BUSINESS ANALYTICS

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

BUSINESS ANALYTICS

– TANUJIT CHAKRABORTY.
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:

X	Y: 1 2 3 c	Marginals
1		N_1
2		N_2
⋮		⋮
r		
	N_1 N_2	N

$$\left.\begin{matrix} P(Y=1 \mid X=1) \\ P(Y=1 \mid X=2) \end{matrix}\right\}$$ 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 pvalue

1. check whether the basic assumptions are satisfied or not.
[Residual plots]
2. Check the existence of outliers / influential observation.

'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) $E(\epsilon_i) = 0 \quad \forall \quad i = 1, 2, \ldots$

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.

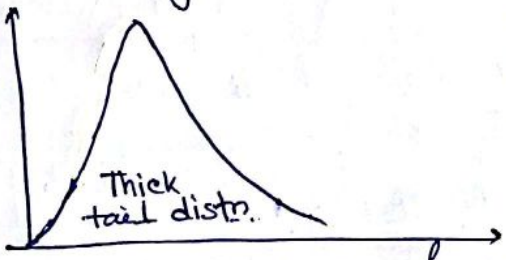
Note: The stratified boxplots may show the following patterns:
- The mean/median of the response may not change as the explanatory variables change.
- Some high/low percentile of the response may change as the explanatory variables.

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. } ⇒ 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.]

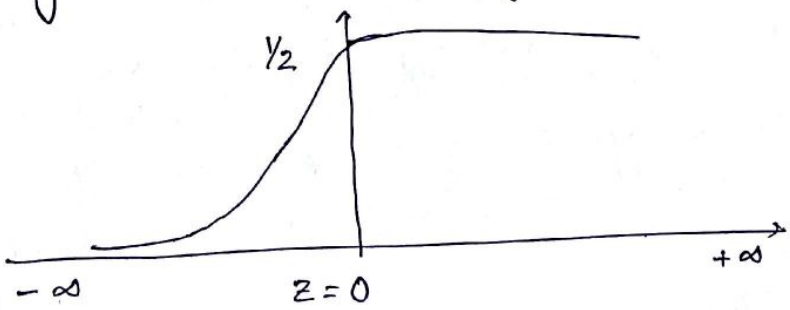
Logistic Regression:- (Binary)

Essentially used for classification and risk analysis.

Classification means when response is categorical.

Explanatory → Numeric and Categorical.

Sigmoid function:



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

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 attempts to quantify risk (Risk event, expenses, probability of risk event).

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

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

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

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

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

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

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

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

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

Logistic regression model:-

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

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

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 2×2 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 = Sensitivity = $P(\hat{y} = 1 \mid y = 1)$ = Prob. of (T/D)

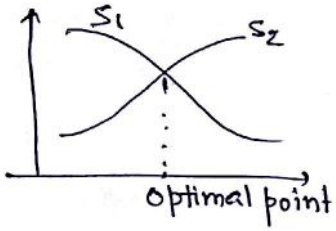
S_2 = Specificity = $P(\hat{y} = 0 \mid y = 0)$ = Prob. of $(\bar{T}/\bar{D})$

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

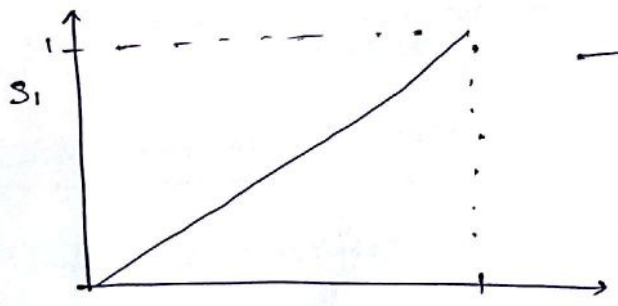
[T = test positive
D = Disease exists
$\bar{T}$ = Test negative
$\bar{D}$ = Disease doesn't exist]

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

S_1 = True positive.



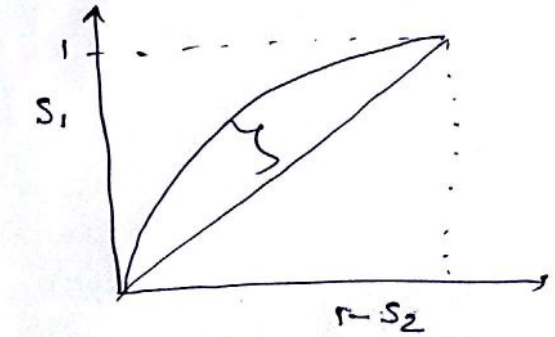
ROC Curve:



— 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	Non-parametric
• Logistic Regression (essentially Binary)	• Naive Bayes
• Linear Discriminant Analysis (LDA)	• Decision Trees (Bagging, Boosting CART, Random forest)
• Quadratic Discriminant Analysis (QDA)	• K-NN
↳ is generally not preferred as the number of the parameters to be estimated increases manifold.	• 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

$\pi_K = Pr(Y = k)$ ⟶ unconditional preposition.

$f_K = P(\underset{\sim}{X} = \underset{\sim}{x} \mid Y = k)$ → 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:-

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

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

$$= \frac{\pi_k f_k}{\sum_{l=1}^{n} \pi_l f_l}$$

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.

$$\hat{\mu}_k = E(\widehat{X \mid Y=k}) = \frac{\sum x_j}{f_k}$$; 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 $\hat{f}(x) \leftarrow 0$
   $r_i \leftarrow y_i$ 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 shrinkage parameter
   (c) Update the residual $r_i \leftarrow r_i - \lambda \hat{f}_b(\underset{\sim}{y})$
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 $\left[1/x, \sqrt{x}, x^2, x^3, \ln x\right]$

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 implication.

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

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

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 $x, x^2, x3$ 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:-

Supervised Analytics → Linear regression → LRM → Stepwise
Linear regression → MLR → Subset Selection
↓
Non-linear models (Splines, natural splines)
↓
Projection Persuit Regression (PPR) ← Shrinkage Method

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 penalty term.

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

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

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.

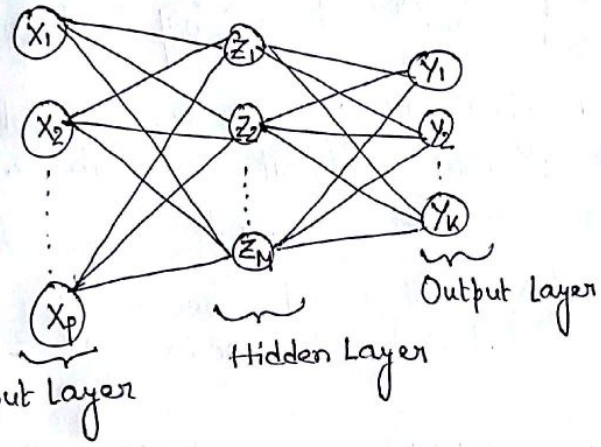
Suppose we have a K-class classification model:

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

Here we estimate $P(Y=j|\underset{\sim}{X})$; $j=1,2,\ldots,K$ 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:

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

$$T_K = \beta_{0k} + \sum_j \beta_{jk} Z_j; \quad K=1,2,\ldots,K.$$

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

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:-

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

$T \rightarrow$ Linear function of Z.

$g_k \rightarrow$ Output function.

Use the NN model for the purpose of prediction only.

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

- 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

$$R(Q) = \sum_{1}^{K} \sum_{1}^{N} \left(y_{ik} - f_k\left(\underset{\sim}{x}_i\right) \right)^2 \quad ; \quad \begin{matrix} N \text{ observations \&} \\ K \text{ response variable.} \\ (\text{Usually } K=1) \end{matrix}$$

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 square of 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.

Smoking	Lung cancer: Yes	Lung cancer: 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 :-

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

$\Leftrightarrow$ Maximizing $P(A_1 A_2 \ldots A_p \mid Y=j) P(Y=j)$

Rule: Allocate that subject to that j where $P(A_1 A_2 \ldots A_p \mid Y=j)$ is maximum.

Naive Bayes Assumption:- $A_1, A_2, \ldots, A_p$ are conditionally independent given $Y=j$.

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

(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 $R_1, R_2, \ldots, R_M$.

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

Let $E(Y \mid \underset{\sim}{X} \in R_j ;\ j=1,2,\ldots,M)$ 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 $E(Y|X)$.

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

Input Data:

X_1	X_2	.....	X_K	Y
x_{11}	x_{12}	.....	x_{1K}	y_1
x_{21}	x_{22}	.....	x_{2K}	y_2
⋮	⋮		⋮	⋮
x_{N1}	x_{N2}	.....	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 $SSE = \sum (y_i - \bar{y})^2$.

   Step2: 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

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

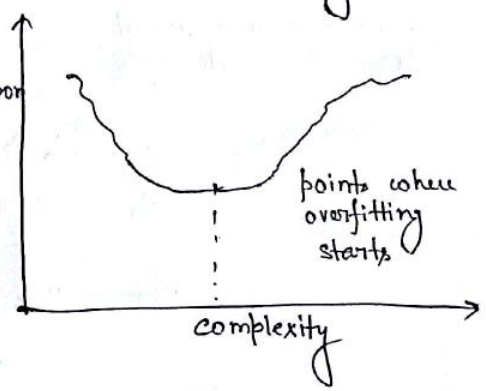
Note: A tree grows very fast. Consequently trees may result in "over fitting" (saturated model).

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 is 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 than 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 leaving 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.

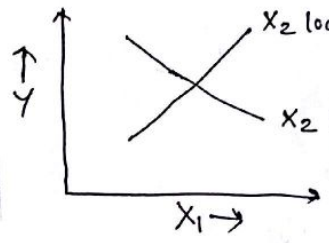
→ **K-fold cross validation:-** Divide the data into K subgroups. Use K-1 subgroups to build the model (tree) and 1 subgroup to test it. Apply it for all the cases.

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

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

→ By drawing diagram 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_{Rjk} ; $k = 1, 2, \ldots, p$.

Let $p_{Rjl} \geq p_{Rjk} \quad \forall\ k$

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

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

(b) Gini Index:- Sum of → $\sum_{k=1}^{p} p_{Rjk}(1 - p_{Rjk})$, p classes.

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

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

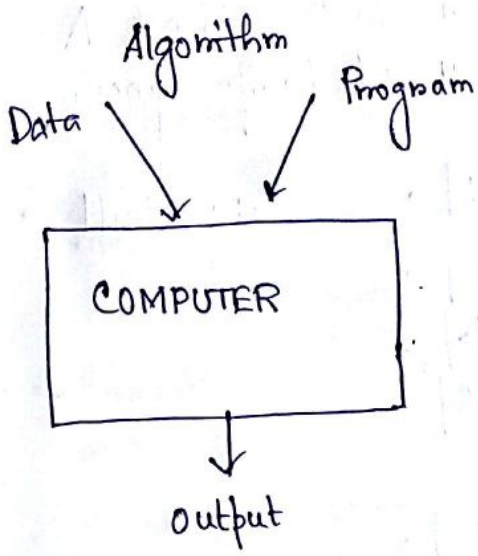
In classification tree we use node purity and in regression tree we use sum of square of error.

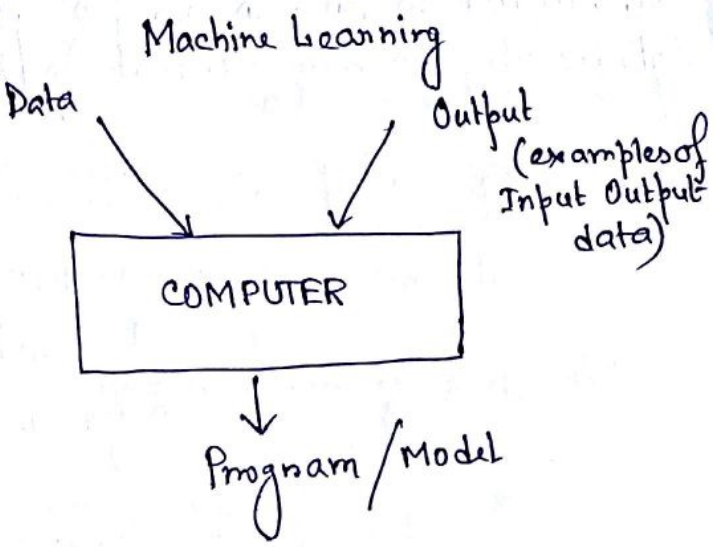
MACHINE LEARNING

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 : Rosenblatt'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:

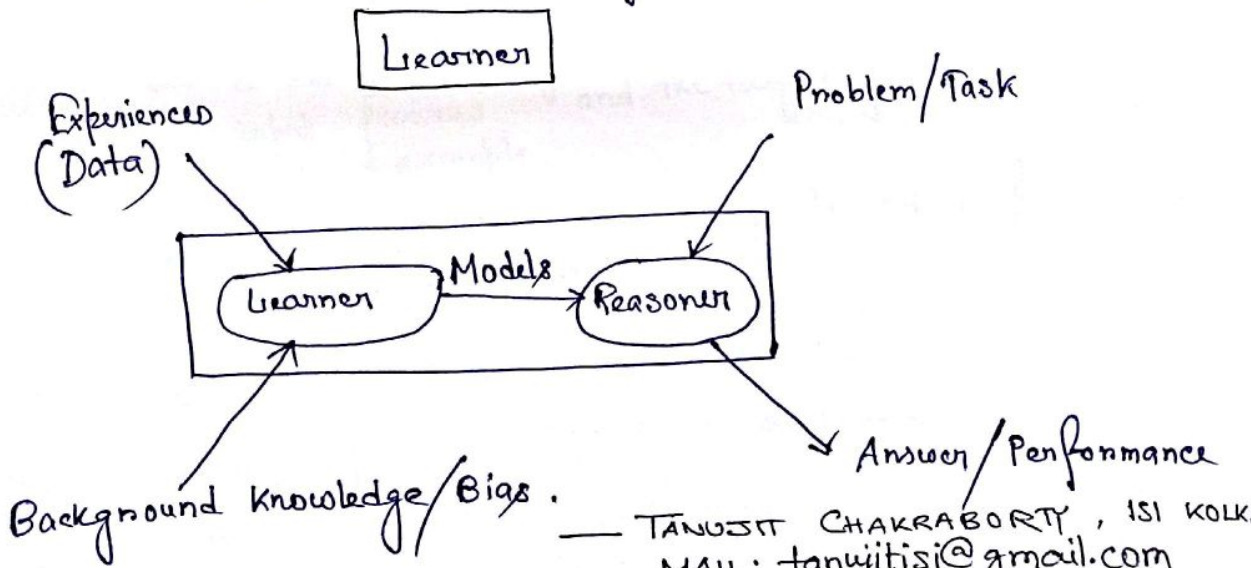
- Predict if a stock will rise or fall in the next few milliseconds.
- Predict if a user will click on an ad or not in order to decide which ad to show.

— 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.