UNIT-II

Linear Correlation- Regression Modelling- Multivariate Analysis- Bayesian Modelling- Inference and Bayesian Networks- Support vector and Kernel Methods- Analysis of time series- Linear System Analysis- Non Linear Dynamics- Rule Induction- Basic Fuzzy and Neural Networks

Linear correlation refers to straight-line relationships between two variables. A **correlation** can range between -1 (perfect negative **relationship**) and +1 (perfect positive **relationship**), with 0 indicating no straight-line **relationship**.Linear correlation is a measure of dependence between two random variables.

Definition

Let X and Y be two random variables. The **linear correlation coefficient** (or Pearson's correlation coefficient) between X and , Y denoted by Corr[X,Y] is defined as follows:

where is the covariance between and and and are the standard deviations Corr[X,Y]=Cov[X,Y]/**σ**[X]**σ**[Y] where Cov[X,Y] is the Covariance[X,Y] .

Note that, in principle, the ratio is well-defined only if **σ**[X]**and σ**[Y] and are strictly greater than zero. However, it is often assumed that Corr[X,Y]=0 when one of the two standard deviations is zero. This is equivalent to assuming that0/0=0 because Cov[X,Y]=0 when one of the two standard deviations is zero.

Interpretation

The interpretation is similar to the interpretation of covariance: the correlation between X and Y provides a measure of how similar their deviations from the respective means are

Linear correlation has the property of being bounded between -1 and 1

$$-1 \leq \text{Corr}[X,Y] \leq 1$$

Thanks to this property, correlation allows to easily understand the intensity of the linear dependence between two random variables: the closer correlation is to 1, the stronger the positive linear dependence between X and Y is (and the closer it is to -1, the stronger the negative linear dependence between X and Y is).

Terminology

The following terminology is often used:

1. If Corr[X,Y]>0 then X and Y are said to be positively linearly correlated (or simply positively correlated).
2. If Corr[X,Y]<0 then X and Y are said to be negatively linearly correlated (or simply negatively correlated).
3. If Corr[X,Y]≠0 then X and Y are said to be linearly correlated (or simply correlated).

4. If Corr[X,Y]=0 then X and Y are said to be uncorrelated.

Correlation of a random variable with itself

Let X be a random variable, then corr[X,X]=1

Symmetry

The linear correlation coefficient is symmetric:

Corr[X,Y]=Corr[Y,X]

Regression Modelling:

It includes many techniques for **modeling** and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). ... In all cases, a function of the independent variables called the **regression** function is to be estimated.

Correlation and linear regression are not the same. Correlation quantifies the degree to which two variables are related. Correlation does not fit a line through the data points. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.

REGRESSION:

Data can be smoothed by fitting data to a function such as with regression.

Linear regression involves finding the best line to fit two variables or attributes so that one attribute can be used to predict the other.

Multiple linear regression:More than two attributes are involved and the data are fit to a multidimensional surface.

Linear Regression:Straight line regression analysis involves a response variable Y and a single predictor variable X.It is the simplest form of regression and models Yas a linear function of X i.e.

$$Y=b+wx$$

Where the variance of Y is assumed to be constants and b,w are regression coefficients specifying the Y-intercept and slope of the line.
Regression coefficient w&b can also be thought of as weight, so that can equivalently write

$$Y=W_0+W_1X$$

These coefficients can be solved by method of least squares,which estimates the best fitting straight line as the one that minimize the error between the actual data and the estimate of the line.

Regression coefficient can be estimated using

$$w_1 = \frac{{}^{|D|}{}_{i=1}\sum (x_i - x)(y_i - y)}{{}^{|D|}{}_{i=1}\sum (x_i - x)^2}$$

Example- Straight line regression using method of least squares.

X years experience	Ysalary(in$1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Distance between two binary variables based on the notion of similarity. For example, the asymmetric binary similarity between the objects ‘i’ and ‘j’, or sum(i,j) can be computed as

$$sum(i,j)=q/q+r+s=1-d(i,j)$$

The coefficient sum(i,j)is called Jaccard coefficient.

MULTIPLE LINEAR REGRESSION

Multiple linear regression model based on 2 predictor attributes or variable A_1 and A_2i.e

$X_1$and X_2→ values of attributes A_1 and A_2 in x.

Multiple regression problems are solved with software packages such as SAS , Spss and S-Plus.

CO-RELATION CO-EFFICIENT

DEFINITION

Let X and Y be two random variables. The linear correlation co-efficient or Pearson's Correlation co-efficient between X and Y denoted by

INTERPOLATION : It is similar to the interpretation of covariance. The correlation between X and Y provides a measure of how similar their deviation from the respective means are

TERMINOLOGY

- If Corr[X,Y] >0 ,then X and Y are said to be positively linearly correlated.
- If Corr[X,Y]<0 ,then it is said to be negatively linearly correlated.
- If Corr[X,Y]≠0 ,then X and Y are said to be linearly correlated.
- If Corr[X,Y]=0 ,then X and Y are said to be uncorrelated.

BAYESIAN CLASSIFICATION

Bayesian classifications are statistical classifiers. They can predict class membership probabilities ,such as the probability that a given tuple belongs to a particular class. Bayesian Classification have exhibited high accuracy and speed when applied to long database. Naive Bayesian classification assume that the effect of an attribute value on given class is independent of value of other attributes. This assumption is called class conditional independence.
It is made to simplify the computations involved ->it is called as Naïve.

BAYES THEOREM

‘ X ’ is considered as evidence. It is hypothesis,such as that the data tuple ’ X ‘ belongs to a specified class ‘ c ‘.
P(H/X) → represents, looking for the probability that tuple ‘ X ‘ belongs to class ‘ c ’ , given that we know the attribute description of ‘ X ’.

P(H/X) is the posterior probability of H conditional in ‘ X ‘.

FOR EXAMPLE,
A Customer is described by the attribute age and income respectively, and that ‘X’ is a 35 year old customer with an income of $40,000. Suppose that ‘H’ is the hypothesis that our customer will buy a computer given that we know the customer’s age and income.
P(H) → Prior-probability, for our example, this is the probability that any given customer will buy a computer regardless of age, income or any other information.
Similarly, P(X/H) is the posterior probability.
P(X) →prior probability of ‘ X ‘
Above probabilities are estimated using Bayes Theorem.

BAYES THEOREM,

How Bayes theorem is used in Naive Bayesian classifier

The Naive Bayesian Classifier or simple Bayesian Classifier work as follows

1. Let ‘ D ‘ be a training set of tuples and their associated class labels. Each tuple is represented by an n-dimensional attribute vector, X=(x_1,x_2,…..) depicting ‘ n ’ measurements made on the tuple from ‘ n ’ attributes A_1,A_2…..

2. Suppose there are ‘ m ‘ classes C_1,C_2,…

   Given a tuple‘ X’, the classifier predict that ‘ X’ belongs to the class having the highest posterior probability , conditioned on ‘ X’ i.e , the Naïve Bayesian Classifier predicts the tuple ‘ X ‘ belongs to the class if and only if

   Thus, we maximize P(/X). The class for which P(/X) is maximized is called maximum posterior hypothesis.

   By Bayes Theorem,

3. As P(X) is constant for all classes, only P(X/)P() need to be maximized. If the class prior probabilities are not known, then it is assumed as

4. If the dataset with many attributes computation is extremely expensive to compute P(X/) to reduce the computation.

   Use,

For each attribute, we look at whether the attribute is categorical or continuous valued. To compute P(X/) consider the following

a) If is categorized, then P(/) is the number of tuples of class in ‘D’ having the value for divided by / , D/ , the number of tuples of class in ‘ D ‘ .

b) If is continuous assumed to have a Gaussian distribution with a mean μ and S.D σ defined by

g(x,μ,σ)=(1/√2πσ)e^-((x-μ)^2)/2σ^2

P(/)=g(,μ,,σ)

5. In order to predict the class label X , P(X/)P() is evaluated for each class .
   The classifier predicts the class label of tuple X is the class if and only if

   $$P(X/)P() > P(X/)P() , \text{ for } i \leq j \leq m , j \neq i$$

   In other words, the predicted class label is the class for which P(X/)P() is the maximum.

Bayesian Belief Networks Or Belief Networks Or Bayesian Networks Or Probabilistic Networks

Bayesian belief network specify joint conditional probability distributions. They allow class conditional independencies to be defined between subset of variables. Trained Bayesian belief networks can be used for classification.

Belief networks is defined by two components – a directed acyclic graph and a set of probability tables.

A Simple Bayesian Belief Network

If an arc is drawn from a node Y to node Z, then Y is the parent or immediate predecessor of Z, and Z is the descendant of Y. each variable is conditionally independent of its non-descendants in the graph, given its parents.

The arcs in the figure allow a representation of casual knowledge. For example, having lung cancer is influenced by a person’s family history of lung cancer, as well as whether or not the person is a smoker.

- Note that the variable positive X-Ray is independent of whether the patient has a family history of lung cancer or is a smoker, given that we know that the patient has lung cancer.
- In other words, once we know the outcome of the variable lung cancer, then the variables family history and smoker do not provide any additional information regarding positive X-Ray.

The arcs also show the variable lung cancer is conditionally independent of Emphysema, given its parents, family history and smoker.

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable Y specifies the conditional distribution P(Y/parents(Y)), where parents(Y) are the parents of Y.

	FH, S	FH, ~S	~FH, S	~FH, ~S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

This shows a CPT for the variable lung cancer. The conditional probability for each known value of lung cancer is given for each possible combination of values of its parents. For instance from the table, the upper leftmost and bottom right most entries we see that

P (Lung Cancer=yes /Family History =yes, Smoker =yes) =0.8

P (Lung Cancer=no /Family History =no, Smoker =no) =0.9

Let X= (x_1, x_2 … x_n) be a data tuple described by the variables or attributes Y_1, Y_2 …. $Y_n$respectively. Recall that each variable is conditionally independent of its non-descendants in the network graph, given its parents. This allows the network to provide a complete representation of the existing joint probability distribution with the following equation:

$$\mathbf{P\ (x_1,\ x_2\ \ldots.\ x_n) = {}^{n}_{i=1}\,P\ (x_i/parents(y_i))}$$

Where P (x_1, x_2 …. x_n) is the probability of a particular combination of values of X, and the values for P (x_i/parents(y_i)) corresponds to the entries in the CPT for Y_i.
A node within the network can be selected as an “output” node representing a class label attribute. There may be more than one output node. Rather than returning a single class label, the classification process can return a probability distribution that gives the probability of each class.

Multi- variate analysis

It is a set of techniques used for analysis of data sets that contain more than one variable, and the techniques are especially valuable when working with correlated variables.

- Here instead of looking at several variables separately, in multivariate analysis we will be looking at them simultaneously and hence we will be able to study the interrelationships between the variables.

Application areas:

- Social science (gender, age, nationality of an individual).
- Climatology (min temp, max temp, rainfall, humidity) on a day.
- Econometrics (input costs, production, profit) of a firm.

- Medical (BP, pulse rate) of persons.
- Administrative (admissions, operations, discharges, deaths) per day in hospital.

Multivariate analysis is classified as

- Classification of individuals.
- Dimension reduction.
- Cause -effect relationship

Cluster analysis:

Clusters are homogenous with itself but different from another cluster. It tells us how the individuals are similar and dissimilar among themselves.

- How the clusters are different is answered by discriminant analysis.

  Discriminant analysis: studies the properties of a given cluster and thereby it identifies the difference between the different clusters.

- Can a newly arrived individual be assigned to one of the cluster?

  Classification comes into part. This is the problem of assigning new individual to the cluster and is referred to as a classification problem.

Discriminant analysis and classification graph

If a new data arrives we should plot the new X and Y data and check in which cluster it lies.

Dimensionality reduction

- PCA (K-L method)
- Factor analysis

PCA searches for K ‘n’ dimensional orthogonal vectors that can best be used to represent the data where k<=n. It combines the essence of attributes by creating an alternating smaller set of variables. The entire data can then be projected into the smaller set. PCA often reveals the relationship that were not previously supported.
The basic procedure is as follows:

- The input data are normalized, so that each attribute falls within the same range. This step helps us to ensure that the entire attribute with large domain will not dominate attributes with smaller domain.
- PCA computes k orthogonal vector that provide a basis for normalized input data. These are unit vectors that each point in the direction perpendicular to the others. These vectors are referred to as the principal components. The input data are linear combination of the principal components.
- The principal components are sorted in the order of decreasing “significance” or strength. The principal components essentially serve as the new set of axes for the data, providing important information about variance.

Figure shows the first two principal components Y1,Y2 for the given set of data originally mapped to the axes X1 and X2. This information helps identify groups or patterns within the data.

- Because the components are sorted according to the decreasing order of significance, the size of the data can be reduced by eliminating the weaker components (i.e.,) those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.
- It can be applied to ordered and unordered attributes, and can handle sparse and skewed data.

Rule Induction

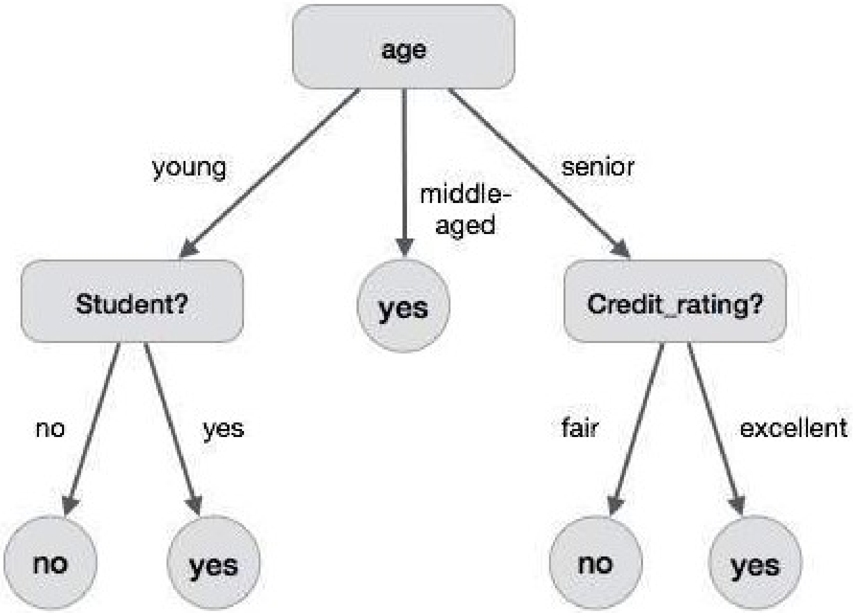
A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buys_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.

The benefits of having a decision tree are as follows:

- It does not require any domain knowledge.
- It is easy to comprehend.

- The learning and classification steps of a decision tree are simple and fast.



The expected information needed to classify a tuple in *D* is given by

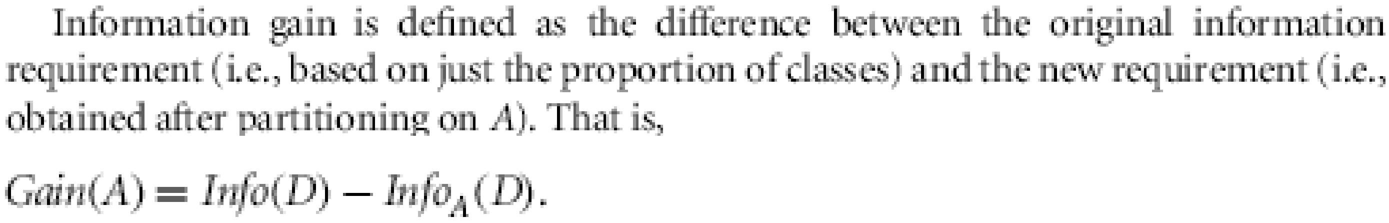
$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

Then, for each attribute A,

where Dj / D is the weight of the jth partition.

Info A (D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information, the greater the purity of the partitions.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j).$$



Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

$$Gain(A) = Info(D) - Info_A(D).$$

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

Here is the Tree Pruning Approaches listed below –

- **Pre-pruning** − The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully grown tree.

Cost Complexity

The cost complexity is measured by the following two parameters –

- Number of leaves in the tree, and
- Error rate of the tree.