Unit III

CHAPTER 4

Statistical Modelling and Machine Learning

Syllabus Topic : Introduction to Model Selection

4.1 Introduction

Model selection is an important task for any application while choosing any of the models various features are to be considered which are as follow :

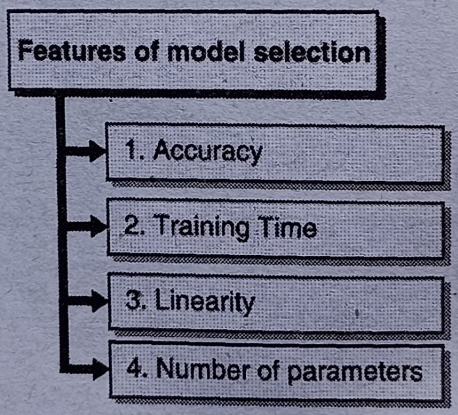


Fig. 4.1.1 : Features of model selection

➜ **1. Accuracy**

Accuracy deals with the perfection in the solution from the given data set how well your algorithm provides an accurate prediction is considered here.

➜ **2. Training time**

Various Machine Learning algorithm works in different ways and the time needed to train the given data set is an important factors if the data set has too much of variance the machine will take time to learn the training data.

3. Linearity

Many machine learning tries to use the linear classification technique where the data set is assumed to be separated by the straight line.

4. Number of parameters

It is a very important feature depending on number of parameters in the given data set the prediction can vary.

Syllabus Topic : Regularization

4.1.1 Regularization

Q. 4.1.1 Explain regularization. (Ref. Sec. 4.1.1) (5 Marks)

- Regularization is used as a solution to get rid out of the overfitting problem multivariate regression, but it can be used with both univariate and multivariate regression too.
- Regularization keeps all the features in multivariate regression but reduces magnitude values of parameters θ_j.

In regularization the cost function is modify as follow,

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda \sum_{i=1}^{n}\theta_j^2\right] \quad ...(4.1.1)$$

Where λ is regularization parameter which controls a trade off between goals.

Mainly there are two goals :

1. To fit the training data.
2. To keep parameter small.

$\sum_{i=1}^{n}\theta_j^2$ is regularization term and it will shrink all θ_s.

If λ is large then we will highly penalize that is parameter θ_s, that is all θ_s will be 0.

Therefore $J_{\theta(x)} = \theta_0$ and hence it will lead to underfitting.

λ should be chosen properly for regularization.

Regularized Gradient descent will be,

$$\theta_j = \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)} - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{m} \theta_j \right] \quad \ldots(4.1.2)$$

Ridge and Lasso Regression are comes under Regularized regression.

Syllabus Topic : Bias / Variance Trade Off E.g. Parsimony

4.1.2 Bias and Variance Trade Off

Q. 4.1.2 Explain bias and variance trade off in detail with help of an example. (Ref. Sec. 4.1.2) **(5 Marks)**

It is an property of different predictive models in which models with a lower bias for parameter estimation have a higher variance for the given data set and vice versa.

- **Bias :** It is an error form the erroneous assumption made during the learning of an algorithm. Higher bias can lead to missing of the relevant data or feature needed for the targeted value in other words it leads to underfitting.
- **Variance:** It is an error form the sensitivity of an algorithm where small fluctuation of samples in the training set can lead to an error. High variance in an algorithm can lead to generation of random noise in the training data and can deviate the output. In other words it leads to overfitting.

Bias-Variance tradeoff is generally faced in supervised algorithms due to which the accuracy and generalization both cannot be adopted in the model.

Any model can be bad or not optimal because of two main reasons :

1. It is not accurate.
2. It does not match the data well.

- The reason for the first is bias and other is variance. If models are made complex then it leads to improvement in the bias but such models are very costly which leads to higher variance, whereas when the model is made more specific to the data then the variance will be reduced but on the other hand it leads to higher bias.

- For example, if a straight line and high degree polynomial curve is considered the straight line will have no variance at all but it will have bias since it is a bad fit line, where as the polynomial curve will have no bias as the curve can be fitted according to the point but this leads to high variance. Bias error is the simplifying assumptions made by a model to make the target function easier to learn.
- Models with high bias are fast to learn and easier to understand but are less flexible, also they have lower predictive performance for the complex problems. Low Bias leads to less assumption about the form of the target function where as High Bias leads to more assumption about the form of target function.
- Variance is the amount that the estimate of the target functions which will change if different training data was used. Algorithm should have some variance.
- Low variance provides small changes to the estimate of the target function with the changes to the training dataset. High variance provides large changes to the estimate of the target function with changes to the training data set. Whenever the model is choosing with low complexity and low variance automatically the high variance is introduced.

Principle of Parsimony

Parsimony principle is used in bias variance trade off and it tell us that as we increase the number of parameters in a model bias goes on decreasing and variance keeps on increasing. Parsimony principle underlies model selection approaches. It also tells that when the numbers of parameters are increased the variance of the system increases and vice versa.

Syllabus Topic : AIC

4.1.3 AIC

Q. 4.1.3 Explain AIC in detail with its mathematical formula. (Ref. Sec. 4.1.3) (5 Marks)

- AIC (Akaike Information Criterion) is similar to a statistical model for estimating the given data set. It is a model selector and given number of models it estimates the quality of each model in comparison with other models and provides with the best model.
- If any model is estimated on a particular set of data AIC score will provide an estimation of that model performance for the new data set. AIC is similar to the in- sample error of the estimated model.
- To select the model using AIC the model which provides smallest AIC value with respect to other is chosen. To avoid the risk of overfitting AIC provides penalty by the term 2*d.

- The term "2 × d" increases as the number of parameters increases and thus it reduces the complexity of the model. AIC is best suited with respect to BIC when the model is more complex.

Mathematical Formulation of AIC

$$AIC = -2(\log - \text{likelihood}) + 2 * d$$

- Where d is the number of model parameters i.e. the number of variables and the intercept.
- Log-likelihood is a measure of model fit and it is obtained from statistical output.

Syllabus Topic : BIC

4.1.4 BIC

Q. 4.1.4 Explain BIC in detail with its mathematical formula. (Ref. Sec. 4.1.4) (5 Marks)

- Bayesian Information Criterion also known as Schwarz criterion is a criterion for selection of an appropriate model out of many available finite set of models. It is partially based on the likelihood function and is also related to Akaike Information Criterion (AIC).
- BIC is used to solve the problem of the overfitting faced by models it does so by introducing the penalty term for the number of parameters or features in the model. The value of penalty term in BIC is more than that used in AIC. BIC is used widely with the time series data set or linear regression.

Mathematical Formulation of BIC

- BIC assumes that the data distribution is done in the exponential family. Following are various parameters and the formula for BIC :

x = the observed data;

n = number of data points in the observed data x or the sample size;

k = number of free parameters to be estimated if the estimated model is considered to be linear regression then k here is the number of regressors and intercepts.

p(x|k) = It is the probability of the observed data for the given number of parameters or in other words it is the likelihood of the parameters for the given dataset.

L = maximum value of Likelihood function for the estimated model.

Formula for BIC

$$ln \cdot p(x \mid k) \approx BIC = -2 \cdot ln \cdot L + k \cdot ln(n)$$

Characteristics of BIC

1. BIC is independent of the constant.
2. It is used to measure the efficiency of model consisting parameters in terms of the predicting data.
3. When there are large numbers of parameters BIC penalized these parameters and hence reduce the complexity of the model.
4. The value obtained after BIC is equal to the minimum description length of that criteria but it has an negative sign.
5. BIC can be used to choose number of cluster depending on the complexity present in the dataset.
6. It is similar to other penalized methods such as RIC and AIC.

Syllabus Topic : Cross Validation

4.1.5 Cross Validation

Q. 4.1.5 Explain cross validation. (Ref. Sec. 4.1.5) (5 Marks)

- It is a technique to evaluate predictive models by recursively making partitions of the original sample into a training set to train the model, and a test set to evaluate it.
- Cross validation is a statistical method used to estimate the skill of machine learning models.
- Cross validation is also known as rotation estimation or out of sample testing. One round of cross-validation involves portioning a sample of data into various subsets and performs the analysis on one subset also known as training set and validating the analysis on other subsets know as the validation set of testing set.
- In order to reduce the variability many methods involves number of rounds of cross validation using different partitions and the validation results are combined to give an estimate of the model's predictive performance.

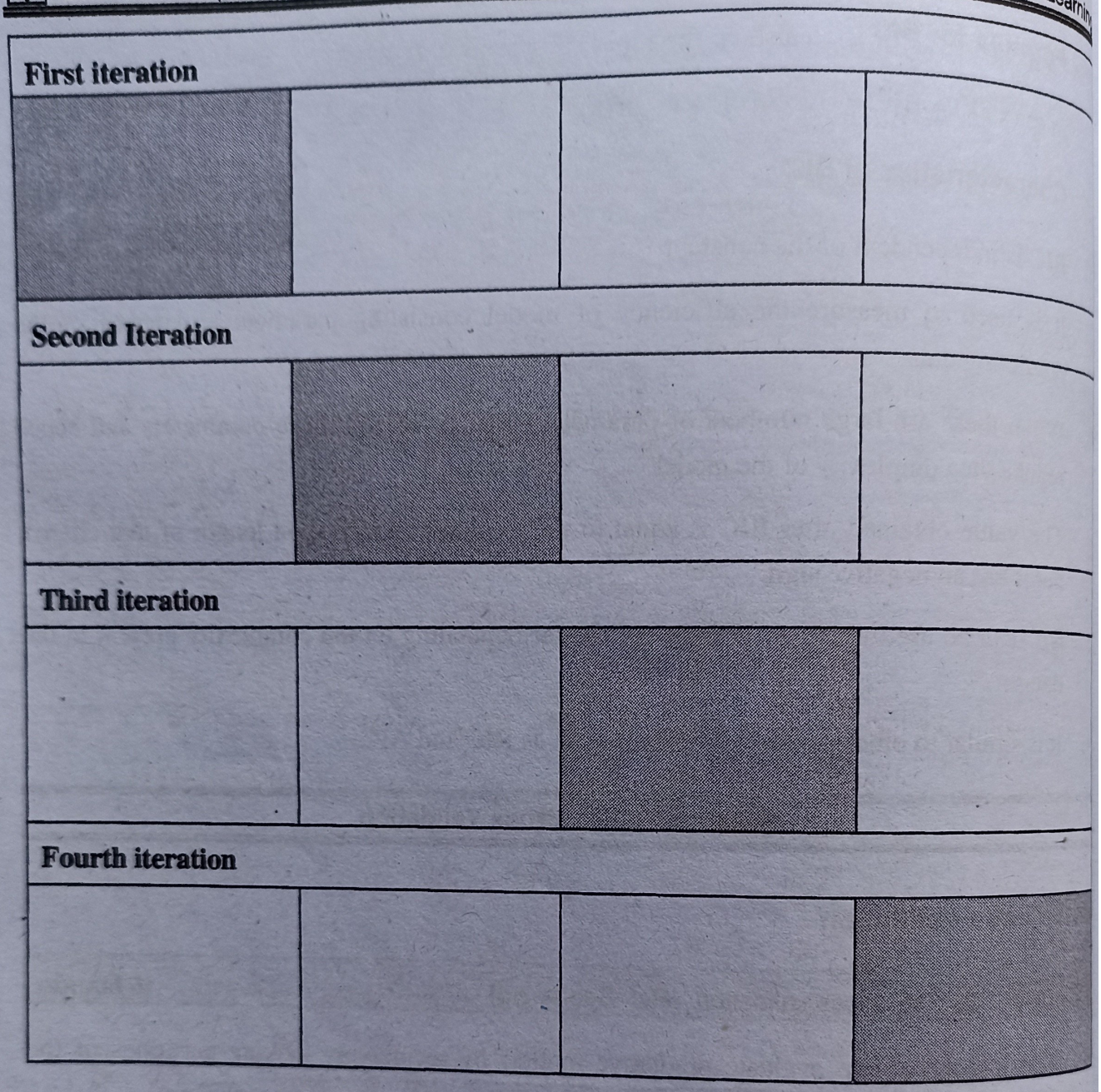

First iteration
Second Iteration
Third iteration
Fourth iteration
Training data set
Testing data set

Fig.4.1.2

- From Fig. 4.1.2 it is clear how the input samples are selected for performing the training and testing in every iteration. The above method can be called as 4 fold cross validation as the process of cross validation is applied 4 times.

Syllabus Topic : Ridge Regression

4.1.6 Ridge Regression

Q. 4.1.6 Explain ridge regression. (Ref. Sec. 4.1.6) (5 Marks)

- Ridge regression is a technique, which comes into picture when the data suffers from multicollinearity (which simply means that independent variables are highly correlated).
- In multicollinearity concept, even though the least squares estimates (OLS) are unbiased, their variances are large which in turn return results in the deviation of the observed value far from the true value.
- By adding a degree of bias to the regression estimates, ridge regression is able to reduce the standard errors.
- The Equation (4.1.3) for linear regression

$$y = a + b * x \quad ...(4.1.3)$$

- Equation (4.1.3) also has an error term. The complete equation becomes:
- y = a + b * x + e (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value].

$$=> y = a + y = a + b_1x_1 + b_2x_2 + ... + e \quad ...(4.1.4)$$

Equation (4.1.4) is used for multiple independent variables.

- In a linear equation, It is possible to fragment prediction errors into two sub-components. The first component is due to the biased and second the second component is due to the variance. Prediction error mostly occurs due to any one of these two or both components.
- Here, we will discuss the error caused due to variance.
- Ridge regression solves the multicollinearity problem through shrinkage parameter λ (lambda). Look at the Equation (4.1.5).

$$\underset{\beta \in \mathbb{R}p}{\operatorname{argmin}} \underbrace{\| y - x\beta \|_2^2}_{\text{Loss}} + \underbrace{\lambda \| \beta \|_2^2}_{\text{Penalty}} \quad \ldots(4.1.5)$$

- In Equation (4.1.5), we have two components. First one is least square term and another one is lambda of the summation of β^2 (beta- square) where β is the coefficient. This is added to the least square term in order to shrink the parameter to have a very low variance.

Important Points to be remembered about Ridge Regression

- In this normality is not to be assumed, the assumptions of this regression are same as least squared regression.
- It shrinks the value of coefficients but does not reach zero, which suggests no feature selection feature.
- This is a type of regularization method and uses L2 regularization.

Syllabus Topic : Penalized Regression E.g. LASSO

4.1.7 Lasso Regression

Q. 4.1.7 Explain lasso regression. (Ref. Sec. 4.1.7) **(5 Marks)**

- LASSO stands for Least Absolute Shrinkage and Selection Operator.
- Lasso regression is a type of linear regression that uses shrinkage, where data values are shrunk towards a central point, like the mean.
- The lasso procedure encourages simple, sparse models. This regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection elimination.
- It is a regression analysis method that performs both variable selection and regularization. Lasso was introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided covariates for use in the final model rather than using all of them.
- Lasso is basically an alternative to the classic least squares estimate to avoid many of the problems with overfitting when we have a large number of independent variables.
- Lasso regression is one of the regularization methods that creates parsimonious models which is to accomplishes a desired level of explanation or prediction with as few predictor

variables as possible in the presence of large number of features, where large means either of the below two things :

1. Large enough to enhance the tendency of the model to over-fit. Minimum ten variables can cause overfitting.
2. Large enough to cause computational challenges. This situation can arise in case of millions or billions of features.

- Lasso regression performs **L1 regularization** which adds a factor of sum of absolute value of coefficients in the optimization objective. Thus, lasso regression optimizes the following :

  Objective = RSS + α * (sum of absolute value of coefficients).

- Here, α (alpha) works similar to that of ridge and provides a trade-off between balancing RSS and magnitude of coefficients. Like that of ridge, α can take various values. Let's iterate it here briefly

  1. $\alpha = 0$: Same coefficients as simple linear regression.
  2. $\alpha = \infty$: All coefficients zero (same logic as before).
  3. $0 < \alpha < \infty$: coefficients between 0 and that of simple linear regression.

- The objective function to minimize is,

$$\min_{\omega} \frac{1}{2n_{samples}} \| X\omega - y \|_2^2 + \alpha \| \omega \|_1 \quad \text{...(4.1.6)}$$

- $\alpha \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of α, the greater the amount of shrinkage.
- The lasso estimate thus solves the minimization of the least-squares penalty with $\alpha \| \omega \|_1$ added, where α is a constant and $\| \omega \| 1$ is the l_1-norm of the parameter vector.

4.2 Exam Pack (Review Questions)

Syllabus Topic : Introduction To Model Selection

Q. 1 List and explain features of model selection. (Refer Section 4.1) **(5 Marks)**

Syllabus Topic : Regularization

Q. 2 Explain regularization. (Refer Section 4.1.1) **(5 Marks)**

➜ **2. Aggregation**

- Aggregation is similar to summarization operation on the data. Data cubes are generated by aggregation process.

➜ **3. Generalization**

- Generalization is similar to categorization where the hierarchy of data is formed.

➜ **4. Normalization**

- Normalization process scales the attribute values to certain upper or lower value or sometime it convert the attribute values to certain range like −1 or +1 etc.

➜ **5. Attribute construction**

- Attribute construction also know as feature construction is used for creation of new attribute values from the other similar values present in the data set.

Syllabus Topic : Dimension Reduction

5.2 Dimension Reduction

Q. 5.2.1 Explain dimensionality reduction with example. (Ref. Sec. 5.2) **(5 Marks)**

- In machine learning algorithms during classification, there may be case when we come across "N" number of dimensions or features /parameters/attributes.
- All these features many times will classify the data into wrong classes due to certain values of the number of features.
- The motivation behind dimensionality reduction or dimension reduction is to cut down (eliminate/remove) such unwanted (not /less important) dimensions or features which will finally classify the dataset into correct class. More number of features can also lead to complex classification.
- Dimensionality reduction can also referred to the process of converting a set of data having vase dimensions into the data with lesser dimensions ensuring that it provides the same or similar information.
- The Fig. 5.2.1 shows 2 dimensions x1 and x2, which are let us say measurements of several object in cm (x1) and inches (x2).

Unit III

CHAPTER 5

Data Transformations

5.1 Introduction

Q. 5.1.1 Explain the concept of data transformation. (Ref. Sec. 5.1) (5 Marks)

Data Transformation is a process of data normalization, aggregation, summarization and generalization. In data transformation, data is transformed or converted into specific format needed for mining.

Various methods involved in transformation are :

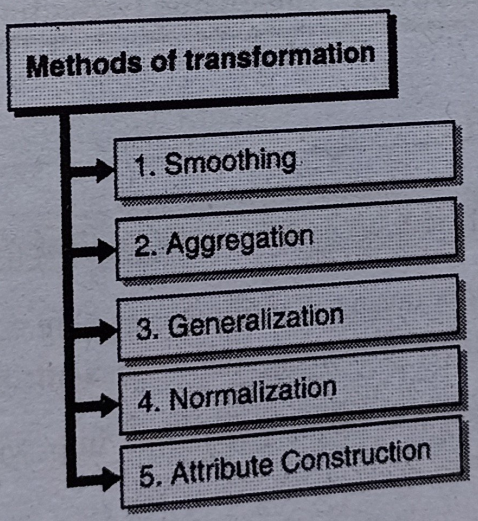


Fig. 5.1.1 : Methods of transformation

1. Smoothing

- Smoothing deals with removal of noise from data and it uses various techniques such as binning, clustering and regression.

☛ Syllabus Topic : Bias and Variance Trade Off E.g. Parsimony

Q. 3 Explain bias and variance trade off in detail with help of an example. (Refer Section 4.1.2) **(5 Marks)**

☛ Syllabus Topic : AIC

Q. 4 Explain AIC in detail with its mathematical formula. (Refer Section 4.1.3) **(5 Marks)**

☛ Syllabus Topic : BIC

Q. 5 Explain BIC in detail with its mathematical formula. (Refer Section 4.1.4) **(5 Marks)**

☛ Syllabus Topic : Cross Validation

Q. 6 Explain cross validation. (Refer Section 4.1.5) **(5 Marks)**

☛ Syllabus Topic : Ridge Regression

Q. 7 Explain ridge regression. (Refer Section 4.1.6) **(5 Marks)**

☛ Syllabus Topic : Penalized Regression E.g. LASSO

Q. 8 Explain lasso regression. (Refer Section 4.1.7) **(5 Marks)**

Chapter Ends...

❑❑❑

- Now, if you were to use both these dimensions in machine learning, they will convey similar information and introduce a lot of noise in system, so you are better of just using one dimension. Here we have converted the dimension of data from 2D (from x1 and x2) to 1D (z1), which has made the data relatively easier to explain.

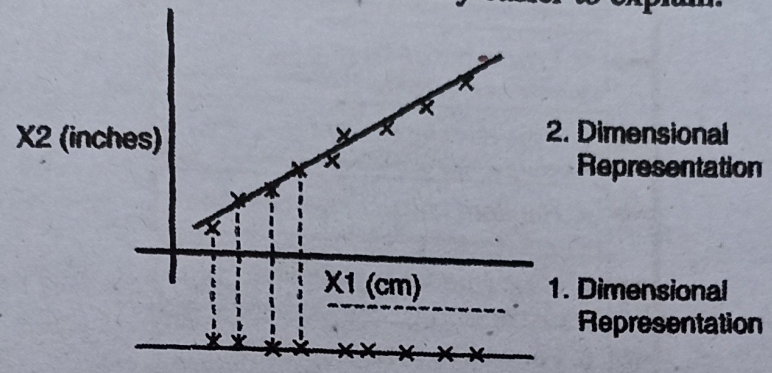


Fig. 5.2.1

- The process of dimensionality reduction can be divided into mainly two types :

1. Feature Selection	2. Feature Extraction

➜ 1. Feature selection

- In this approach the subset of the original set of variables/ features /attributes are formed. These subsets are modelled to the classification problem and the best subset is chosen which will classify the dataset into the proper classes.

There are three ways to do this

1. Filter method
2. Wrapper method
3. Embedded method.

➜ 2. Feature Extraction

In feature extraction method transformation of the data from a high dimensional space to fewer dimension space is done. The best technique for feature extraction is Principal Component Analysis (PCA). PCA algorithm works with linear data set.

5.2.1 Methods for Dimensionality Reduction

Q. 5.2.2 Explain various methods for dimensionality reduction. (Ref. Sec. 5.2.1) (5 Marks)

- There are various methods for dimensionality reduction technique which are as follow :

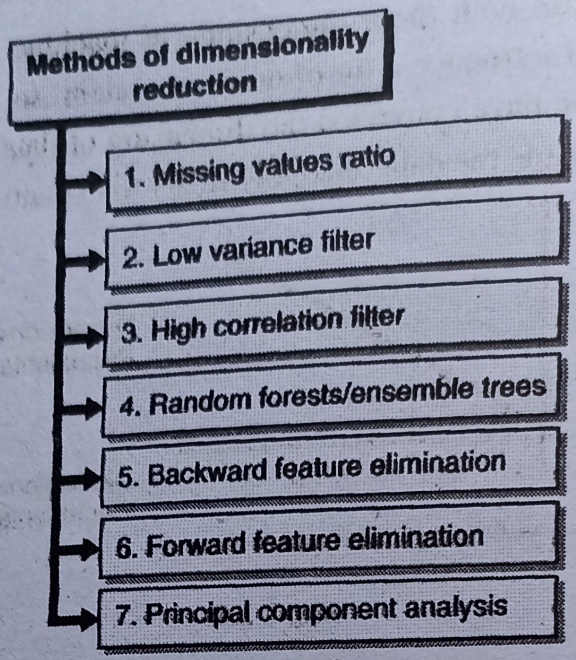


Fig. 5.2.2 : Methods of dimensionality reduction

➜ 1. Missing values ratio

- From a given dataset if some columns or attribute or feature has too many missing values then it is not much useful feature.
- By considering certain threshold values like for 1000 records if threshold value is considered as 50%, means if any column having more than or equal to 50% missing data then it can be considered useless and removed.

➜ 2. Low variance filter

- In this technique if suppose there are attributes/ features with similar type of data that is the values of that column does not vary at great extend or the values are in same range then that feature again becomes useless as there is low variation or low variance in the value.
- These types of columns can be identified by considering some threshold like if there are 50 % of those columns having similar range of the value then they are useless and such columns can be removed.

➜ 3. High correlation filter

- In the dataset there might be some columns which may have same values in different format.

Syllabus Topic : Feature Extraction

5.3 Feature Extraction

Q. 5.3.1 Explain Feature Extraction using PCA (Ref. Sec. 5.3) (5 Marks)

In feature extraction method transformation of the data from a high dimensional space to fewer dimension space is done. The best technique for feature extraction is Principal Component Analysis (PCA). PCA algorithm works with linear data set. PCA is explained in detail in chapter no 7.

Syllabus Topic : Smoothing and Aggregating

5.4 Smoothing and Aggregating

5.4.1 Smoothing

Q. 5.4.1 Explain Smoothing with various techniques used for it. (Ref. Sec. 5.4.1)(5 Marks)

Smoothing deals with removal of noise from data .Noise is unwanted quantity in signal and it is random. Various techniques included in smoothing are as follow :

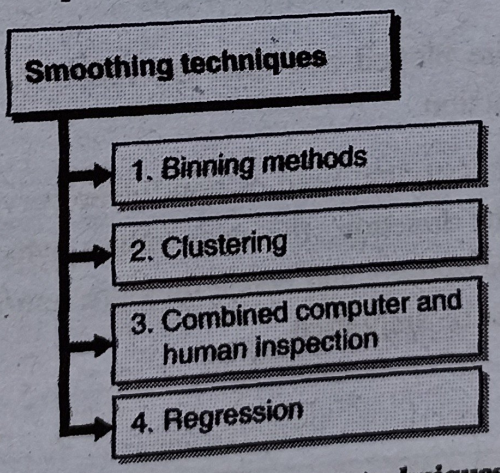


Fig. 5.4.1 : Smoothing techniques

➔ **1. Binning methods**

- In Binning method to smooth the data, firstly the data is needed to be sorted and for sorting the data the neighbourhood values or the surrounded values are compared.

Note : Both backward and forward feature elimination techniques are time consuming and computationally expensive. They are applicable to dataset with less number of features.

➜ 7. Principal component analysis

- It is a dimension reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.
- It is a mathematical procedure that transforms a number of possibly or correlated variables into a smaller number of uncorrelated variables called as principal components.PCA will be explained in detail in the last section of this chapter.
- Apart from these techniques Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA) are also dimensionality reduction techniques.

5.2.2 Advantages of Dimensionality Reduction

- It helps for data compression and due to this the storage space can be reduced.
- As the numbers of features are reduced the computation time is reduced.
- When there are "N" numbers of features, some of them can be giving the same data in different format (Redundant Data), removing such duplications will create the unique data set.
- Reducing the dimension also means reducing from 3D to 2D in some cases for the given dataset it is easier to study a 2D model then a 3D model.
- Due dimensionality reduction the time as well as space required for performing the operations is reduced and hence we can say that dimensionality reduction increases the performance of the classification problem.

5.2.3 Disadvantages of Dimensionality Reduction

- Sometimes removing some feature can lead to data loss, it happens if the user is new to particular data set and he might have less idea about the data set.
- Dimensionality reduction technique like PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA method fails with few dataset where the mean and covariance are not enough and also for all the type of data format it is not possible to find the covariance and mean.
- As there is no fix methodology to remove certain feature many can get different classes even for same data set.

- For example suppose there is a dataset used for prediction for house price and has say some 50 features or attributes, now suppose if there is one row with area of house in square feet and one row with the area of house in square meter, then these both the columns actually provides same data and hence only one can kept and other can be removed.
- In order to keep the track of all such columns correlation coefficient is removed between the column by using pearson's product moment coefficient and the Pearson's Chi square value techniques.

➜ 4. Random forests/ensemble trees

- Decision trees or random Forest trees are useful methods for the selection of the best features.
- Set of sub-trees are made based on the target variable or feature and then the sub-tree providing highest gain for that target attribute or feature is selected as the best classifier so out of "n" sub-trees formed with various features only one with the highest gain is selected as the best classifier all others are removed which leads to reduction of many features.

➜ 5. Backward feature elimination

- In this technique initially all the features are considered as the input features and then one by one some features are removed and its effect is seen, if removal of any feature leads to increase in the gain then that feature is permanently removed for certain classification and if that feature leads to decrease in the gain that feature is again added back to the classifier.
- Backward feature elimination is an iterative process, on every iteration it is checked for where to remove or keep the feature.
- The process is ongoing and stops when maximum gain or best classification is done.

➜ 6. Forward feature elimination

- This technique is completely reverse of the backward feature elimination. It has only one similarity that it is also iterative process.
- Here initially only one feature is considered, on every iteration one more feature is added, after adding that feature if the system provides increase in gain than that feature is kept else that feature is eliminated.

- The sorted values are then distributed into number of buckets also known as bins. In order to group the values in the bins or buckets binning method compare or consult with the neighbourhood values and hence they perform local smoothing.

➜ **2. Clustering**

- In clustering some seed values are considered at the initial state and then the remaining data values from the set are then accordingly placed into the cluster.
- Each cluster will have specific set of values or data which is closest to the initially selected seed value for specific cluster.
- The value which does not come under any cluster is the outlier or anomaly and it is considered to be noise, these outliers are then removed and data is smoothed.

➜ **3. Combined computer and human inspection**

- Many times combined effort of human and computer can be used for removing the outlier. In an application computer can be used to detect the pattern and then human can identify which to be considered and which to be not.
- The patterns those are rejected can be considered as the outliers and hence can be removed and the data can be smoothed.

➜ **4. Regression**

- In regression data values are plotted and then the function is evaluated which will fit maximum values on that function.
- Such a function can be linear (straight line) or non linear(curve) accordingly data is fitted on the shape defined by the function, the data values which are not covered by the function are considered to be outliers and hence they are removed as the unwanted/ noisy values and hence the data can be smoothed.

5.4.2 Aggregation

Q. 5.4.2 What is Aggregation in data transformation ? (Ref. Sec. 5.4.2) (5 Marks)

- Aggregation is similar to summarization operation on the data. Data cubes are generated by aggregation process.
- For example in a retail shop daily sales data can be aggregated and the monthly and annual total sale amount can be calculated, which can further be used to identify profit and loss for the sales in that shop.

- Data aggregation generally is used on large amount of data which does not give information as whole but after aggregation can give certain information.
- Data aggregation is mainly used for gathering, utilization and presentation of the data which is available and can be presented on the internet.
- Aggregation is generally done in order to get the summary and by using it the reports can be generated as and when needed. The tools or the functions used for aggregation should be well defined any wrong assumption can lead to wrong summary or report.
- Aggregation can be done as and when needed but there are certain time intervals when the aggregation becomes compulsions.

Reporting period

It is the time when the reports are to be generated and presented to the higher authorities in the business. Depending on the type of business the reporting period can be on daily, weekly, monthly, quarterly and yearly basis.

Granularity

Granularity is a period over which data points for a particular resource or set of resources is collected. For example if time series or real time data is used and you need to get the information about the particular share during 10:00 am to 10:05 am then this period of 5 minutes is said to be granule. Granularity can be in any range in minutes, in hours, in days, in weeks or even in months.

Polling period

Polling period provides time duration which determines how often a particular data has been sampled. Polling period and granularity period can be same or different. Consider the granularity period is of 20 minutes and polling period is say 5 minutes then it can be concluded that in 20 minutes of the granularity period (20/5) 4 time sampling or aggregation is done.

5.5 Exam Pack (Review Questions)

Syllabus Topic : Dimension Reduction

Q. 1 Explain the concept of data transformation. (Refer Section 5.1) **(5 Marks)**

Q. 2 Explain dimensionality reduction with example. (Refer Section 5.2) **(5 Marks)**

Q. 3 Explain various methods for Dimensionality Reduction. (Refer Section 5.2.1) **(5 Marks)**

☛ **Syllabus Topic : Feature Extraction**

Q. 4 Explain Feature Extraction using PCA. (Refer Section 5.3) **(5 Marks)**

☛ **Syllabus Topic : Smoothing and Aggregating**

Q. 5 Explain Smoothing with various techniques used for it. (Refer Section 5.4.1) **(5 Marks)**

Q. 6 What is Aggregation in Data Transformation? (Refer Section 5.4.2) **(5 Marks)**

Chapter Ends...

❑❑❑

Least square method

- In regression for the given data set we find the best line which can fit the data samples. There are number of possibilities to draw the regression line, but to get a best fit line number of trails are to be made.
- Least square method is one of the important statistical techniques used to find the regression line or best fit for the given model. It is used to compare relation between two variables.

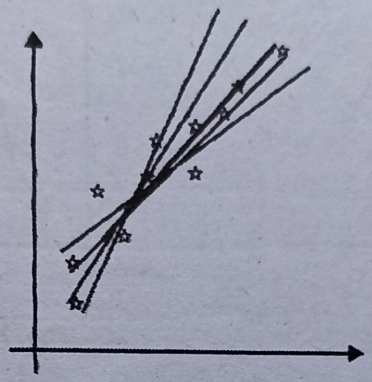


Fig. 6.2.1 : Possible regression lines

Steps to be followed for least square method are

Consider the data set as $(x_1, y_1)\ (x_2, y_2)\ (x_3, y_3)\ ....\ (x_n, y_n)$

Step 1 : Calculate the mean of the sample for x and y as,

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Step 2 : Calculate the slope of the line for best fit using the Equation (6.2.1)

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{X}) - (y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} \quad ...(6.2.1)$$

- A supervised technique used for predicting the value when the data is continuous or real valued is known as regression. Regression consists of dependent variable know as Y or output variable and independent variable known as X or input variable.
- Examples of the task in which the regression can be used are :
  - Predicting the house price.
  - Predicting age of person.
  - Predicting nationality of person.
  - Predicting stock price of the company etc.
- Some of the important features of linear regression are as below :
  - Linear regression is a fast and easy to model technique and is mainly useful when relationship is modelled and is not extremely complex.
  - Linear regression is not feasible for fewer amounts of data.
  - Linear regression is very sensitive to outliers.
  - Regression is based on hypothesis and it can be linear, quadratic, polynomial, non-linear etc.
  - Regression is divided mainly into two types which are :

1. Simple regression
2. Multiple regressions

➜ **1. Simple regression**

It has only one feature and it is simple to use and develop. Simple regression is further classified into simple linear regression and simple non-linear regression.

➜ **2. Multiple regressions**

It has two or more than features and it is little complex to develop with respect to simple regression.

Syllabus Topic : Linear Models

6.2.1 Linear Models

Least square method is used to find the best fit line in the linear models.

Unit III

CHAPTER 6

Supervised Learning

6.1 Introduction

- In supervised learning algorithms there is a trainer which trains the sample. This algorithm consists of a target/ outcome variable and the set of predictors.
- Target variable are dependent variables were as the predictors are independent variables. In supervised learning data is labelled.
- The trainer train the sample until the accuracy of the system is achieved. There are various algorithms in supervised learning which are as follow :

1. Regression
2. Classification
3. Decision tree
4. Random forest
5. KNN
6. Logistic regression
7. Support vector machine
8. Neural network etc.

Syllabus Topic : Regression

6.2 Regression

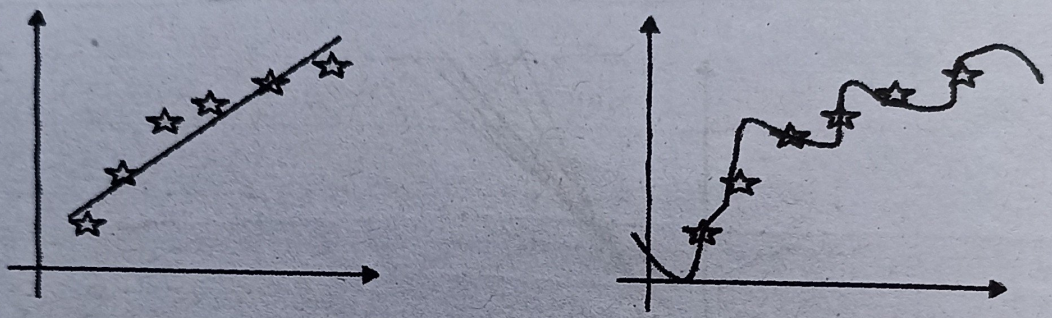
Q. 6.2.1 Explain Regression in detail with help of an example. (Ref. Sec. 6.2) (5 Marks)

Step 3 : Compute the y-intercept of the line.

$$b = \bar{Y} - m\bar{X} \qquad \ldots(6.2.2)$$

Step 4 : Use the slope m and the y-intercept b to form the Equation of the line.

Fig. 6.2.2 shows linear and non-linear plots for plotting the dataset. Fitting data on the line is the linear model where as fitting data on the curve is non linear model.



(a) Linear Model (b) Non-Linear Model

Fig. 6.2.2

Syllabus Topic : Regression Trees

6.2.2 Regression Trees

Q. 6.2.2 Explain regression tree with example. (Ref. Sec. 6.2.2) **(10 Marks)**

- Decision tree is a tree like structure which breaks down the dataset into smaller subsets. It is an iterative process and on every iteration tree increments itself. Decision tree mainly consists of nodes and links, nodes are further of two types decision nodes and leaf nodes.
- Decision nodes are the nodes which on some condition are further divided into various other decision nodes or leaf nodes. Leaf nodes are also known as terminals and they cannot be further divided. Decision trees can handle both categorical as well as numerical data.
- Decision tree when developed for regression is known as regression tree. Decision tree can be developed for the classification also and main difference in both is in regression tree standard deviation reduction is done where as in classification tree gain is calculated.
- Consider the Table 6.2.1 to understand the decision tree :

Now let us see the standard deviation reduction. Standard deviation reduction is generally based on the low value of standard deviation once a dataset is split on an attribute. Here the aim to find a decision tree is to identify the attribute which returns the highest value of standard deviation reduction.

Following are the steps carried out for forming the tree

Step 1: Calculate the standard deviation of the target.

As here target is hours played and hence the standard deviation for Hours played = 9.32(calculated earlier).

Step 2: Calculate the standard deviation for each branch, and then find the standard deviation reduction.

For Outlook

		Hours Played (standard deviation)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR =1.66		

For Temperature

		Hours Played (standard deviation)
Temperature	Cool	10.51
	Hot	8.95
	Mild	7.65
SDR = 0.17		

For Humidity

		Hours Played (standard deviation)
Humidity	High	9.36
	Normal	8.37
SDR = 0.28		

For Windy

		Hours Played (standard deviation)
Windy	False	7.87
	True	10.59
SDR = 0.29		

- Standard deviation is used for calculating the homogeneity of a numerical sample, in case if in any problem all numerical samples are homogenous then the standard deviation is zero. Standard deviation can be calculated for single or multiple attributes.

☛ Standard deviation for single attribute

- For Table 6.2.1 let's consider the hours played attribute.
- Now from the data table the total no of values in the hours played attributes are 14 and hence count = n = 14.

Average is calculated as $\frac{\text{sum of the samples}}{\text{no of samples (count)}}$ and hence in this case,

$$\text{Average} = \bar{x} = 39.8$$

$$\text{Standard deviation} = s = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = 9.32$$

Next the Coefficient of Deviation (CV) is calculated it is used to decide when the branching should be stop.

$$CV = \frac{s}{x} * 100\ \% = 23\%.$$

☛ Standard deviation of multiple attributes

Let us consider here two attributes target and predictor for this standard deviation is calculated as follow,

$$S(T, X) = \sum_{c \in X} P(c)\, S(c)$$

		Hours Played (standard deviation)	Count
Outlook	Overcast	3.49	4
	Rainy	7.78	5
	Sunny	10.87	5
			14

Hence the standard deviation here is calculated as,

$$S(\text{Hours Played, Outlook}) = P(\text{Sunny})*S(\text{Sunny}) + P(\text{Overcast})*S(\text{Overcast}) + P(\text{Rainy})*S(\text{Rainy})$$

$$= (4/14)*3.49 + (5/14)*7.78 + (5/14)*10.87 = 7.66$$

Table 6.2.1

Predictors				Target
Outlook	Temp	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

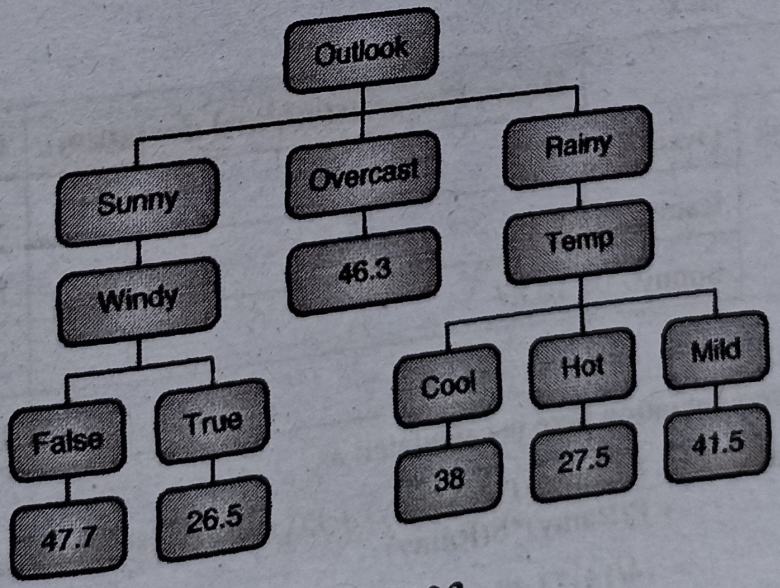


Fig. 6.2.3

Outlook : Sunny

Temp	Humidity	Windy	Hours Played	Hours Played (standard deviation)	Hours Played (Average)	Hours Played (Coefficient of deviation)
Mild	High	FALSE	45	10.87	39.2	28%
Cool	Normal	FALSE	52			
Cool	Normal	TRUE	23			
Mild	Normal	FALSE	46			
Mild	High	TRUE	30			

Now for individual attributes from above table we calculate standard deviation reduction and select the best attribute.

For Windy

	Hours Played (standard deviation)	Count	
Windy	False	3.09	3
	True	3.50	2

$$SDR = 10.87 - ((3/5) * 3.09 + (2/5) * 3.5) = 7.62$$

For Humidity

		Hours Played (standard deviation)	Count
Humidity	High	7.50	2
	Normal	12.50	3

$$SDR = 10.87 - ((2/5) * 7.5 + (3/5) * 12.5) = 0.370$$

For Temperature

		Hours Played (standard deviation)	Count
Temperature	Cool	14.50	2
	Mild	7.32	3

$$SDR = 10.87 - ((2/5) * 14.50 + (3/5) * 7.32) = 0.678$$

The above attribute cannot be further divided into further branches as the count is 3 or less than 3 and hence the leaf nodes are generated.

For Outlook : Rainy

Outlook	Temp	Humidity	Windy	Hours Played
Rainy	Hot	High	FALSE	25
Rainy	Hot	High	TRUE	30
Rainy	Mild	High	FALSE	35
Rainy	Cool	Normal	FALSE	38
Rainy	Mild	Normal	TRUE	48

Note : Here we will use the coefficient of deviation for a branch which is smaller than a certain threshold, consider here the threshold as 10% which actually means the two less instances are present for certain branch.

Step 4.2 : From the above tables it is seen that overcast subset does not need any further splitting as it has the value for coefficient of deviation as 8% which is less than the assumed threshold which is 10%.

		Hours Played (standard deviation)	Hours Played (Average)	Hours Played (Coefficient of deviation)	Count
Outlook	Overcast	3.49	46.3	8%	4
	Rainy	7.78	35.2	22%	5
	Sunny	10.87	39.2	28%	5

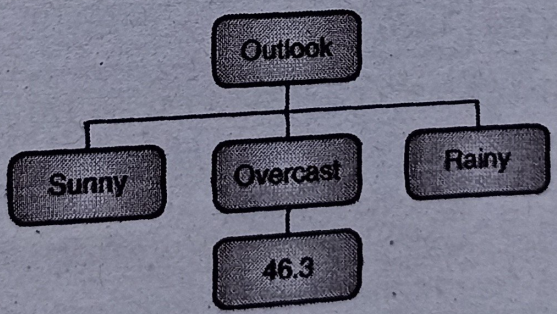


Fig. 6.2.4

Step 4.3 : Split further the branches which are more than threshold.

As seen from the above table sunny branch to further split.

Now combined standard deviation reduction mentioned in the above tables is calculated as,

$$SDR(T, X) = S(T) - (T, X)$$

SDR (Hours Played, Outlook) = S (Hours Played) - S (Hours Played, Outlook)

$$= 9.32 - 7.66$$

$$= 1.66$$

Step 3 : Choose the decision node with the largest value of standard deviation reduction.

From all the tables in step 2 standard deviation reductions for hours played and outlook is maximum and hence it is selected.

		Hours Played (standard deviation)
Outlook	Overcast	3.49
	Rainy	7.78
	Sunny	10.87
SDR =1.66		

Step 4 .1 : In this step the dataset is divided with respect to the values of selected attribute. It is a recursive process and in it is done until complete data is processed.

For Outlook : Sunny

Outlook	Temp	Humidity	Windy	Hours Played
Sunny	Mild	High	FALSE	45
Sunny	Cool	Normal	FALSE	52
Sunny	Cool	Normal	TRUE	23
Sunny	Mild	Normal	FALSE	46
Sunny	Mild	High	TRUE	30

For Outlook : Overcast

Outlook	Temp	Humidity	Windy	Hours Played
Overcast	Hot	High	FALSE	46
Overcast	Cool	Normal	TRUE	43
Overcast	Mild	High	TRUE	52
Overcast	Hot	Normal	FALSE	44

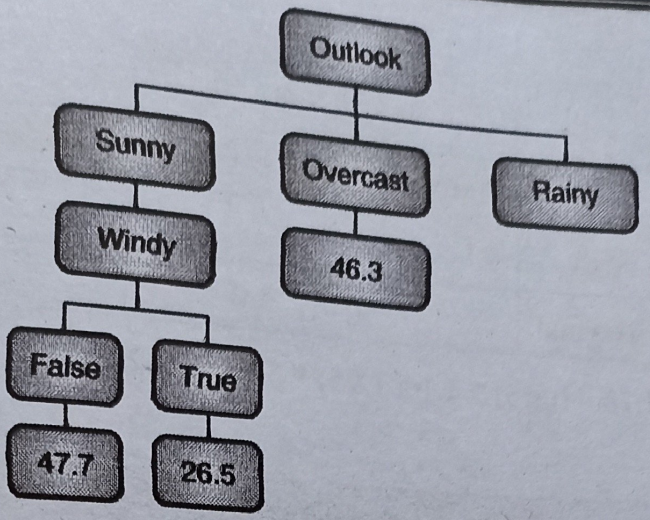


Fig. 6.2.5

Now the rainy branch with coefficient of deviation 22% is considered for split.

Outlook : Rainy

Temperature	Humidity	Windy	Hours Played	Hours Played (standard deviation)	Hours Played (Average)	Hours Played (Coefficient of deviation)
Mild	High	FALSE	25	7.78	35.2	22%
Cool	Normal	FALSE	30			
Cool	Normal	TRUE	35			
Mild	Normal	FALSE	38			
Mild	High	TRUE	48			

Now for Individual attributes from above table we calculate Standard deviation reduction and select the best attribute.

For Windy

		Hours Played (standard deviation)	Count
Windy	False	5.6	3
	True	9.0	2

$$SDR = 7.78 - ((3/5) * 5.6 + (2/5) * 9.0) = 0.82$$

For Humidity

		Hours Played (standard deviation)	Count
Humidity	High	4.1	3
	Normal	5.0	2

$$SDR = 7.78 - ((3/5) * 4.1 + (2/5) * 5.0) = 3.32$$

For Temperature

		Hours Played (standard deviation)	Count
Temperature	Cool	0	1
	Mild	6.5	2
	Hot	2.5	2

$$SDR = 7.78 - ((1/5) * 0 + (2/5) * 6.5 + (2/5) * 2.5) = 4.18$$

The above attribute also cannot be further divided and hence the splitting process has to be stop.

Final decision tree shown in Fig 6.2.6,

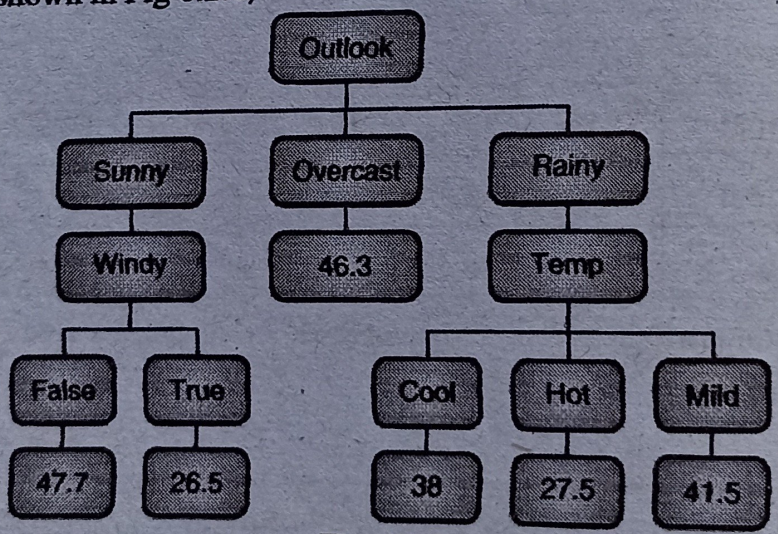


Fig. 6.2.6

Syllabus Topic : Time Series Analysis

6.2.3 Time Series Analysis

Q. 6.2.3 Explain time series analysis with its components. (Ref. Sec. 6.2.3) **(5 Marks)**

- Forecasting methods differ from time and situation.

Classification of various forecasting methods

1. **Model based**
   - Linear regression.
   - Auto regressive models.
   - ARMA.
   - Logistic regression.
   - Econometric models.
2. **Data Driven**
   - Naive forecast.
   - Smoothing.
   - Neural nets.

Syllabus Topic : Classification

6.3 Classification

Q. 6.3.1 Explain classification in detail with help of an example. (Ref. Sec. 6.3) (5 Marks)

- Supervised learning is a type of machine learning technique in which a teacher or a reference is required.
- Classification belongs to supervised learning technique. Classification is considered to be most common task in machine learning. In classification there are finite numbers of classes and these classes can be denoted as C_1, C_2 ... C_n.
- Input in the classification can be of any size and the aim of the classification is to map that input or samples to the respective classes. Classification deals with assigning various observations into separate categories.

Binary classification

- As the name Binary (means two) if there are only two possible classes then it is said to be Binary Classification. In a simplest case there can be two classes which can be represented as positive and negative or +1 or −1.

- Most widely used time series analysis model is Autoregressive Moving Average (ARMA) model. It is divided into two parts which are Autoregressive (AR) and Moving Average (MA) part.
- The model is also widely referred as ARMA (p, q) model where p and q are orders of auto regressive and moving average parts respectively.

AR (p) is given by using the formula,

$$X_t = c \sum_{i=1}^{P} \phi_i X_{t-i} + \varepsilon_t$$

Where, $\{\phi_1, \ldots, \phi_p\}$ are the parameters to be estimated and c is constant where as variable ε_t is white noise. MA (q) is given by the formula,

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{P} \theta_i \in_{t-i}$$

Where, $\{\theta_i .. \theta_i\}$ are the parameters of the model and μ is the expectation of X_t and the ε_t, ε_{t-1}, ... refers to white noise error.

Syllabus Topic : Forecasting

6.2.4 Forecasting

Q. 6.2.4 What do you mean by Forecasting ? (Ref. Sec. 6.2.4) (5 Marks)

- It is the process of making prediction of the future based on the present and the past data most commonly by analysis of trends. Prediction is similar term to the forecasting but not exactly the same.

☛ **Steps involved in forecasting are as follow**

1. Define the goal or business objective.
2. Get the required data.
3. Explore and visualise the series.
4. Pre-process the data.
5. Partition the series.
6. Apply suitable forecasting method.
7. Evaluate and compare the performance of the system.
8. Implement the final forecast system.

- Time series analysis consists of set of methods used to analyze various data facts or statistics from various characteristics of the data.
- Time series analysis is used for continuous data for example economic growth of an organization, share prices, sales, temperature, weather etc.
- Time series analysis model has time factor "t" as an independent variable and the target is a dependent variable denoted by Y_t. The output from the time series model is a predicted value of Y at the given time t. Time series is the process of recording of the data at regular interval of time.

Time series components

There are various time series components which are as follow :

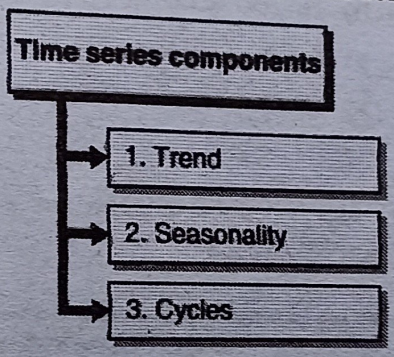


Fig. 6.2.7 : Time series components

1. Trend

- It is considered to behaviour of the feature at a particular amount of time, it can be categorized as increasing trend, decreasing trend or constant trend.
- When the particular feature value increases in particular amount of time it is increasing trend, similarly if it decreases it is decreasing trend and when it does not change over the period of time then it is constant trend.

2. Seasonality

Seasonality is a pattern which repeats at the constant frequency. For example here the demand for the umbrellas will be in rainy season only.

3. Cycles

- Cycles are type of seasonality pattern but it does not repeat at regular frequency. Cycle can be generally considered as the task completion time.
- For example, in iterative model of software engineering every iteration can have different time requirement, but the every task has to undergo all stages in a single iteration.