validation.

For value estimation problems, the sum of squared errors for the test data is found.

For classification problem, the rate of classification error may be computed for the test data.

Variable Importance Measures : Let the initial node impurity be D_0. Note that this is the level of impurity (deviance or cross-entropy) for the entire training data.

Suppose, there were n splits with node impurities $D_1, D_2, \cdots, D_n$. The gains (reduction of impurity) are $D_0 - D_1, D_1 - D_2, \cdots D_{n-1} - D_n$.

Note that $D_0 - D_1 + D_1 - D_2 + \cdots + D_{n-1} - D_n = D_0 - D_n$ or the overall gain.

Thus $$\sum_{j=0}^{n-1} \frac{D_j - D_{j+1}}{D_0 - D_n} = 1$$

The importance of the $(j+1)$th split, $j=0,1,2\ldots(n-1)$ is given by $\dfrac{D_j - D_{j+1}}{D_0 - D_n}$

As every split is defined by one variable, this measure gives us ~~the~~ a measure of variable ~~import~~ importance.

Concepts of bagging, random forest and boosting :

The techniques of bagging, random forests and boosting use trees as building blocks to construct more powerful prediction models.

Bagging : The decision trees introduced in the previous chapters suffer from high variance. Thus, if we split the training data into two parts at random and fit a decision tree to both halves, the resulting trees may be quite different. In contrast, a procedure with low variance will yield similar results if applied repeatedly to distinct data sets.

Linear regression tend to have low variance when the ratio of n (the sample size) to p (the number of variables) is moderately large.

4.10.2017

Comments on bagging : The trees used in bagging are grown deep and are not pruned. Thus individual trees have high variance and low bias. Averaging these trees reduces variance.

For bagging we often we often average hundreds or even thousands of trees.

Usage of bagged trees :

Regression trees : Average of the predicted values is taken as the final estimate.

Classification trees : Majority vote, i.e. the most commonly occurring class among the B predictions is taken as the predicted class.

Error estimation : ~~A method called ou~~ The bagged trees use some portion of the data. The observations not used to fit a given bagged tree are referred to as out-of-bag observations.

~~An~~ A method to estimate error of a bagged tree is to estimate error for each of the Out Of Bag (OOB) observation.

4.10.2017

Variable Importance Measure for bagged trees :

We have already discussed about a way to measure importance of variables in a quantitative manner. Note that in the case of a single tree the splitting rules are visible. Further, we are aware that the earlier splits typically contribute to larger ~~redu~~ improvement of node purity and hence the variables involved in earlier splits are likely to be more important. Thus, in the case of a single tree, quantitative estimates of variable importance may not be of much practical value.

The measure of variable importance, however, is very ~~import~~ important for bagged trees. As bagged trees are average of many trees, the importance of individual variables would not be easily known.

Method : The method adoped is as follows :

Step 1 : For each of the B trees find the decrease of deviance / cross entropy / Gini Index. Suppose the decrease for variable i in tree j is D_{ij}, $i=1,2\cdots,p$ and $j=1,2\cdots B$ where p is the number of variables and B the number of trees fitted.

4.10.2017

Step 2 Find $\bar{B}_i = \frac{1}{B}\sum_{j=1}^{B} B_{ij}$, $i = 1, 2 \cdots p$

Step 3 Find variable importance V_i as

$$V_i = \frac{\bar{B}_i}{\sum \bar{B}_i}$$

Random Forests : Random forests ~~imp.~~ provide an improvement over bagged trees by decorrelating the trees.

In random forests, a random sample of m predictors is chosen from the entire collection of p predictors.

Usually $m = \sqrt{p}$

The random forests often improve performance as selecting a small sample of predictors help decorrelating variables. In the case of bagging a few important variables appear on top of all the bagged trees. Thus the bagged trees are likely to be correlated.

This situation is avoided in random forests as only a few variables are used for the purpose of building trees in any given iteration.

25 4.10.2017

Boosting : Boosting is a slow learning method where small trees are fitted on modified versions of the original data set. The algorithm for boosting is given below:

Let $\hat{f}(\underset{\sim}{x})$ be the estimated value of y given $\underset{\sim}{x}$.

[Note: We discuss only the value estimation problem as boosting in classification setting is rather complex]

Algorithm :

~~Step 1 $\hat{f}(\underset{\sim}{x}) \leftarrow 0$, $r_i \leftarrow y_i$ for all i in the training~~

Algorithm :

Step 1 (Initialization): $\hat{f}(\underset{\sim}{x}) \leftarrow 0$; $r_i \leftarrow y_i$ for all i in the training set. [Note that $(y_i, \underset{\sim}{x}_i)$ gives the ith row of the data matrix and $\underset{\sim}{x}_i = (x_{i1}, x_{i2} \cdots x_{ip})$]

Step 2 (Processing): For $b = 1, 2 \ldots B$ repeat

a) Fit a tree with d splits ($d+1$ terminal nodes) to the training data $(r, \underset{\sim}{X})$. Let the ~~tree~~ fitted tree be $\hat{f}^b(\underset{\sim}{x})$

b) Update $\hat{f}(\underset{\sim}{x})$ as $\hat{f}(\underset{\sim}{x}) \leftarrow \hat{f}(\underset{\sim}{x}) + \lambda \hat{f}^b(\underset{\sim}{x})$.

λ is called the shrinking parameter

c) Update residuals, $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$

4.10.2017

Step 3 Output the boosted model

$$\hat{f}(\underset{\sim}{x}) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(\underset{\sim}{x})$$

Parameters of boosted model : Boosting has 3 tuning parameters as given below

1 Number of trees B : Unlike bagging and random forest boosting can overfit if B is too large. We use cross-validation to select B

2 Shrinking parameter λ : Small positive numbers need to be taken. Often values like 0·01 or 0·001 are chosen. Smaller λ requires larger B

3 Number of splits in each tree, d : Shallow trees are chosen. Often d=1 works well. In this case each tree is a stump consisting of a single split. Occassionally d=2 is chosen.

Advantages and Disadvantages of Trees : Decision trees used for the purpose of value estimation (regression) and classification have a number of advantages over classical models like regression:

4.10.2017

Advantages of trees

a) Trees are very easy to explain — often easier than linear regression

b) It is believed by many that trees closely resemble process of human decision making

c) Trees can be displayed graphically

d) Trees can easily handle qualitative predictors

Disadvantages

a) Fully grown trees often overfit the data

b) Pruned trees may not have much predictive power

c) Bagging, random forests and boosting improve predictive accuracy but interpretation may not be easy

30.10.2017

Projection Pursuit Regression

We fit a regression of the form

$$f(\underset{\sim}{X}) = \sum_{m=1}^{M} g_m(\omega_m' \underset{\sim}{X})$$

Essentially we are fitting an additive model on the derived features $V_m = \omega_m' \underset{\sim}{X}$

The ω_m's are unit vectors and may, therefore, be treated as direction vectors.

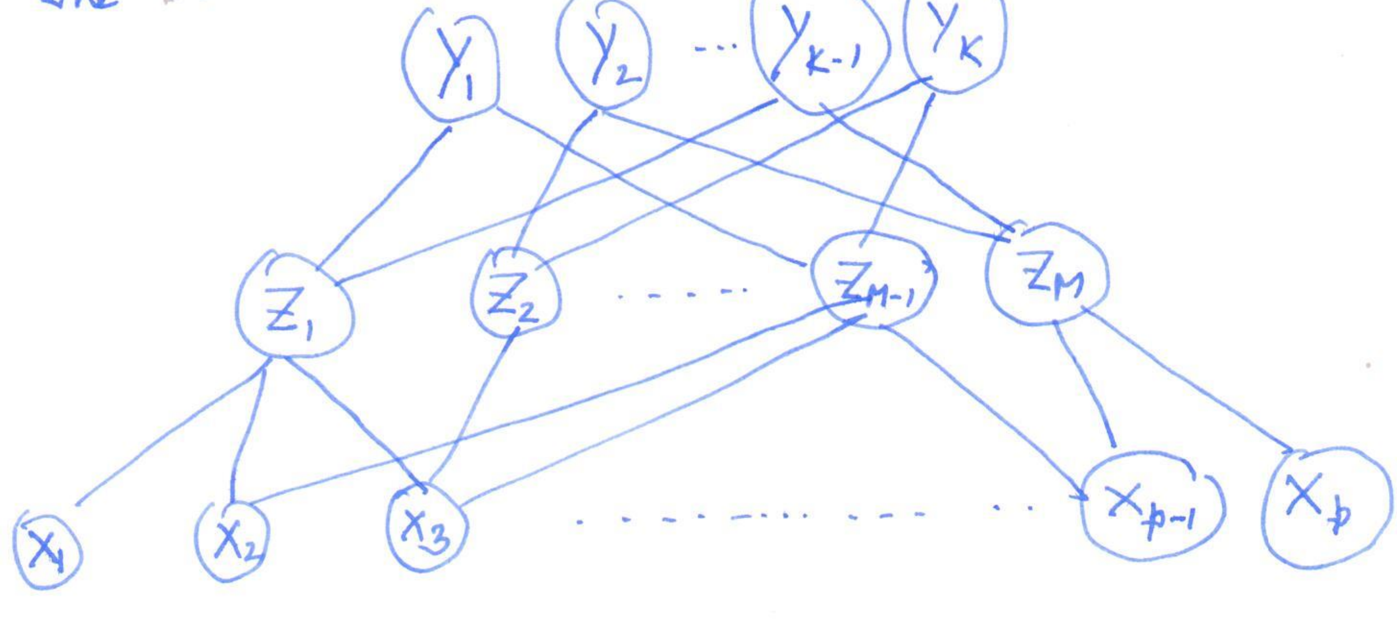
The fn. g is estimated from the data.

This is known as universal approximator.

Neural Networks

Non-linear statistical models much like PPR.

The network:



Neural Networks

Essentially the model works as follows:

First, the derived variables are created from ~~the~~ the input variables. The derived variables are sigmoid functions of linear combination of input variables. Linear fns. of the derived variables are created & these are subsequently used for value estimation/classification.

$$Z_m = \sigma(\alpha_{0m} + \alpha_m' \underset{\sim}{X}), \quad m = 1, 2, \cdots M$$

$$T_k = \beta_{0k} + \beta_k' \underset{\sim}{Z}, \quad k = 1, 2, \cdots K$$

$$f_k(\underset{\sim}{X}) = g_k(T)$$

Usually $g_k(T) = \dfrac{e^{T_k}}{\sum_{j=1}^{K} e^{T_j}}$

Note: When σ is the identity function, the entire model collapses into a linear model.

Fitting the NN model:

The NN model has the following parameters:

Let $\underset{\sim}{X} = (x_1, x_2, \cdots x_p)$

Then $\alpha_m = (\alpha_{0m}, \alpha_{1m}, \cdots, \alpha_{pm})$; $m = 1, 2 \cdots M$

$\beta_k = (\beta_{0k}, \beta_{1k}, \cdots \beta_{Mk})$, $k = 1, 2 \cdots K$

$\Rightarrow$ We have $M(p+1) + K(M+1)$ weights

.10.2017

Neural Networks

We use SSE as the measure of fit for regression

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(y_{ik} - f_k(\underset{\sim}{x}_i) \right)^2,$$

For classification we use the cross-entropy (deviance)

$$R(\theta) = -\sum_{k=1}^{K} \sum_{i=1}^{N} y_{ik} \ln f_k(\underset{\sim}{x}_i)$$

———— × ————

Neural Networks

Starting Values : Near zero random weights are selected. The starting model is ~~ti~~ nearly linear. Starting with large values generally lead to poor solution.

Overfitting : NN ~~with~~ has $M(p+1)$, $K(M+1)$ parameters. In order to avoid overfitting we use a loss + penalty approach.

We add a penalty to the error function

$$R(\theta) + \lambda J(\theta)$$

$$J(\theta) = \sum \beta_{km}^2 + \sum \alpha_{ml}^2,$$

31.10.2017

Cluster Analysis

Grouping items on the basis of certain variables.

Hierarchical & Non-hierarchical clustering

→ Agglomerative clustering : All items are single point clusters. The two closest clusters are combined into a new cluster. The ~~process sequencing~~ sequence of joining is shown in the dendogram.

Linkage methods : The linkage method defines how the distance between clusters is computed.

~~Notice that~~ Suppose C_1 and C_2 are two clusters having n_1 and n_2 items. The distance between C_1 & C_2 is a function of the distances between the individual element.

~~&i~~ Let d_{ij} be the distance between ith item in C_1 and jth item in C_2

Single linkage :- The distance between C_1 & C_2 is taken as $\min\{d_{ij} / i \in C_1, j \in C_2\}$

Complete linkage : $D = \max(d_{ij})$

31. 10. 2017

Cluster Analysis

Note: Single & complete linkage methods are based upon the similarity of 'most similar' and 'most dissimilar' pairs. (When do we need these?)

Average linkage :- Distance between C_1 & C_2 is taken to be the average of d_{ij}, i.e. $\bar{d} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij}$.

Average linkage tends to produce clusters with small within cluster distance and tends to create clusters with approximately same within cluster variance.

Ward's Method : ~~The distance is taken as the sum of squares between two clusters~~ The objective at each stage is to minimize the increase in total within cluster error sum of squares.

Cluster Analysis

31. 10. 2017

Hierarchical Clustering

In certain cases a cluster resembles an evolutionary tree. This is quite natural in biological applications. Other areas where hierarchical classifications are appropriate might be social systems or taxonomy development.

Example : Suppose we are trying to segment customers. (?)

Agglomerative Methods : We start with n single member clusters. We then group the individual items to form the cluster.

Single Linkage :- This is called the nearest-neighbour technique.

asisgangopadhyay@yahoo.com

1

Cluster Analysis

5.11. 2017

Measurement of proximity :

A clustering investigation starts with a n×n matrix that provides a ~~measure of a~~ quantitative measure of similarity or dissimilarity between the individual elements.

~~We Note that~~

Suppose an individual X_i is characterized by a $p \times 1$ vector where the p elements are $x_{i1}, x_{i2} \cdots x_{ip}$.

When all x_{ij}'s are measured in at least interval scale the distance between X_i and X_j may be measured in terms of $(x_{ik} - x_{jk})$, $k = 1, 2 \cdots p$.

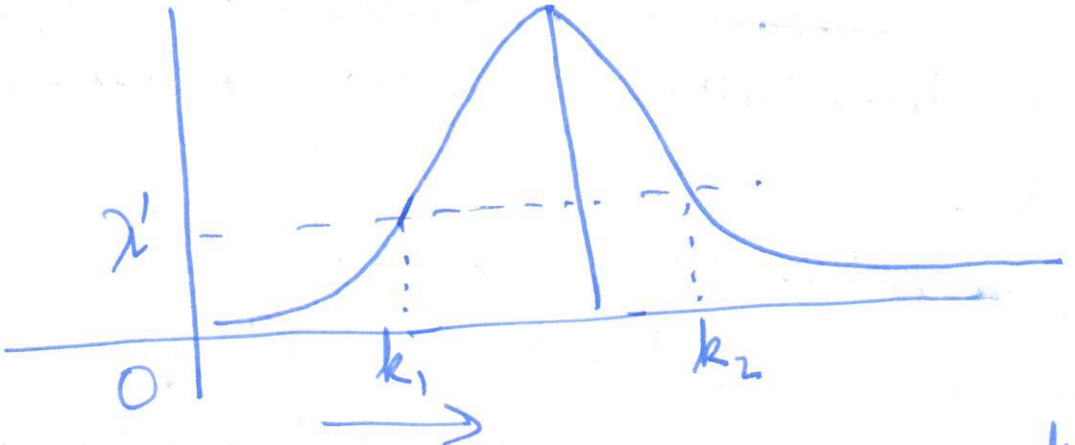
We may consider $(x_{ik} - x_{jk})^2$ or $|x_{ik} - x_{jk}|$ or any other similar numeric function.

Difficulties arise when some or all of the x_{il}'s are categorical.

8 LR Test 9 1.11.2017

Note that $\left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}\cdot\frac{\hat{\sigma}^2}{\sigma_0^2}}$ is a fn. of the form $f(x) = x^c e^{-cx}$

This is a unimodal fn. with max at 1 of the following form



$f(x) < \lambda' \Rightarrow x < k_1$ or $x > k_2$

Wald Test
LR Test
AIC/BIC & model selection

Cluster analysis ~~1~~ ② 6.11.2017

Measures of proximity

i) Binary characteristics – Matching coefficients

Jaccard Coeffs

ii) Categorical characteristics with more than 2 classes

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk}$$

where ~~α~~ $s_{ijk} = \begin{cases} 1 & \text{if } x_{ik} = x_{jk} \\ 0 & \text{otherwise} \end{cases}$

$$d_{ij} = 1 - s_{ij}$$

iii) Numerical ~~data~~ characteristics :

Euclidian Measure - $d_{ij} = \sqrt{\sum_{k=1}^{p} \omega_k^2 (x_{ik} - x_{jk})^2}$

City Block $d_{ij} = \sqrt{\sum_{k=1}^{p} \omega_k |x_{ik} - x_{jk}|}$

Minkowski $d_{ij} = \left(\sum \omega_k^r |x_{ik} - x_{jk}|^r\right)^{1/r}$

Must satisfy triangular inequality

Must be scaled

(5) LR test 1.11.2017

Example: Suppose we are testing $H_0: \mu \le \mu_0$ vs. $H_1: \mu \geq$

σ unknown.

Let the entire parameter space be Θ

$$\Theta = \{(\mu, \sigma) / -\infty \le \mu < \infty, \sigma > 0\}$$

Let the parameter space under H_0 be Θ_0.

$$\Theta_0 = \{(\mu, \sigma) / -\infty < \mu \le \mu_0, \sigma > 0\}$$

~~Let $\hat{\mu} = \bar{x}$ under~~

Note that under Θ, $\hat{\mu} = \bar{x}$

Note that if $\bar{x} \le \mu_0$, Max L_0 = Max L_1

If $\bar{x} > \mu_0$ then the previous analysis applies with the change that H_0 is rejected when

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > K$$ instead of the absolute value.

Example: Suppose we are testing $H_0: \sigma^2 = \sigma_0^2$ vs. $H_1: \sigma^2 \neq \sigma_0^2$

$\Theta_0 = \{(\mu, \sigma) : -\infty < \mu < \infty, \cancel{\sigma >} \sigma = \sigma_0\}$, $\Theta = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$

$$\text{Max } L_0 = \left(\frac{1}{\sigma_0 \sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma_0^2} \Sigma (x_i - \bar{x})^2} = \left(\frac{1}{2\pi \sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma_0^2}}$$

$$\text{Max } L_1 = \left(\frac{1}{\hat{\sigma}^2 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2\hat{\sigma}^2} \Sigma (x_i - \bar{x})^2}$$

$$= \left(\frac{1}{2\pi \hat{\sigma}^2}\right)^{n/2} e^{-\frac{n}{2\hat{\sigma}^2} \cdot \hat{\sigma}^2} = \left(\frac{1}{2\pi \hat{\sigma}^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

Cluster Analysis

5. 11. 2017

Jaccard Coefficient: Note that d may not really provide any useful information. The fact that ~~two individuals~~ neither of two individuals share a characteristic, may not imply that they are similar.

Note: Choice of similarity measure, eg. Matching coefficient or Jaccard coefficient can lead to different assessment of similarity.

Similarity measures for categorical data with more than 2 levels:

Let ~~there be~~ p characteristics ~~have~~ be categorical variables with more than 2 levels.

Let $s_{ijk} = \cancel{1} \begin{cases} 1 & \text{if individuals } i \,\&\, j \text{ are same wrt characteristic } p \\ 0 & \text{otherwise.} \end{cases}$

Then the similarity between i and j is given by

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}$$

(7) LR Test

1.11.2017

$f(x) = x^c e^{-cx}$

$f'(x) = c x^{c-1} e^{-cx} - c e^{-cx}$

$= c e^{-cx}(x^{c-1} - 1)$

$f'(g(x)).g'(x)$

$c e^{-cx}(x^{c-1} - 1) = 0$

$\Rightarrow x^{c-1} = 1 \Rightarrow x = 1$

$\lambda' < k \Rightarrow \frac{\hat{\sigma}^2}{\sigma_0^2} < k_1$ or $\frac{\hat{\sigma}^2}{\sigma_0^2} > k_2$

Example: We wish to test

$H_0: \sigma^2 = \sigma_0^2$ vs. $H_1: \sigma^2 \neq \sigma_0^2$

Let $X_1, X_2 \dots X_n$ be iid $N(\mu, \sigma^2)$, μ, σ unknown

Let ~~ω b~~ ω be the parameter space under H_0 (i.e. the restricted parameter space)

Let Ω be the unrestricted parameter space.

$$L = \frac{\sup_{(\mu,\sigma^2)\in\omega} f(x, \mu, \sigma^2)}{\sup_{(\mu,\sigma^2)\in\Omega} f(x, \mu, \sigma^2)}$$

Cluster Analysis 5. 11. 2017

Similarity / dissimilarity measure for binary data :

Suppose p characteristics of an entity are binary.

~~When we compare two individuals~~

The result of comparison between two individuals may be may be compiled in a 2X2 table as follows.

		Individual 2		
	Outcome	1	0	Marginal
Individual 1	1	a	b	a+b
	0	c	d	c+d
Marginal		a+c	b+d	$p=a+b+c+d$

The level of similarity may be measured in terms of the cell frequencies. Two important measures are :

a) Matching coefficient: $\dfrac{a+d}{a+b+c+d}$

b) Jaccard coefficient: $\dfrac{a}{a+b+c}$

⑧ LR Test

1.11.2017

When $(\mu, \sigma^2) \in \omega$, $\hat{\mu} = \bar{x}$ ~~and $\hat{\sigma}^2 = \sigma_0^2$~~ σ_0^2 given

$$\Rightarrow \sup_{(\mu,\sigma^2)\in\omega} f(\underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i / \bar{x}, \sigma_0^2)$$

$$= \left(\frac{1}{\sigma_0^2 \cdot 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_0^2}\sum(x_i-\bar{x})^2}$$

When $(\mu, \sigma^2) \in \Omega$, $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$

$$\Rightarrow \sup_{(\mu,\sigma^2)\in\Omega} f(\underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i / \bar{x}, \hat{\sigma}^2)$$

$$= \left(\frac{1}{\hat{\sigma}^2 2\pi}\right)^{\frac{n}{2}} e^{-\frac{1}{2\hat{\sigma}^2}\sum(x_i-\bar{x})^2}$$

$$= \left(\frac{1}{\hat{\sigma}^2 \cdot 2\pi}\right)^{\frac{n}{2}} e^{-\frac{n}{2}}$$

$$\Rightarrow L = \frac{\sup_{(\mu,\sigma^2)\in\omega} f(\underset{\sim}{x})}{\sup_{(\mu,\sigma^2)\in\Omega} f(\underset{\sim}{x})}$$

$$= \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} \cdot \frac{e^{-\frac{n}{2}\cdot\frac{\hat{\sigma}^2}{\sigma_0^2}}}{e^{-\frac{n}{2}}}$$

$$L < \lambda \Rightarrow \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}\cdot\frac{\hat{\sigma}^2}{\sigma_0^2}} < \lambda'$$

① Shrinkage Method ① 8.11.2017

Ridge: $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij}\right)^2 + \lambda \sum \beta_j^2$

$\lambda \geq 0$ is the tuning parameter.

When $\lambda = 0$, Ridge Regression ~~bec~~ becomes MLR

When $\lambda \to \infty$ the variance becomes 0 but bias increases.

β_0 is not shrunk.

~~When all x_i~~ Take ~~$x_i' = x_i - \bar{x}_i$~~ z_i ~~y_i~~ $= x_i - \bar{x}_i$

Then $\bar{z}$ ~~$\bar{y}$~~ $= \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

In this case $\hat{\beta}_0 = \frac{1}{n} \sum y_i = \bar{y}$

Scale Invariance: The LS coefficients are scale invariant, i.e. $X_j \hat{\beta}_j$ is same irrespective of the scale of ~~X~~. x_j.

Ridge coefficients are not scale invariants.

We use $\tilde{x}_{ij} = \dfrac{x_{ij}}{\sqrt{\frac{1}{n} \sum (x_{ij} - \bar{x}_j)^2}}$

Shrinkage Method ② 8.11.2017

Why does shrinkage method perform better?

As $\lambda \uparrow$ the variance ~~of the prediction~~ reduces at the expense of a small increase of bias.

Lasso: Ridge regression does not achieve parsimony. The penalty $\lambda \sum \beta_j^2$ shrinks all $\beta_j \to 0$

Note: Ridge regression ~~does not~~ may not lead to reduction of accuracy. However, interpretability may be a problem as p does not reduce.

Lasso: Minimize $\sum(y_i - \beta_0 - \sum_{j=1}^{p} \beta_{ij} x_j)^2 + \lambda . \sum|\beta_j|$

Note: Ridge imposes l_2 penalty norm whereas ~~Lasso~~ lasso imposes l_1 penalty.

Note: Lasso performs variable selection. The l_1 norm has the effect of forcing some coefficients to ~~zero~~ zero.

Shrinkage Method 8.11.2017

Alternative formulation for ridge & lasso

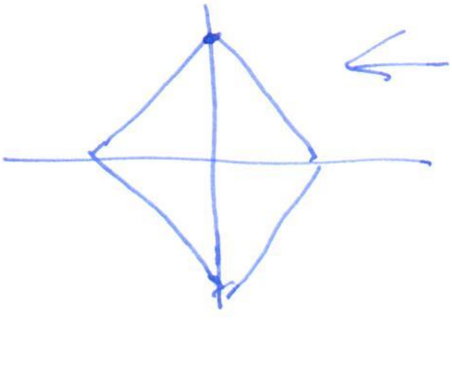
Lasso: $\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$ subject to $\sum_{j=1}^{p}|\beta_j| \le s$

~~and~~ and Ridge: $\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$

subject to $\sum_{j=1}^{p} \beta_j^2 \le s$

These formulations are equivalent in the sense that for every λ one can find an s that gives the same set of coefficient estimates & the other way round.

Special case: When $p = 2$, lasso coefficients correspond to the least RSS for (β_1, β_2) falling in the diamond described by $|\beta_1| + |\beta_2| \le s$



← A region defined by $|\beta_1| + |\beta_2| \le s$

Shrinkage Method

8.11.2017

Similarly ~~for~~ ridge regression estimates have the smallest RSS out of all points ~~th~~ that lie within the circle defined by

$$\beta_1^2 + \beta_2^2 \le s$$

Subset selection

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \sum_{j=1}^{p} I(\beta_j \neq 0)$$

Comparing ridge & lasso :

No clear winner.

When some of the coe