APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

1

Unit 3: Applied Machine Learning

Syllabus of Unit-III: Classification Techniques: (10 Hours)

- Naïve Bayes Classification:
  - ❖ Fitting Multivariate Bernoulli Distribution,
  - ❖ Gaussian Distribution and
  - ❖ Multinomial Distribution
- K-Nearest Neighbors
- Classification Trees
- Linear Discriminant Analysis
- Support Vector Machines:
  - ❖ Hard Margin and Soft Margin,
  - ❖ Kernels and Kernel Trick
- Evaluation Measures for Classification Techniques

Unit 3: Decision Tree Classification

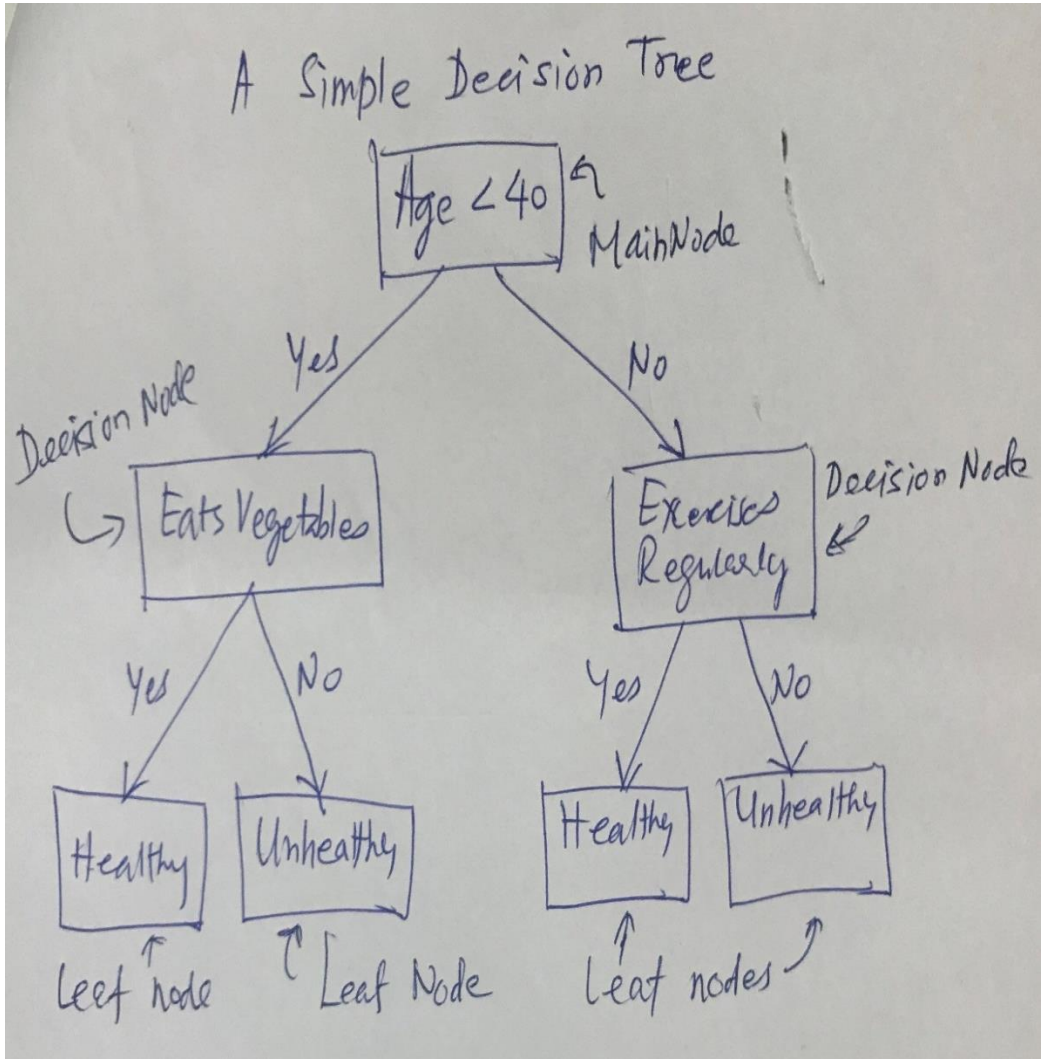
Decision Tree Classification:

- Among the various algorithms for ML problems, the best and most efficient algorithm for a any given data set is the main point of performance while developing a good model.
- A Decision Tree (DT) is supervised ML algorithm which can be used for both classification and regression.
- DT is one such algorithm for good classification and regression both.
- DTs usually implement exactly the human thinking ability while making a decision.
- The logic behind the DT can be easily understood as it shows a flow type structure/inverted tree structure.
- This makes it easy to visualize and extract information.

Unit 3: Decision Tree Classification

Decision Tree Classification:

- A decision Tree (DT) is like a inverted tree with nodes, which split data into branches till it achieves a threshold value.
- It contains a main node, decision nodes, edges and leaf nodes.
- It can have multiple level of decision nodes.
- The number of levels it contains decision nodes is called The Depth of the Tree. Leaf node level is included in depth calculation.



Unit 3: Decision Tree Classification

Building a Decision Tree:

- Start with a selection of a best attribute, which has to be selected as Root Node/Main node, and also attributes as Decision Nodes (for each level).
- This is done with the help of a technique known as Attribute Selection Measure (ASM).
- There are two techniques for ASM:

  (i) Information Gain, and (ii) Gini index.
- Information Gain technique is a measure of changes in Entropy after splitting or segmenting the data set based on a selected attribute.
- It tells how much information a feature/attribute provides.

Unit 3: Decision Tree Classification

Building Decision Tree: Information Gain

- The DT always tries to maximize the Information at every step of branching in its building process.
- A feature or attribute having the highest information gain splits the data first forming main node or root node.
- This is followed by other attributes in each branch in the decreasing order of information gain forming decision nodes.
- Information gain depends on the entropy and the weighted average entropy as:

  Info. Gain = Entropy(S) – Weighted Average Entropy(All Features)

Where Entropy(S) = - P(yes)*$\log_2$(P(yes)) – P(no)*$\log_2$(P(no))

Here, S = Total number of samples,

P(yes) = Probability of yes outcomes (i.e. belongs to the class),

P(no) = probability of no outcomes (i.e. does not belong to the class).

Unit 3: Decision Tree Classification

Building Decision Tree: Gini Index

- Gini Index is named after the name of a statistician who introduced this index.
- Gini index is a measure of impurity of a class while creating a Decision Tree.
- It is used mainly in CART algorithm (Classification And Regression Tree)
- An attribute having a low Gini index value is preferred to that having high Gini index value.
- It creates only Binary Splits. It is defined as:

  Gini index = 1 - $\sum_j (P_j^2)$, where P_j $are\, probability\, of\ jth\ attribute$.
- Probability is computed by normalizing the frequency distribution of attribute values

Unit 3: Decision Tree Classification

Building Decision Tree: Steps

1. Select the best feature using ASM and make it a Root Node.
2. Break the data set into smaller subsets based on a selected value of this attributer in the root node.
3. Continue the tree building process by repeating the above process recursively for each decision node until one of the following condition is achieved:
   (a) All examples/instances belonging to same attribute/value are grouped together.
   (b) There are no more attributes left out.
   (c) There are no more examples/instances are also left out.

Unit 3: Decision Tree Classification

Advantages of Decision Tree:

- It is simple to implement and it follows a flow chart type of structure; like human decision making process.
- It proves to be vey useful for decision-related problem;

  for example Decision Support System (DSS) development.
- It helps to find all possible outcomes for a given problem.
- There is very little (no need) for data-cleaning compared to other ML algorithms, as no mathematical modelling are used.
- DT can handle both numerical as well as categorical data values.

Unit 3: Decision Tree Classification

Disadvantages of Decision Tree:

- Too many layers (i.e. large depth) of the Decision Tree makes it extremely complex system.
- Pruning of the tree branches may be required by reducing the branches, and levels in tree.
- First poorly correlated attributes should be pruned.
- The many layers (i.e. higher depth) may result in Overfitting.
- As the number of classes increase, the computational complexity of the Decision Tree increases.

Unit 3: Decision Tree Classification

Random Forest Algorithms:

- The issue of overfitting can be eased or resolved to a large extent by using Random Forest algorithm.
- Random Forest is a conglomerate of several Decision Tress.
- Decision Trees are the models used to make predictions by going through each and every feature in the data one-by-one.
- This means the Random Forest can be defined as a collection of multiple Decision Trees.
- Random Forest are a collection of such Decision Trees being grouped together and trained together that use random orders of features in the data sets.
- Instead of relying on one decision tree, the Random Forest takes the decisions from each and every tree.
- Based on majority of prediction votes it gives the final prediction output.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

2

Unit 3: kNN Classification

kNN (k-Nearest Neighbour) Classifier:

- It is the simplest classification method, but extremely powerful one.
- kNN method of classification was first discovered in 1950s.
- Until 1960s, this method did not gain popularity as it was supposed to be compute intensive.
- With the availability of higher compute resources, it is widely used now.
- The kNN method is based on the philosophy of learning by analogy.
- Learning by analogy is by comparing a given test data example with training examples that are similar to it.
- In this kNN method, k stands for number of nearest neighbour samples.
- In kNN method, the unknown or unlabelled samples which needs to be predicted is judged on the basis of the training data sets elements.
- These training data samples are similar to the unknown data element.

Unit 3: kNN Classification

kNN Classifier: Feature Space concept

- The training data examples are described by m-attributes/features and it is represented as a point in the **m-Dimensional feature space**.
- The proximity/closeness of any given test data point, i.e. sample point, in the m-D feature space to the data points of different classes are compared.
- There are two proximity measures, viz. l_1 i.e. Manhattan distance and l_2 i.e. Euclidean distance.
- That is, for a given unknown example (of the test data), the kNN classifier searches the feature space for k-Number of nearest neighbour features.
- Assuming that the probability of the given data point/example is high, the test data point is classified as the class of k-Nearest Neighbours in the feature space.
- This asserts that the class of an unknown sample strongly resembles that of k Nearest Neighbour samples.

Unit 3: kNN Classification

kNN Classifier: Distance Measure

- The closeness is defined in terms of distance metric of Euclidean Distance.
- The Euclidean distance between two points in m-D feature space is given by:

  $\text{Dist}(X_1,X_2) = \sqrt{[\sum_{i=1}^{m}[(x_{1i} - x_{2i})^2\ ]}$,

  where X_1 and X_2 are the data points with co-ordinates x_{1j}, and x_{2j}, j = 1 to m.
- Note:

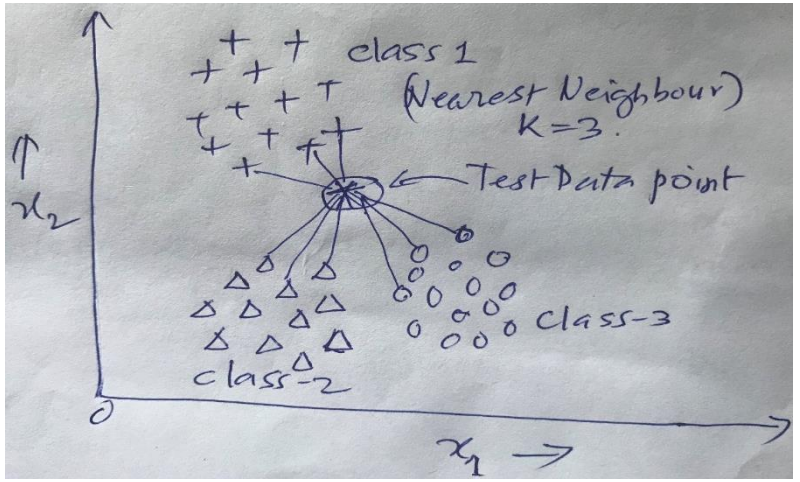
  (i) Always the positive (+ ve) square root is considered as distance measure or metric.

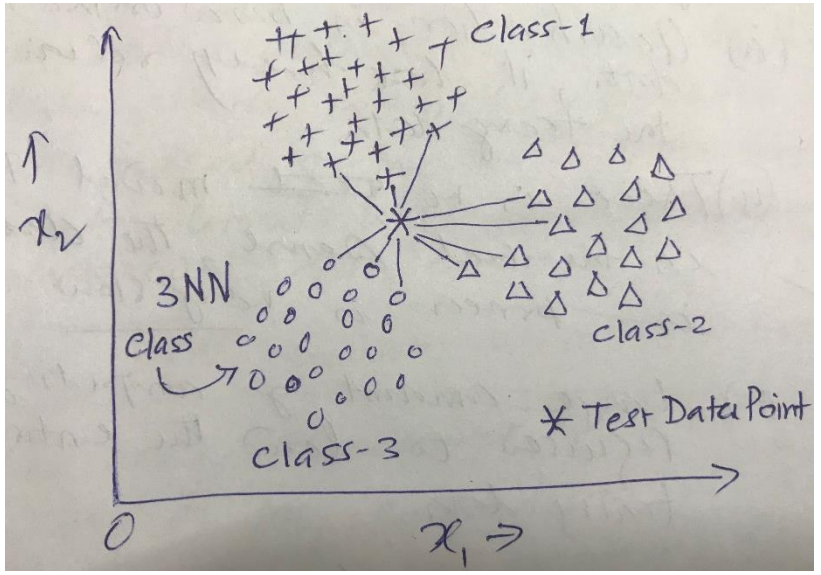
  (ii) The normalization of values of each attribute/feature is essential before using their values in computation of the Euclidean distance.
- The data values are usually converted to Standard Normal range from 0 to 1.

Unit 3: kNN Classification

k-NN Classifier:

- In this method the training data per se (i.e. on face value) are not used for any modelling purpose.
- Actually no model development is carried out.
- Location of Test data/samples points in mD space are compared with training data/samples locations directly based on Euclidean distances computed.
- The Euclidean distances of each of the test data points are computed from all the training points in the feature space.
- Arrange the distances computed in the increasing order along with class labels.
- Then find that class whose first k-points have shortest values.
- Assign the test data point to such a class.
- Extend this to all test data points.





Unit 3: kNN Classification

k-NN Classifier: Value of k (Number of Nearest Neighbours))

- How to find a good value of k in kNN Classifier?
- The kNN method is applied for k-values ranging 1 to some max value k_{max} (in range of 3 to 7).
- Compute the various parameters of the classification results (such as accuracy, specificity, sensitivity etc. for each classification)
- These parameters describe True positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).
- Based on these, generate the Confusion Matrix and then compute from it various quantities viz. Accuracy, Precision, Recall, Sensitivity.
- These are supposed to describe the classifier performance.
- Choose that value of k for which accuracy is high and sensitivity is high compared specificity.
- In most of cases, the k value in the range of 3 to 7 is found to give good classification accuracy.
- This is due to localization of classes generally, especially so in case of imagery data.

Unit 3: kNN Classification

kNN Classifier :

- The kNN method can also be used for numerical prediction.
- In this case described so far, the kNN method returns the average of the numerical values of the kNN class data points.
- In case of non-numeric features, the nominal values are compared to find the closest features.
- The kNN method is a Lazy learner.
- Unlike Eager learners, this method directly uses the training data as they are.
- Thus, in kNN there is no learning happening in the real sense. Hence it is a Lazy learner.
- The Eager learners follow the general steps of ML, i.e. perform an abstraction of the information obtained from the input data and the follow it through by generalization step.

Unit 3: kNN Classification

kNN Classifier :

Strength of kNN method:

(i) It is Extremely simple method, but powerful method.
(ii) It is Very effective in certain situations, e.g. Recommender system design and Similar Documents searching.
(iii) It is very fast. Almost much less time required for the training phase.

Weaknesses of kNN method:

(i) Lazy learner.
(ii) Classification relies heavily on the training data as they are, which invariably contain errors.
(iii) There is no model trained in the real sense.
(iii) Large amount of computing resources are required to load the entire data, like it is done in batch mode.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

3

Unit 3:Classification

Linear Discriminant Analysis (LDA):

- LDA is commonly used feature extracting technique like PCA (Principal Component Analysis) or SVD (Singular Value Decomposition).
- The objective of LDA is to transform a data set into lower dimensional features space (Dimensions reduction Technique).
- In PCA, the interest is in reducing the data dimensionality, may be at the cost of some variability or information content.
- But, the focus of LDA is not to capture the data set variability.
- Instead, LDA focusses on class separability, i.e. separating the features based on class separability.
- The above step avoids overfitting of a ML method.
- LDA computes Eigen values and Eigen functions within a class and interclass scatter matrices.

Unit 3:Classification

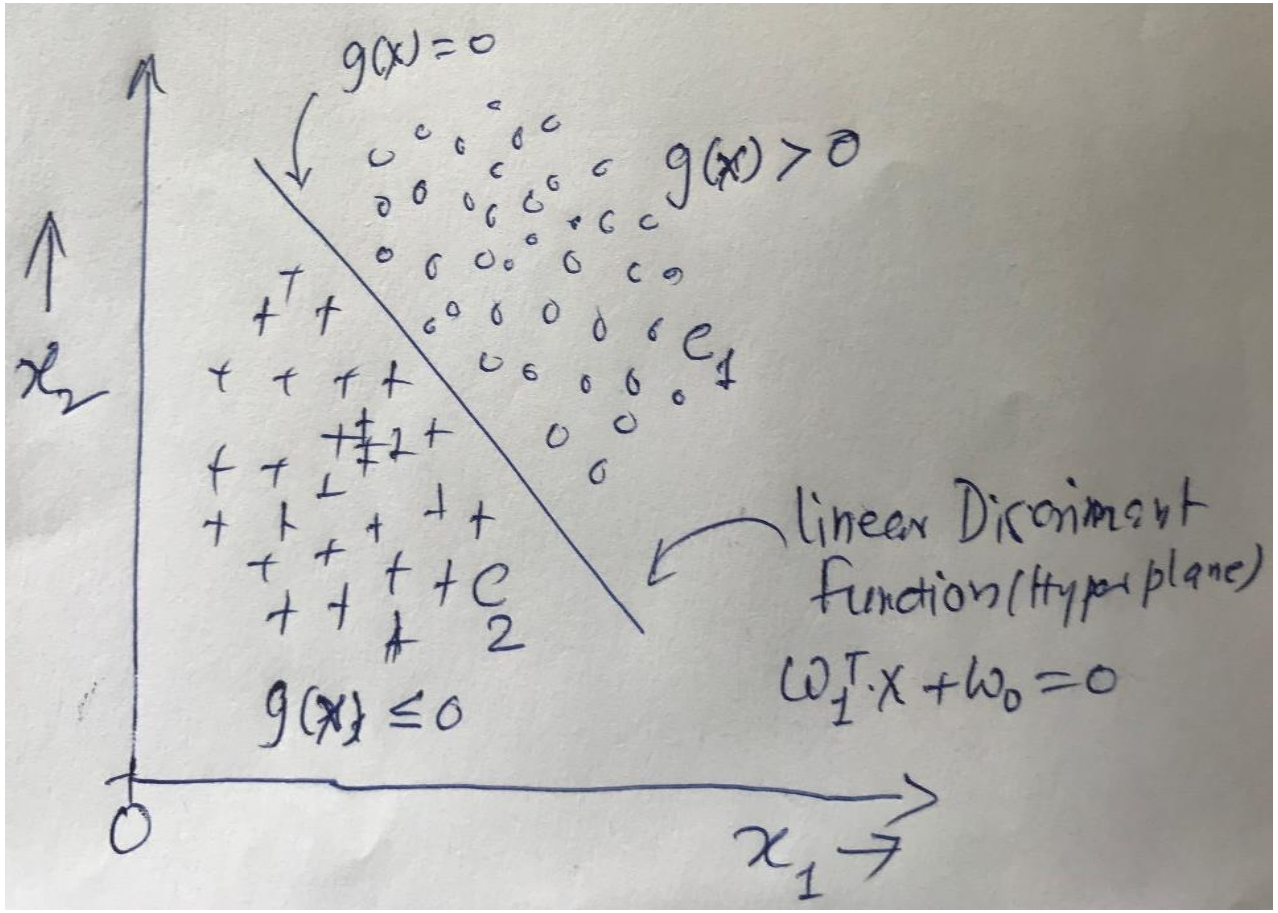
Method of Linear Discriminant Analysis:

- We know that the instances/examples of a class are represented as a point in the m-Dimensional features space.
- In Linear Discriminant Analysis (LDA), we assume that:

  "the group of instances of a class in feature space are linearly separable from other groups of classes in feature space".
- Linearly separable means the line (2D) or plane (3D) or hyperplane (mD) of dividing boundary can be represented by a linear sum of the attributes or features.
- This means the methods of representing the dividing planes can be modelled using simple linear algebra.
- Machine Learning Library of Scikit-learn, known as "linalg", contains Suit of functions for this type analysis.

Unit 3: Method of Linear Discriminant Analysis

Geometry of Linear Discriminant Function:



Unit 3:Classification

Method of Linear Discriminant Analysis:

- The LDA is based on approach that estimates parameters of discriminant boundary directly.
- It does Not consider the distributions of points in clusters.
- In classification, one class needs to be separated or discriminated from other classes based on the clustering properties of the features.
- We define a set of discriminant function $g_i(x)$, i = 1 to k, where k is number of clusters or classes.
- Then choose a class assigned as C_i such that:

  $g_i(x) = \max[g_i(x)]$, that maximum of $g_i(x)$ function for class C_i.

Unit 3:Classification

Method of Linear Discriminant Analysis:

- In the simplest case, the function $g_i(x)$ are assumed to be linear in X, and given by:

  $g_i(x|w_i,w_{i0}) = w_i^T x + w_{i0} = \sum_{j=1}^{m} [w_{ij}\, xj] + w_{io}$

- In the above equation, the RHS is the weighted sum of several factors or terms.
- The magnitudes of the weights show the importance of that factor and sign shows whether the effect is +ve or –ve.
- If it is +ve, that factor is supposed to be a enforcing factor, and if it is –ve, that corresponding factor is said to be inhibiting.

Unit 3:Classification

Method of Linear Discriminant Analysis:

- Steps:
  - (1) Compute mean vectors of the individual classes (S_w).
  - (2) Compute the intra-class (S_w) and also inter-class scatter matrices S_B.
  - (3) Compute Eigen values and Eigen-functions of S_w^{-1} and S_B
  - (4) Identify the top k eigenvectors having the top k eigenvalues.

  $S_w = \sum_{i=1}^{C} S_i$ and

  $S_i = \sum_{x \in Di}^{n} [(x - mi)(x - mi)^T]$, where m_i is the mean vector of the i^{th} class.

  $S_B = \sum_{i=1}^{C} [N_i (m - mi)(m - mi)^T]$

Here, m_i = mean vector of the i^{th} class, m = the overall mean of the data, and N_i = the sample size of the i^{th} class.

Note: Scatter matrix is the scatter plot in higher dimensional space (more than two dimensions).

Unit 3: Method of Linear Discriminant Analysis

Method of Linear Discriminant Analysis:

- Its complexity can be increased to quadratic discriminant function, given by:

$$g_i(x|W_i, w_i, w_{i0}) = X^T W_i x + w_i x + w_{io}$$

- Note the use of both capital and lower case w in the above equation.
- This above formulation has both bias/variances.
- The quadratic model requires larger volume of training data.
- And it may end up overfitting on small data sizes.
- Larger the number of terms in the model, larger the requirement of the training data.
- When a linear model is not adequate, it can be intrinsically complex model and require to be accounted for non-linearity related complexities.
- In such a case, Support Vector Machine algorithm can be used.

Unit 3: Method of Linear Discriminant Analysis

Method of Linear Discriminant Analysis:

- In case of model underfitting, an alternate way would be adding higher order terms (such as product terms).
- If x_1 and x_2 are two attribute variables, new additional variables can be generated, such as x_1x_2, x_1^2, x_2^2, in addition x_1 and x_2.
- Let $z = [x_1, x_2, x_1x_2, x_1^2, x_2^2]^T$ to be a five dimensional feature space, z.
- Then, the generalized discriminant function can be defined as:

  $g_i(x) = \sum_{j=0}^{k} w_j \phi_{ij}(x)$
- Here ϕ_{ij} are called basis functions.
- They can be: $\sin(x_i)$, $\text{Exp}(-(x_i - m)^2/c)$, $\log(x^2)$ etc.

Unit 3: Method of Linear Discriminant Analysis

Geometry of Linear Discriminant Function:

LDF for two classes:

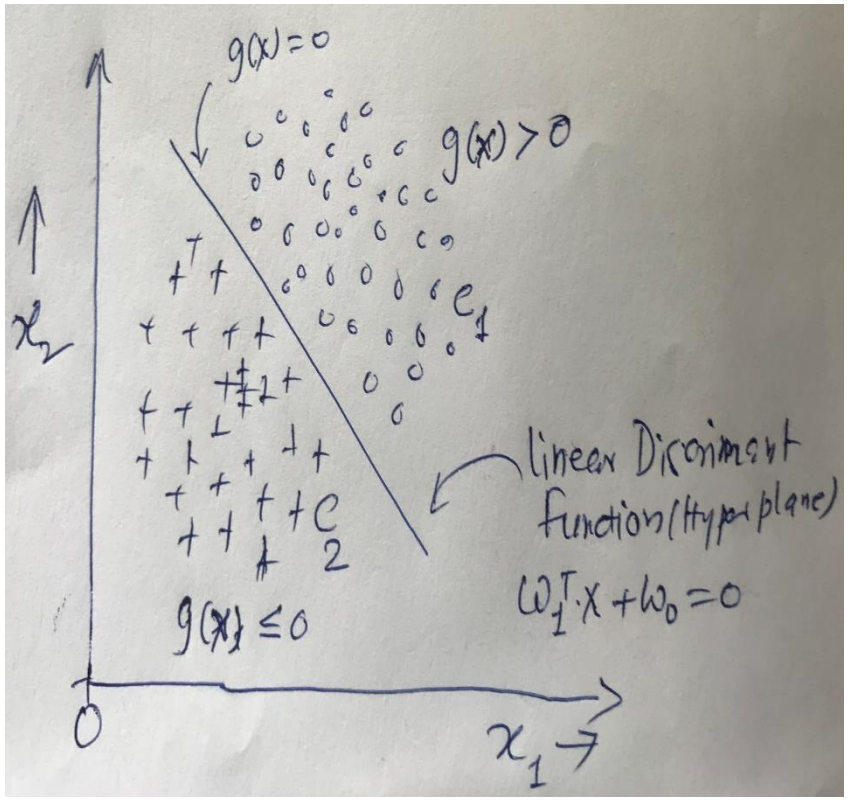
- Two classes case is the most simple case of LDA.
- For this case of classes, the LDF can be written as:

$$g(x) = g_1(x) - g_2(x) = (w_1^T X + w_{10}) - (w_2^T X + w_{20})$$
$$= (w_1^T - w_2^T) + (w_{10} - w_{20})$$
$$= w^T X + w_0$$

- And for this case, the classification scheme will be as:

$$\text{Class} = \{c_1 \text{ if } g(x) > 0 \mid c_2 \text{ if } g(x) \leq 0\}$$

- This defines a hyperplane with weights as w_1 and bias w_0 as threshold or bias.



Unit 3: Method of Linear Discriminant Analysis

Geometry of Linear Discriminant Function:

DF Multiple classes:

- When there are more than the two classes ($k > 2$) , then there are k discriminant functions, given by

  $g_i(x) = w_i^T X + w_{io}$

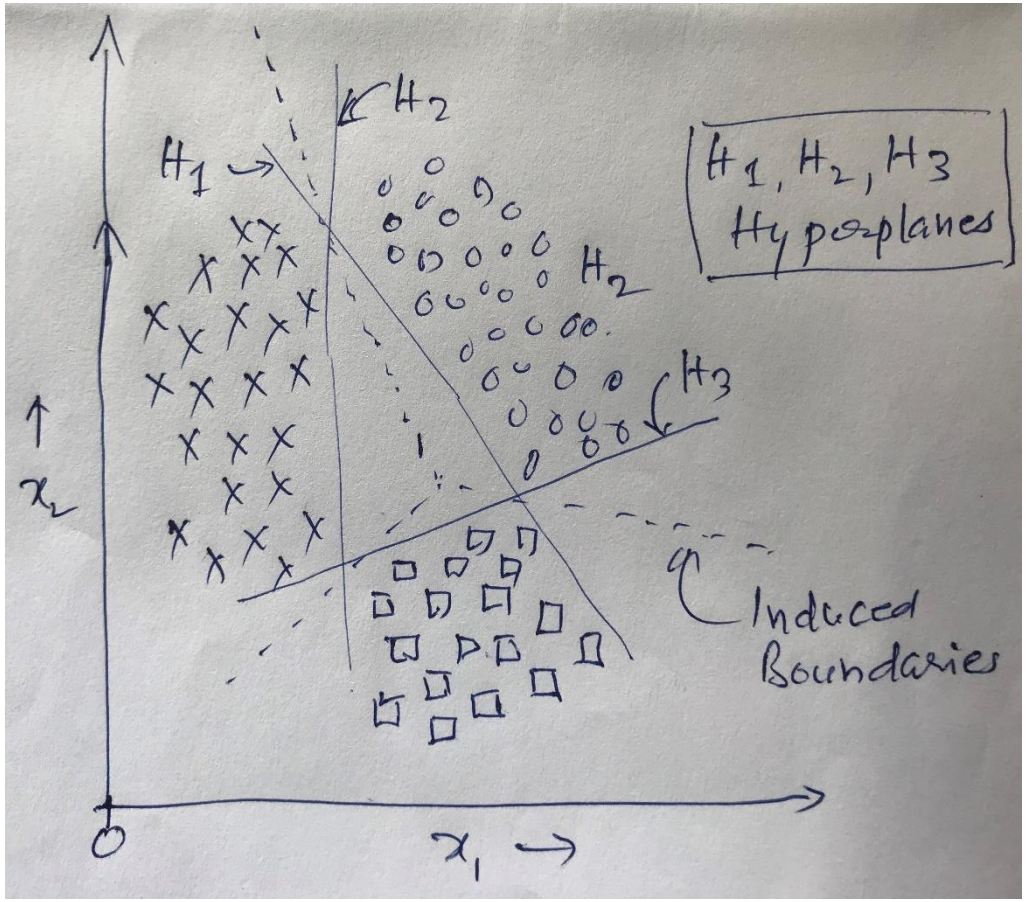
- The parameters w_i and w_{i0} are computed and assigned to a class as:

  $g_i(x | w_i, w_{i0})$

  $= \{ > 0 \text{ if } x \in C_i \mid \leq 0 \text{ otherwise}\}$

  for all X in the training data.

- Each hyperplane separates the examples of C_i from examples of the other classes.



Unit 3:Classification

Method of Linear Discriminant Analysis:

- In many applications or cases, linear discriminant function is quite accurate in the classes separability.
- When classes are Gaussian distributed which share co-variances matrix, the optimal discriminant is linear.
- LDA can be used even when this Gaussian assumption is not strictly valid and the model parameters can be computed w/o making any assumptions of the class densities.
- LDA is always preferred to more complex discriminant functions.
- The problem of finding a LDA function is to search for the parameter values that minimise the error function.

Unit 3: Method of Linear Discriminant Analysis

Geometry of Linear Discriminant: LDF Multiple classes:

- Using above type of discriminant function is equivalent to assuming that all classes are linearly separable.
- Thus, the existence of hyperplane need to be assured.
- But this is not the case always. In such cases, there could be overlaps.
- Such cases are considered as reject class.
- But, the usual approach is to assign such outliers to the class having highest discriminant function value.
- Choose C_i, if $g_i(x) = \max \{g_j(x)\}$, j = 1 to k.
- As $\frac{|gi(x)|}{||Wi||}$ is the distance of the discriminant hyperplane from the input point, the above rule assigns the data point to that class whose hyperplane is most distant.

  This is called a Linear Classifier.
- Geometrically, it divides the feature space sometimes into k decision regions.

Unit 3: Method of Linear Discriminant Analysis

Geometry of Linear Discriminant:

Pairwise Separability:

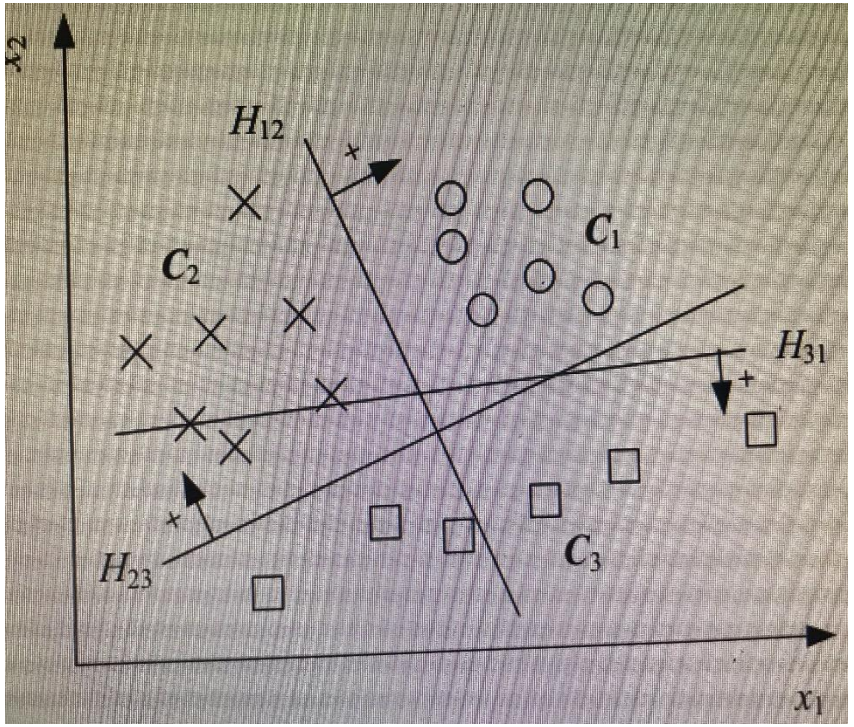
- If classes are not linearly separable, one approach is to divide them into a set of linear problems.
- One possibility is pairwise separation.
- This uses k(k-1)/2 linear discriminant , given by:

  $g_{ij}(x|W_{ij}, W_{ij0}) = W_{ij}^T X + W_{ijo}$
- The parameters W_{ij}, $j \neq i$ are computed during the training so as to have:

  $g_{ij}(x) = \{ > 0$, if $x \in C_i \mid \leq 0, x \in C_j \mid$ Anything (reject class,

  otherwise$\}$

  Where i and j = 1-k, and $i \neq j$.
- If the classes are not linearly separable and pairwise separable, then pairwise separability can be used.
- Thus, this leads to non-linear separation of classes.
- This is another example of breaking down complex non-linear problem into set of linear problems.



APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

4

Unit 3:Classification

Dimensional Reduction Techniques:

- Many ML problems involve thousands or even millions of features for each training instances.
- Such a situation makes training a slow process.
- Also, makes it much harder to find a good solution.
- This problem is often known as “Curse of Dimensions”.
- In practice, it is often possible to reduce the number of features (dimensionality) considerably.
- But, reducing features dimensionally does loose some information (just like compressing an image to jpg).

Unit 3:Classification

Dimensional Reduction Techniques:

- In some cases, the reducing of dimensionality of the training data may filter out some noise and unnecessary details.
- This can result in higher performance (or just speed up training process).
- The Dimensionality reduction is also extremely useful for data visualization.
- Reducing dimensionality to 2 or 3 makes it possible to plot a high-D training data on a graph.
- This can gain some important insights by visual inspection, such as detecting patterns, clusters etc.
- There are two main approaches to Dimensionality reduction,

  (i) Projection and (ii) Manifold Learning.
- Popular Dimensionality Reduction Techniques are:

  (i) PCA, (ii) Kernal-PCA and (iii) LLE (Locally Linear Embedding).

Unit 3: Support Vector Machine Classifier

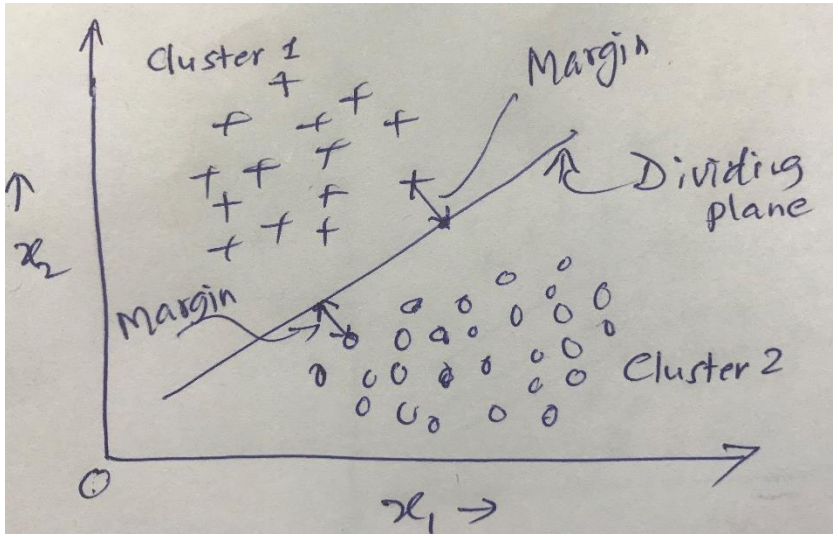
Support Vector Machine (SVM) Classifier:

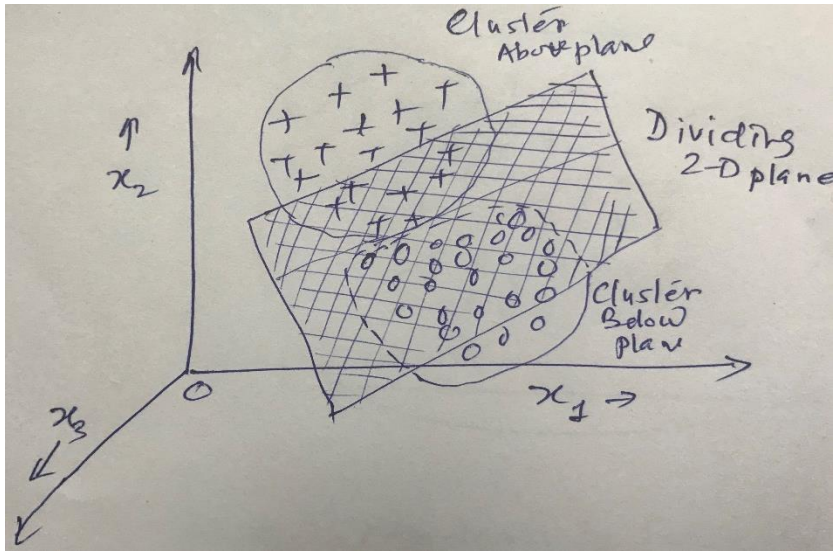
- It is an algorithm for both linear Classification and Regression.
- It is based on concept of surface in 2-D feature space (and hyperplane in multidimensional space).
- It draws a boundary in feature space of multidimensional feature space separating classes.
- The output prediction is one of two conceivable classes existing in the training data, i.e. it is basically a binary classifier.
- For m dimensional (mD) feature space, the SVM algorithm builds a (m-1)D hyperplane or hypersurface that discriminates the examples of two output classes.

Unit 3: Support Vector Machine Classifier

SVM Classifier:

- If the data examples are already linearly separable in feature space, then instances of each class fall on two sides of a straight line drawn in the 2-D feature space.
- When this concept is extended to higher-D space, the dividing boundary transforms to plane surface and hyperplane surface.
- The dividing plane will be (m-1)-D hyperplane in the m-D feature space.
- The goal of SVM algorithm to find this hyperplane in the m-D feature space.





Unit 3: Support Vector Machine Classifier

SVM Classifier:

- Finding (m-1)-D hyperplane is done for the training data and then it is used to classify the test data.
- Theoretically, there will be many hyperplanes possible that separate the two classes.
- The real challenge of the SVM algorithm is to find the optimal hyperplane giving minimum error in classification.
- The space separating the closest data points of two classes from separating hyperplane is known as **Soft Margin**.

Unit 3: Support Vector Machine Classifier

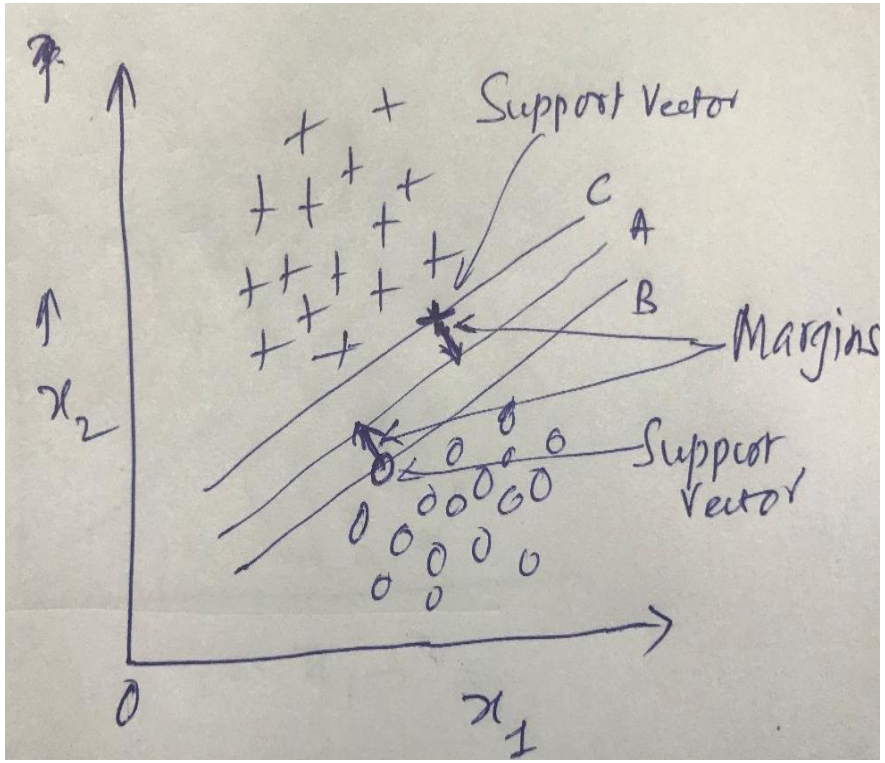
SVM Classifier:

- A **hard margin** in terms of SVM means that the SVM model is inflexible in classification and tries to fit the data exceptionally well resulting in Overfitting.
- Hard margins result in case of overlapping data points in mD feature space.
- By introducing curves or curvilinear features in hyperplanes, the model tries to fit data better and better ways.
- This becomes compute intensive and takes lot of computing time, and also there is risk of overfitting.

Unit 3: Support Vector Machine Classifier

Support Vectors in SVM:

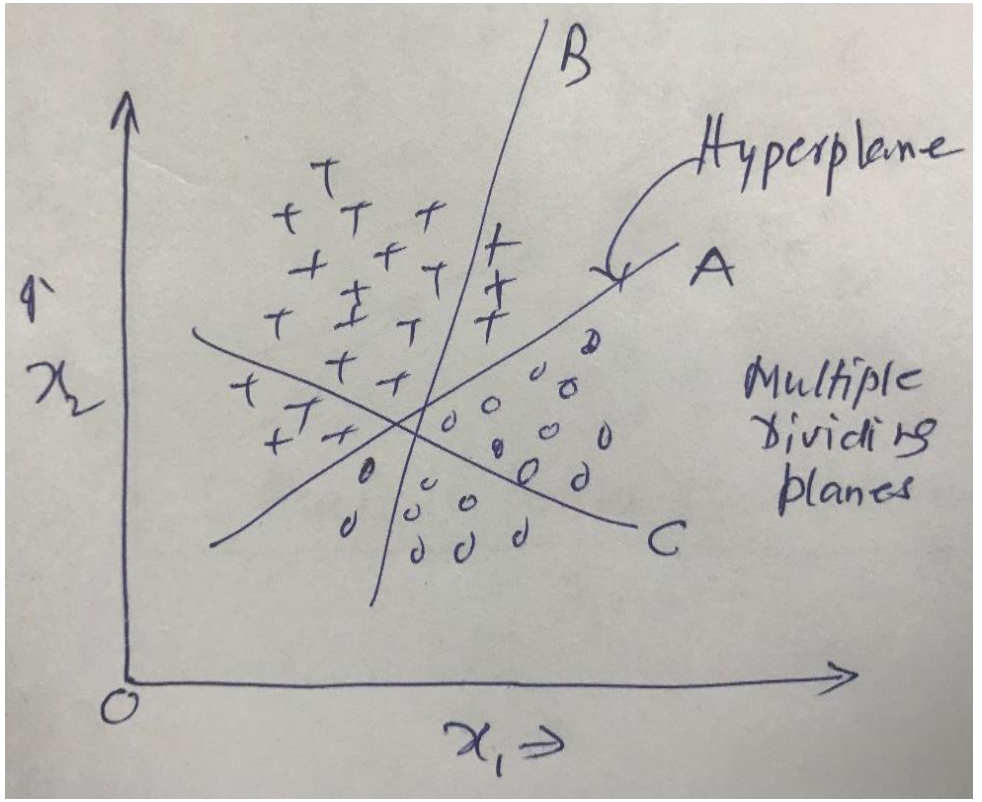
- Each example is a vector, and it is represented as a point in m-D feature space.
- Those examples which lie near (closer to) the identified set of dividing lines or hyperplanes are known as **Support Vectors.**
- If all or any of the support vector data points are removed, the dividing line or hyperplane will be altered.
- The location of hyperplanes are determined mainly by these Support Vectors.
- The separation between the support vectors of the two classes determines the **Soft Margin**.



Unit 3: Support Vector Machine Classifier

Determining the Correct Hyperplane:

- As stated above there could be multiple hyperplane dividing the training data into two classes.
- We need to identify or select the one which has least soft margins (distances from Support Vectors of each class).
- In the figure, the hyperplanes B and C have large errors, whereas A has least error.
- Therefore, A is the Optimum Hyperplane.



Unit 3: Support Vector Machine Classifier

Margins:

- In 2-D feature space, the hyperplane dividing the two classes is a line and it is given by:

$$c_0 + c_{1*}x_1 + c_{2*}x_2 = 0$$

- And in m-D feature space, the hyperplane is given by:

$$c_0 + c_{1*}x_1 + c_{2*}x_2 + \ldots\ldots + c_{m*}x_m = 0 \quad \text{or}$$

$$\vec{c}.\vec{x} + c_0 = 0.$$

- The more distant the data points lie from the hyperplane, the more confident or more separable the classes are. These have large soft margins.
- The shortest distance between the hyperplane and the support vectors is known as Margin.Each class has its Margin.

Unit 3: Support Vector Machine Classifier

Determining the Correct Hyperplane:

- In some cases, some outliers of one class fall in the region of another class.
- In such a case, a hyperplane which strictly separates it from the class, may have less margin.
- The hyperplane having less margin may properly classify the outlier, but it is not a optimum hyperplane.
- This type of classification mostly does not generalize well (i.e. fails in predictions). Such a model overfits the training data.
- The margin of this type of hyperplane is known as **HARD MARGIN**.
- In case of large extent of overlaps due to outliers, the hard margin gets smaller and the hyperplane becomes more complex.
- This results into the SVM overfitting training data and poor performance in prediction.
- We know that too much of overfitting must be avoided always due to its poor predicting capabilities.

Unit 3: Support Vector Machine Classifier

In Nutshell:

- The hyperplane should separate the data points of any two classes in best possible way, but with optimal soft margins.
- It should maximize the margins.
- If there is need to choose or prioritise between lower margin and higher classification accuracy, the higher classification accuracy must get priority, i.e. higher accuracy must be preferred.
- That is, the chosen hyperplane should reduce the misclassifications.

APPLIED MACHINE LEARNING

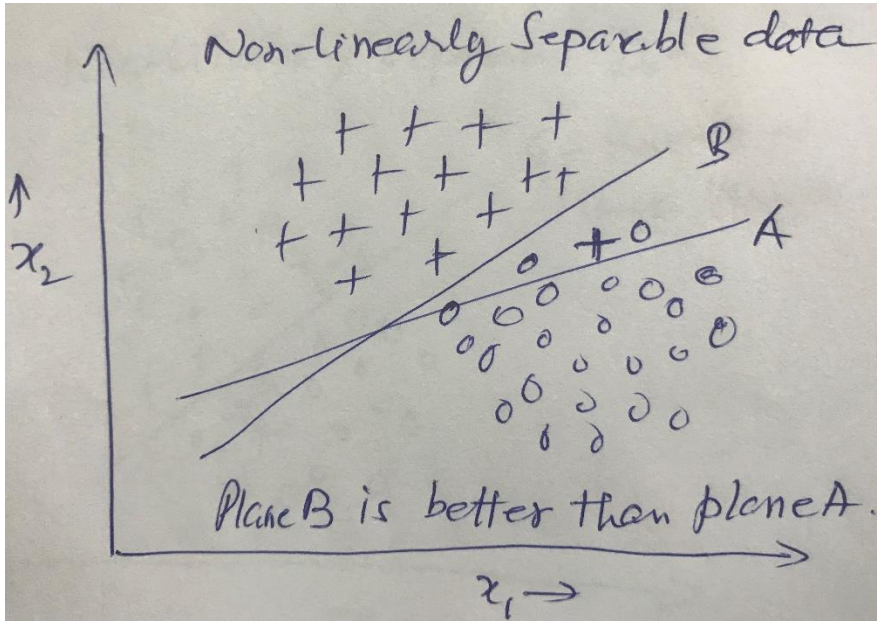
UNIT-III: CLASSIFICATION

5

Unit 3: Support Vector Machine Classifier

Maximum Margin Hyperplane (MMH):

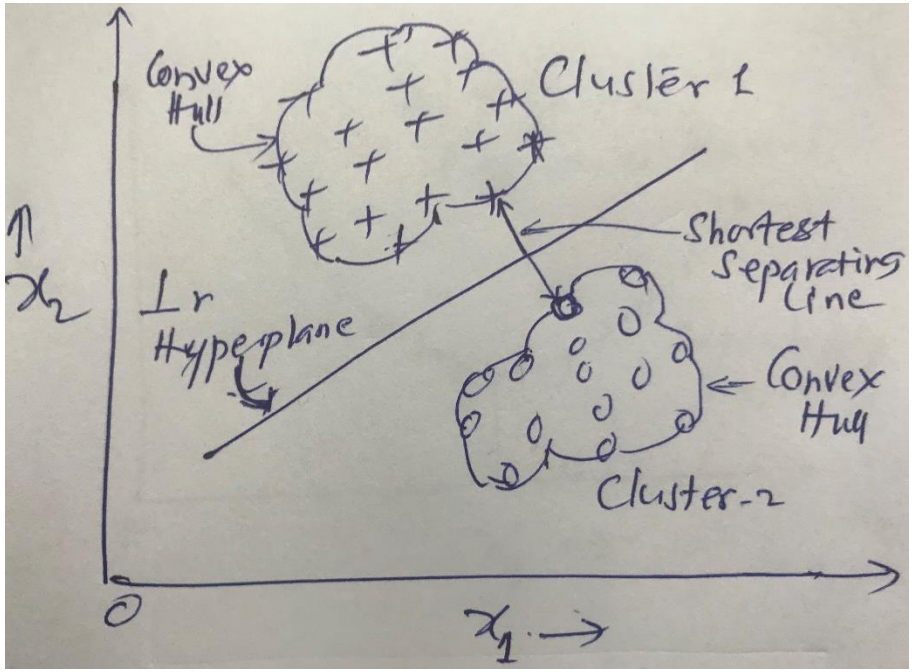
- The MMH is that hyperplane which gives largest margins (preferably soft) separating the two classes.
- The search for the such a MMH helps in best classification and higher level of generalisation.
- There should be at least one support vector from each class.
- Modelling a problem using SVM is nothing but identifying Support Vectors and using MMH corresponding to the problem space.



Unit 3: Support Vector Machine Classifier

Finding MMH for linearly Separable Classes:

- An outer boundary needs to be drawn for data points of each class.
- In 2-D space, the outer boundary is known as Convex Hull.
- The MMH is then the Perpendicular hyperplane bisecting the shortest line joining the support vectors.
- In multi-dimensional space, the same method is followed.
- Then a hyperplane passing through closest support vectors and perpendicular to line joining support vectors is determined.



Unit 3: Support Vector Machine Classifier

Finding MMH for linearly Separable Classes:

- The hyperplane surface separating the classes will be a plane perpendicular to it and dividing equally.
- The equation of hyperplane is given by:

  $C.X + C_0 = 0$

  For one class, it is $C.X + C_0 \leq 1$ or 0 and

  For other class, it is $C.X + C_0 \geq 1$ or 0
- The distance between the two planes is $2/|C|$.
- For maximizing this distance, the magnitude $|C|$ has to be minimised.
- The task of the SVM algorithm is to solve the optimization problem of setting or minimizing of $|C|/2$, or maximising $2/|C|$.

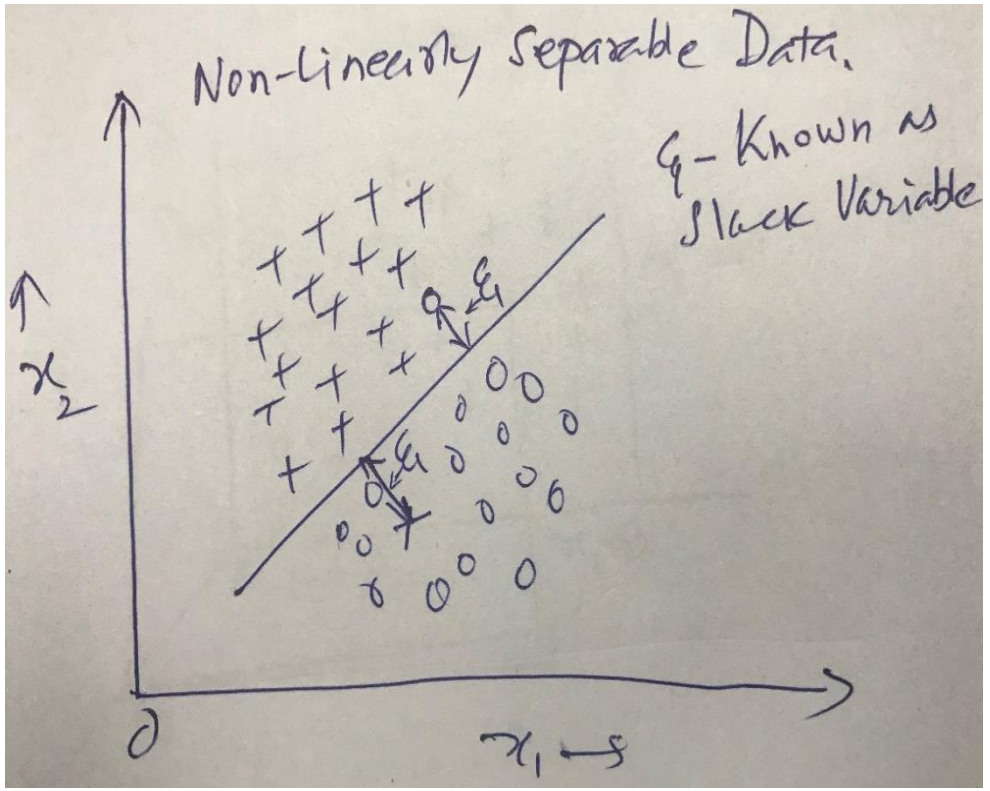
Unit 3: Support Vector Machine Classifier

Finding MMH for linearly Non-Separable Classes:

- In case of non-linearly separable data, there are overlapping of data points from one class on the other class.
- For identification of MMH, a slack variable ξ is added to provide some soft margin for data instances falling on the other classes.
- To find the MMH in this case, the quantity to be optimised (i.e. minimized) is then:

  $\min(C^2/2) + q * \sum_{i=1}^{m}(\xi_i)$

- Here, q, the hyperparameter, is the cost value imposed on all such outlier instances.



Unit 3: Support Vector Machine Classifier

Kernel Functions:

- One way to deal with non-linearly separable data is using a slack variable and a optimization function to minimise the cost function (as seen before).
- SVM has a technique called kernel trick to deal with non-linearly separable data.
- These are the functions which transform non-linearly separable data from lower dimensional to higher dimensional input space.
- Such functions are called Kernel functions.

Unit 3: Support Vector Machine Classifier

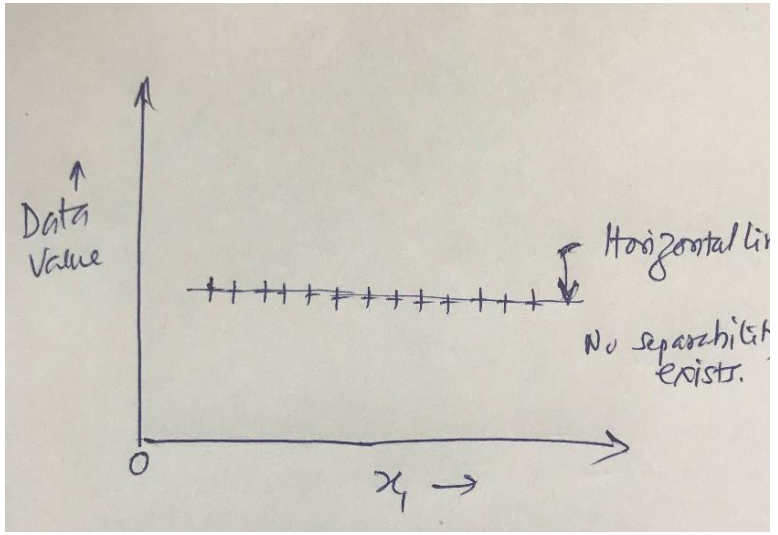
Kernel Tricks in SVM:

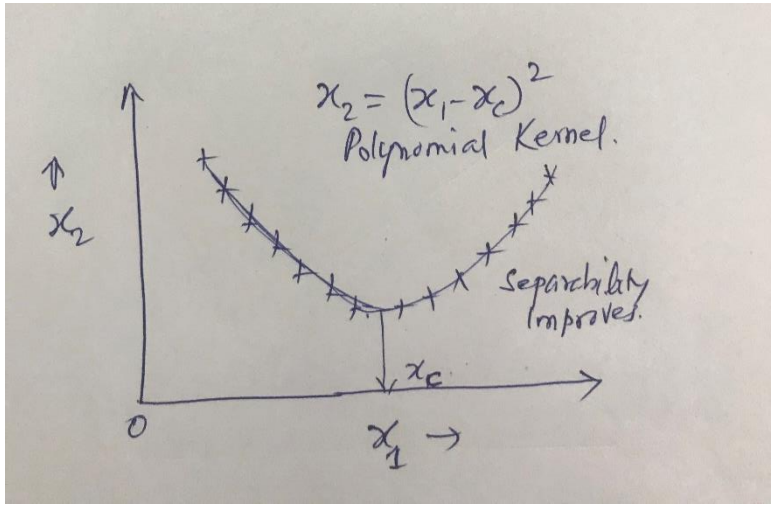
- Although linear SVM classifiers are efficient and work well in many cases.
- Some cases or practically many data sets are not even close to being linearly separable.
- One approach to make them separable is to add more features, i.e. to increase the dimensionality of the data.
- The features in the data are non-linear data sets.
- It is expected that the non-linear features may become separable in the higher dimensional feature space.
- Note this method is opposite to feature reduction techniques, such as PCA, SVD etc.

Unit 3: Support Vector Machine Classifier

Kernel Tricks:

- By adding higher dimensions, such as, a polynomial feature, for example, can result in linearly separable data.
- Example: A data set with a single feature with constant value is not separable.
- Now, after adding another feature, in the 2D feature space, they become separable.
- The 2nd feature could be $x_2 = (x_1-x_c)^2$.

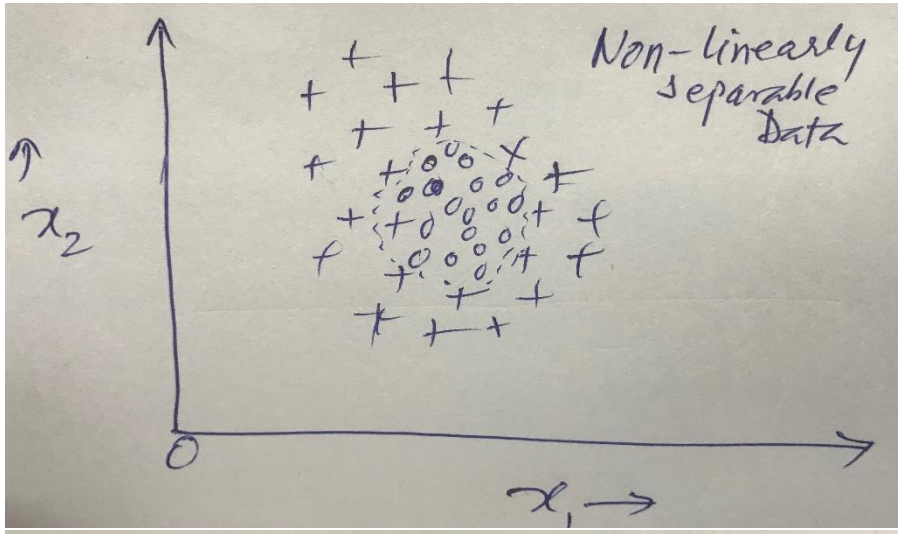


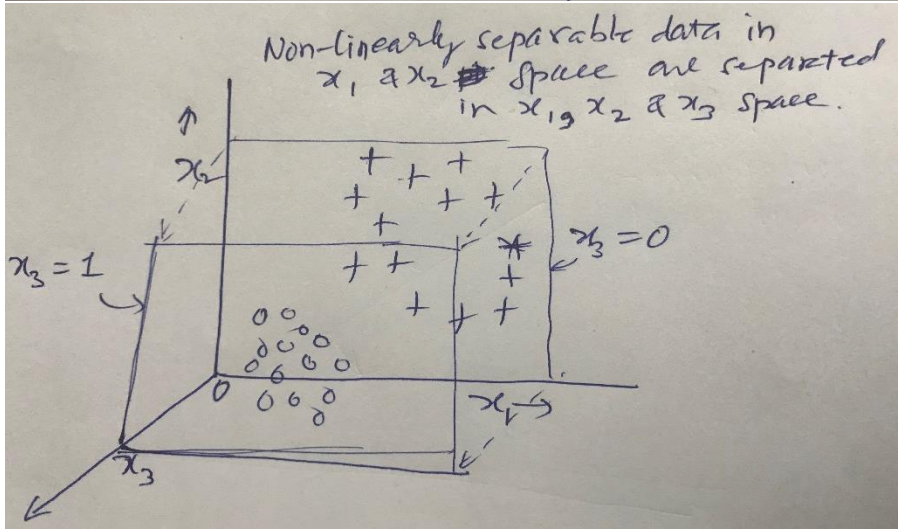


Unit 3: Support Vector Machine Classifier

Polynomial Kernel:

- Though adding a polynomial kernel is simple to implement, a low degree polynomial kernel may not deal with highly complex data sets.
- Adding high polynomial degree kernels may result into huge number of features.
- In example shown, x3 with value of 1 is added.
- This kernel trick makes it possible to get the same result in classification as that of addition of many polynomial features w/o actually having to add them.





Unit 3: Support Vector Machine Classifier

Kernel Tricks:

- Some common Kernel functions are:

(i) Linear Kernel: To transform $K(x_i, x_j)$ ➔ $x_i.x_j$

(ii) Polynomial Kernel: To transform $K(x_i, x_j)$ ➔ $(x_i.x_j + 1)^d$

(iii) Sigmoid Kernel: To transform $K(x_i, x_j)$ ➔ $\tanh(kx_i.x_j)$

(iv) Gaussian RBF Kernel: To transform $K(x_i,x_j)$ ➔ $\exp[-(x_i^2+x_j^2)/2\sigma^2]$

Where d = degree of the polynomial, and σ = Standard deviation of $(x_i^2+x_j^2)$

- The success of SVM depends on the:

(i) Selection of appropriate kernel function, and

(ii) Also selection of the appropriate kernel parameters.

Unit 3: Support Vector Machine Classifier

Strength and Weakness of the SVM Algorithm:

Three main strengths of the SVM algorithm are:

(i) It can be used for both regression and classification.

(ii) It is a robust classifier, i.e. it is not impacted by noise or outliers.

(iii) The predictions are more reliable and promising, and rarely overfitting.

Unit 3: Support Vector Machine Classifier

Strength and Weakness of the SVM Algorithm:

- The weakness of the SVM algorithm are:

(i) It is applicable more effectively only to binary classification (i.e. when there are only two main classes).

(ii) The model is quite complex, difficult and impossible to understand in case of high dimensional feature space.

(iii) It is slow in case of larger data sets and also in case of large number of features.

(iv) It is quite memory intensive algorithm.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

6

Unit 3: Binomial Distribution

- Before discussing Binomial Theorem, first we need to understand what are the Bernoulli Trials.
- Bernoulli Trials:
  - If an experiment consists of a series of trials, the outcome of each trail can be considered as being either as SUCCESS (s) or FAILURE (f), True(T) or False(F), or Yes(Y) or No(N) etc.
  - A trial having this property is called “Bernoulli Trial”.
  - Each of the outcome of these trials are independent.
  - The theoretical basis of Binomial distribution is the Binomial Theorem.
  - Outcome of tossing of a coin (Bernoulli trial) is Bernoulli event, as only two outcomes are possible, head or tail.
  - But outcome of a dice throw is NOT, in principle, a Bernoulli trial.
  - Can be considered so if outcome of one event (e.g. face 5 up) is only considered.

Unit 3: Binomial Distribution

- **Binomial Theorem:**
  - It deals with expansion of power functions of sum of TWO variables.
  - For any two real numbers a and b, and for any positive integer n, the expansion of $(a+b)^n$ is given by Binomial Theorem:

    $(a+b)^n = \int_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k}$
  - Here $\binom{n}{k}$ = nCk. This is number of combinations of k taken at a time from n.
  - If n = 2, then $(a+b)^2$ = aa + ab + ba + bb = aa + ab +ab + bb

    $= a^2 + 2ab + b^2$ (Coeffs: 1,2,1)
  - Similarly, if n = 3, then $(a+b)^3$ = aaa + aab + aba + baa + abb + bab + bba + bbb = $a^3 + 3a^2b + 3ab^2 + b^3$ (Coeffs: 1,3,3,1)
  - The numerical coefficients (1,2,1), (1,3,3,1) and (1,4,6,4,1) in each above cases are the combinations given by nCk

Unit 3: Binomial Distribution

Binomial Theorem:

- If X denotes the number of successes obtained in n Bernoulli trials, then X is a random variable with Binomial Distribution.
- Let X be a random variable with Binomial Distribution with parameters n and p.

  It (i.e. the Binomial Distribution) is defined as: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$
- Here p is like a and (1 - p) is like b in the Binomial Theorem
- In the above expression, n is number of Bernoulli trials and x = 0,1,2,3, ….,n
- From this Binomial Theorem, we see that the total probability (over all x) will be given by:

  $\Sigma \binom{n}{x} p^x (1-p)^{n-x} = [p + (1 – p)]^n = 1^n = 1$
- Mean = np and variance = np(1-p).

Unit 3: Multinomial Distribution

Multinomial Theorem:

Multinomial distribution is related to type of events that are characterized by the following property:

- Each observation/outcome of an event can be classed as falling into exactly one of the categories or bins.
- These are supposed to be mutually exclusive events.
- Interest is in the number of observation of events falling into each category or bins.
- There are three problems areas:

(i) Testing to see whether a set of observations is having a specified probability distribution.

(ii) Testing for independence between any two variables used to classification or categorization.

(iii) Finding the proportions of each events falling in different categories.

Unit 3: Multinomial Distribution

Multinomial Theorem:

- Before giving the definition of multinomial random variable, it is required to define “multinomial trial”.
- Definition of Multinomial trial: If a trial with multiple probabilities p_1, p_2, p_3, ……., p_k results in exactly one of k possible outcomes, then the trail is known as Multinomial Trial.
- Since the trials will result into exactly any one of the k possible outcome, the sum of all probabilities is equal to 1.
- $\sum p_i = 1$, where i goes from 1 to k.

Unit 3: Multinomial Distribution

Multinomial Theorem:

- The multinomial random variable arises whenever a series of independent and identical multinomial trials are performed.
- Definition of Multinomial Random Variable:
  - Let n denote independent and identical multinomial trials with probabilities p_i, i = 1 to k, and
  - Let X_i, i = 1 to k denoting number of trials that result in outcomes i = 1,2,3,……,k,
  - Then, the k-tuple (X_1, X_2, X_3, ….X_k) is called multiple random variable with parameters (n, p_1, p_2, p_3, ….,p_k).

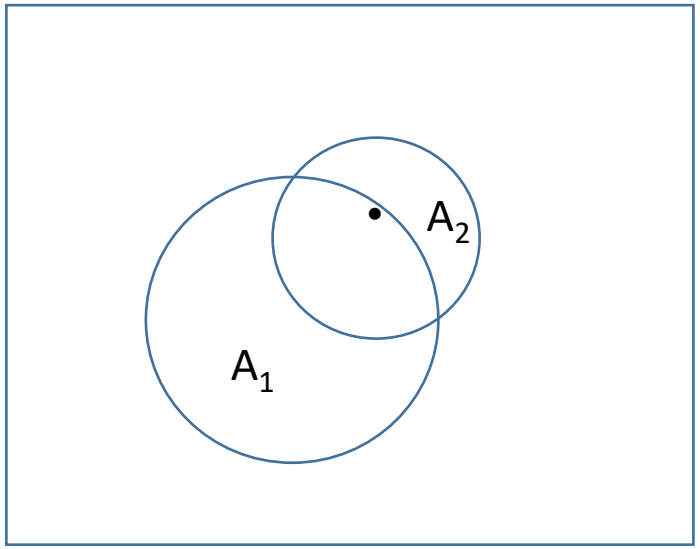
Unit 3: Bayes Theorem

Conditional Probability:

- It is defined as the probability of occurrence of an evet A_2 with the condition that another event A_1 has already occurred.
- Conditional Probability is Denoted by $P[A_2|A_1]$. Probability of event A2 subject to condition that Event A1 has occurred.
- It is given by: $P[A_2|A_1] = \frac{P[A1 \cap A2]}{P[A1]}$

  i.e. $P[A_2|A_1]*P[A_1] = P[A_1 \cap A_2]$ = intersection part in Venn Diagram

Venn Diagram



Unit 3: Bayes Theorem

Conditional Probability:

Example:

- If Event A_1 is hitting of a bullet and Event A_2 is dying of bullet injuries, then $P[A_2|A_1]$ is probability that bullet hits a person and the person dies.
- In absence of their dependence, i.e. they are independent:

  $P[A_2|A_1] = P[A_2]$ i.e. it does not depend on whether A_1 has occurred or not.
- There is No intersection area in Venn diagram of events A_1 and A_2.

Unit 3: Bayes Theorem

Independence of events

- Two events are considered as independent of each other if the knowledge of the fact that one has occurred gives us no clues as to the likelihood that the other event will occur.
- Such events are termed as Independent Events.
- If X and Y are discrete events and X taking value of x and Y taking value of y and the events are A_1 and A_2, then they are independent if and only if

  $P[A_1 | A2] = P[A_1]*P[A_2]$,

  i.e. $P[X = x \text{ and } Y = y] = P[X=x]*P[Y=y]$

  $f_{xy}(x,y) = f_x(x)*f_y(y)$

Unit 3: Bayes Theorem

Independence of events

- This means that joint density is expressed as products of individual densities.
- Definition: If X and Y are random variables with joint density f_{xy} and individual densities f_x and $f_{y,}$ then X and Y are independent if and only if $f_{xy}(x,y) = f_x(x)*f_y(y)$
- Example: In Tossing of two coins, the outcome of different combinations like both heads, both tails, one head and one tail. Second example could be throwing of dices.
- In case of Multiple independent events: $P[A_1|A_2|A_3|.....] = P[A_1]*P[A_2]*P[A_3]*.....$

Unit 3: Bayes Theorem

Bayes Theorem:

- By Reverend Thomas Bayes in 1761 (~260 years back).
- It is on how the Conditional Probability is related to the individual probabilities.
- As per the Bayes Theorem:
  The conditional probability P[A|B] is related to P[B] by
  $P[A|B] = \frac{P[B \cap A]}{P[B]}$ where P[B] > 0
- $P[B \cap A]$ is the Intersection area in the Venn Diagram
- The Bayes Theorem can also be written as:
  $P[B|A] = \frac{P[B \cap A]}{P[A]}$ where P[A] > 0
- From above two, the following equality is deduced:
  $P[A|B]*P[B] = P[B|A]*P[A] = P[B \cap A]$ (as $P[B \cap A] = P[A \cap B]$)
- The $P[A \cap B]$ and $P[B \cap A]$ refer to the same overlapping portion in Venn Diagram.

Unit 3: Bayes Theorem

Generalization of Bayes Theorem:

The Generalized Bayes Theorem:

- If $E_1, E_2, \ldots\ldots, E_k$ are mutually exclusive (i.e. independent from each other) and exhaustive events (all except event B in the Venn diagram) and B is any event, then

$$P[E_1|B] = \frac{P[B|E_1]*P[E_1]}{\{P[B|E_1]*P[B|E_2]* \ldots\ldots\ldots\ldots\ldots*P[B|E_k])\}}$$

- Note: P[A] denotes the probability of the occurrence of event A. The probability of non-occurrence of event A is denoted by P[A']. Sum of the two must be equal to 1.

  That is, P[A]+P[A'] =1 and P[A'] = 1 – P[A].

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

7

Unit 3: Bayes Classification

Bayes and Naïve Bayes Classification:

- The Bayes probability rule is given by:

$$P[A|B]=\frac{P[B|A]*P[A]}{P[B]}$$

- In above eqn. A and B are conditionally related events and P[A|B] denotes the probability of event A occurring when event B has already occurred with probability P[B].
- The task is to find best classification using the probability dependencies knowledge in the training data.
- The a priori knowledge about the probabilities of various classifications is called Prior in the context of Bayes theorem.
- The probability of a particular classification holds for a data set based on the Prior is called Posteriori probability.

Unit 3: Bayes Theorem

Independence of events

- This independence means that joint probability density can be expressed as products of individual probability densities.
- Definition: If X and Y are random variables with joint density f_{xy} and individual densities f_x and f_y, then X and Y are independent if and only if $f_{xy}(x,y) = f_x(x)*f_y(y)$
- Example: In Tossing of two coins, the outcome of different combinations like both heads, both tails, one head and one tail. Second example could be throwing of dices.
- In case of Multiple independent events: $P[A_1|A_2|A_3|.....] = P[A_1]*P[A_2]*P[A_3]*.....$
- This means joint probability is product of probability of individual events.

Unit 3: Bayes Theorem

Bayes Theorem:

- By Reverend Thomas Bayes in 1761 (~260 years back).
- It is on how the Conditional Probability is related to the individual probabilities.
- As per the Bayes Theorem:
  The conditional probability P[A|B] is related to P[B] by
  $P[A|B] = \frac{P[B \cap A]}{P[B]}$ where P[B] > 0
- $P[B \cap A]$ is the Intersection area in the Venn Diagram
- The Bayes Theorem can also be written as:
  $P[B|A] = \frac{P[B \cap A]}{P[A]}$ where P[A] > 0
- From above two, the following equality is deduced:
  $P[A|B]*P[B] = P[B|A]*P[A] = P[B \cap A]$ (as $P[B \cap A] = P[A \cap B]$)
- The P[A∩B] and $P[B \cap A]$ refer to the same overlapping portion in Venn Diagram.

Unit 3: Bayes Theorem

Generalization of Bayes Theorem:

The Generalized Bayes Theorem:

- If $E_1, E_2, \ldots\ldots, E_k$ are mutually exclusive (i.e. independent from each other) and exhaustive events (all except event B in the Venn diagram) and B is any event, then

$$P[E_1|B] = \frac{P[B|E_1]*P[E_1]}{\{P[B|E_1]+P[B|E_2]+ \ldots\ldots\ldots\ldots\ldots+P[B|E_k])\}}$$

- Note: P[A] denotes the probability of the occurrence of event A. The probability of non-occurrence of event A is denoted by P[A'].
- Sum of the two must be equal to 1.

  That is, P[A]+P[A'] =1 and P[A'] = 1 – P[A].

Unit 3: Naïve Bayes Classification

Bayes and Naïve Bayes Classification:

- These Bayesian classifiers are based on probabilities of the occurrences of the classes and those of the features or attribute variables.
- The attribute data probabilities are computed from their normalized frequency distributions.
- These training data probabilities can be used to compute an observed probability of each class based on the feature or attribute values.
- This application of observed probabilities of feature of training data is like applying a priori knowledge of probabilities to an outcome.
- If so, the Bayesian classification is known as Naïve Bayes Classifier.

Unit 3: Naïve Bayes Classification

Bayes and Naïve Bayes Classification:

- In Bayesian classifier, the maximum probability of classification is considered given the observed training data.
- This maximum probability classification is called “Maximum A Posteriori (MAP)” probability based classification.
- The Maximum A POSTERIORI (MAP) is given by:

$$MAP = \max\{P[A|B]\} = \max\left\{\frac{P[B|A]*P[A]}{P[B]}\right\}$$

- If P[B] is constant and equal to 1, and mostly so, the MAP becomes:

$$MAP = \max\{P[A|B]\} = \max\{P[B|A]*P[A]\}$$

- If every possible class has equal a prior probability,, then

  MAP is proportional to max{P[B|A]}, when i.e. P[A] is constant

- This is case of Maximum Likelihood Classification (MXL).

Unit 3: Naïve Bayes Classification

Steps in the Naïve Bayes Classification:

- Construct a frequency table of X. It is constructed for each attribute/feature against target outcome (Y).
- Identify the cumulative probability for two possible class outcomes on the basis of all the attributes:

  (i) Simply multiply probabilities of all favorable attributes to derive positive class, and

  (ii) Multiply probabilities of unfavorable features, to derive negative classes.

- Compute probability through normalization by applying:

$$P(+ve) => \frac{P(+ve)}{P(+ve)+P(-ve)}, \text{ and } P(-ve) => \frac{P(-ve)}{P(+ve)+P(-ve)}$$

- Here P(+ve) is the overall probability of favorable features, and P(-ve) is that of unfavorable features.

Unit 3: Naïve Bayes Classification

Bayes and Naïve Bayes Classification:

- One of the strengths of the Bayesian classifiers is that they use all the available features for prediction, however weak the effects may be.
- It could be termed as Lazy Learner.
- Because it assumes that even if a few parameters have small influence individually on the outcome of classification, the combined effect could be large.
- But is always advisable to do feature engineering using dimensionality reduction techniques. That be Eager Learner.
- Thus, it has higher probability of meeting actual real life outcome.

Unit 3: Naïve Bayes Classification

Properties of Bayesian Learning methods:

- As the method is simple and easy, it is possible to classify new test instances by combining the predictions of multiple classification schemes by assuming suitable weights.
- The out of classification heavily depends on the probability distribution of the feature variables.
- If not known, some background knowledge or previous data or assumptions about the data are required to be known.
- It assumes that even if a few parameters have small effects individually on the outcome, the combined effect could be large.

Unit 3: Naïve Bayes Classification

Properties of Bayesian Learning methods:

- Prior knowledge of features is combined with observed data for computing the final probability of classes.
- The Bayesian approach to learning is more flexible than other approaches, as each observed pattern can influence the probability of the outcome.
- This method can perform better than other methods when validating the predictions.
- The Naïve Bayes Classifier provides a simple and powerful way to consider the influence of multiple attributes on the classification outcome.
- It is able to do the computations through independency assumptions.
- Therefore, it refines the uncertainty of outcomes on the basis of the prior knowledge.

Unit 3: Naïve Bayes Classification

Bayes and Naïve Bayes Classification:

Some Examples:

- Text based data classification, such as SPAM or junk mail filtering, author identification or categorization of topics.
- Medical diagnosis, in which a set of symptoms are related to probability of a disease.
- Network security, such as a detecting an illegal intrusion or anomaly in computer networks.
- Hybrid Recommender Systems: They use the Bayesian classifier and collaborative filtering techniques. This system applies ML for filtering unknown information for prediction.
- Online Sentiment analysis: It uses single words and word pairs to determine sentiments.
- Certain words and word combinations and their frequency occurring indicates either positive or negative, happy or sad, emotional or depressed states.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

8

Unit 3: Classification Accuracy

Validation of Simple Linear Regression Models:

- The Linear Regression model can be validated by F-Test, which is the outcome of analysis of variance(ANOVA).
- The F-Test is given by: F-Fest = $$\frac{(SSE_r - SSE_f)/(k-m)}{SSE_f/(n-k-1)}$$
- Here k number of attributes used in full model, m is actual number of attributes in the reduced model, n is number of examples.
- Also, r stands for residual and f stands for full, and SSE_r and SSE_f are SSE for residual and full models.

Unit 3: Classification Accuracy

Validation of Simple Linear Regression Models:

- Model F-Test value should be greater than table F-Test value for corresponding values of m, k, n values.
- It indicates the overall significance of whether the Regression model provides a good fit to data with all independent variables.
- The model that does not contain independent variables implies that no relationship exists with independent attributes.
- This type of model w/o independent variables is called the “Intercept only” model with correlation coefficient = 0.

Unit 3: Classification Accuracy

Validation of Simple Linear Regression Models:

- A set of Hypothesis testing is formulated to test the significance of inclusion of independent variables.
- A Hypothesis is a precise and declared statement that is related to the outcome of a study. In the present case, it is the Regression Model.
- The Null-Hypothesis (H0) states that the coefficients are zero, meaning that the dependent variable is not related to the independent variables.
- Therefore, Null-Hypothesis (H0) and Alternate Hypothesis (H1):
  - H0: $\beta_j = 0$, for all j = 1 to m. But $\beta_0 \neq 0$, i.e. it is the intercept only model.
  - H1: $\beta_j \neq 0$ for all j = 1 to m.
- In this Hypothesis, the P-Value is set to determine the F-Statistics.
- P-Value is the desired significance level.

Unit 3: Classification Accuracy

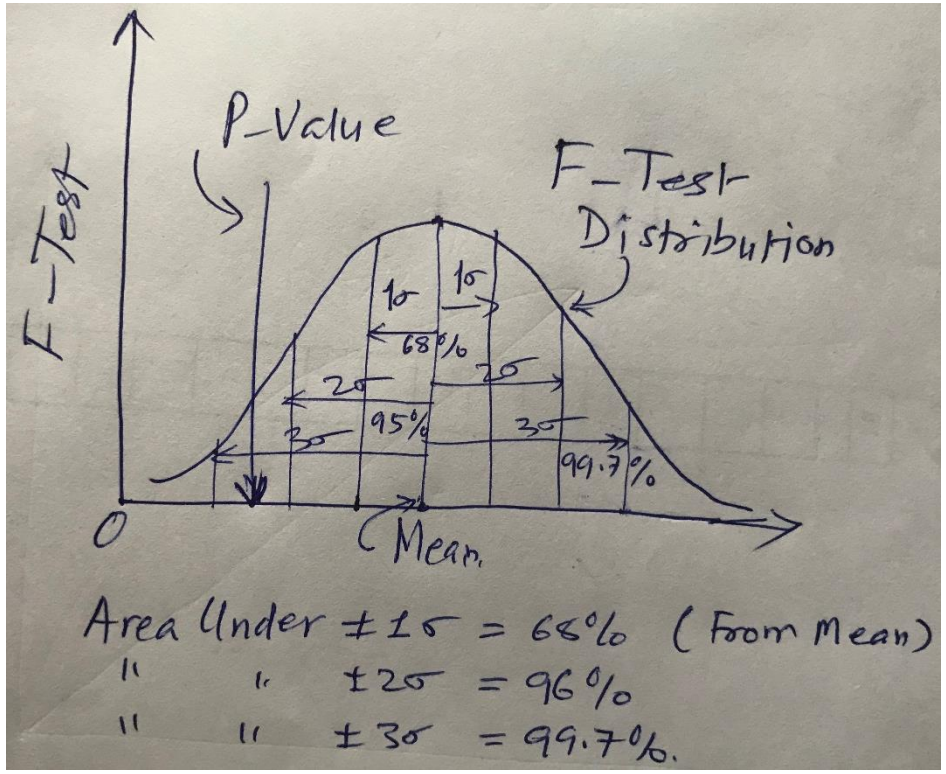
Validation of Simple Linear Regression Models:

- In this Hypothesis, the P-Value is set to determine the F-Test Statistic.
- The min and max interval of F-Test Statistic is determined by the α value set for the accuracy.
- The α is the probability of making an error.
- The α values are typically 0.01, 0.05, 0.10 for 1%, 5%, 10% level of Significance, or for 99%, 95% and 90% accuracy level, respectively.
- At low DoF, the F-Test statistic has right skewed distribution.
- For large DoF, F-Test statistic shows near Normal Distribution.

Unit 3: Classification Accuracy

Validation of Simple Linear Regression Models:

- If the P-value < α-value, then there is evidence to conclude that the coefficients are non-zero.
- Then, the Regression Model fits the data better with all features/attributes, rather than the model w/o all features.
- If P-value is > α-value, then the F-test results are not significant.
- This implies that the Regression Model w/o features fit the data very well.
- This means that the Regression Model is invalid.



Unit 3: k Fold Cross Validation

K Fold Cross Validation Method: Model Validation.

- Normally the data set is divided in at least two sets, viz. training set and Testing set.
- Sometimes a small portion is reserved as Validation set.
- Training set size is kept large for good model development, as model reliability depends on the data size, i.e. number of samples or examples in training set.
- Nearly 80 to 90 percent of the data is randomly selected from the total data set and designated as training data.
- From the remaining 10 to 20 percent, sometimes a small percent (about 2 to 5%) samples are reserved for model testing/validation during actual model development.
- If not needed the whole 10 to 20 % samples data are reserved as Testing Data.
- In our discussion, we assume no validation set is chosen.

Unit 3: k Fold Cross Validation

k Fold Cross Validation Method:

- In k-Fold cross validation method, the data set is divided in to k sections from the randomly arranged data set.
- For example, if k = 10, then there 10 sections selected from the randomly arranged data set.
- Out of these 10, any 9 are selected as training samples and 1 is selected as testing data set.
- Model accuracy evaluation is done based on the model application on the test data.
- Above exercise is carried out 10 times by selecting any 9 segments at a time as training set and left out 1 segment as testing set.

Unit 3: k Fold Cross Validation

K Fold Cross Validation Method:

- The above exercise will result in 10 sets of results on accuracy evaluation.
- But 10 is statistically too small or not a significant a data set for reliable statistical parameter estimations.
- Nearly minimum 30 to 40 are required for reliable or significant Statistical tests.
- If the data set is very big in size, it may be possible to generate 30 to 40 training segments values by dividing data into 30 to 40 segments with large good number of examples in each segment.
- But may not be possible when data set is small.
- In case of small data sets, reshuffle the data into 3 to 4 times and apply 10-fold validation to generate statistically significant 30 to 40 test results.
- Reshuffling is done with a new set of random numbers every time.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

9

Unit 3: Classification Accuracy

Classification Model Performance:

- The classification accuracy of a model can be tested or evaluated using various performance metrics.
- They are: (i) Confusion Matrix, (ii) Area under the Curve (AUC) of Receiver Operator Characteristics (ROC), (iii) Gini Test and (iv) Kolmogorov-Smirnov test.
- WE will be concentrating the first two methods only, i.e. Confusion Matrix and AUC/ROC.
- Confusion Matrix: This method provides the overall accuracy, sensitivity and Specificity of classification model.
- This method is most popular. It can be used for both binary as well as multiple class classification.

Unit 3: Classification Accuracy

Classification Model Performance:

- The confusion matrix has two columns and two rows (i.e. 2x2 Matrix) in case of binary classification.
- The rows are Trues and False and columns are Positive and Negative (Predicted).
- The cells of the matrix in rows are True Positives, False Positives and False Positive and False Negative.
- If a classification model predicts the matching positive class correctly, then the outcome is called True Positive (TP).
- If the model fails to predict a correct class i.e. it predicts wrongly, then the outcome is False Negative (TN)
- If the model predicts a positive class incorrectly, then the outcome is False Positive (FP).
- If it predicts a non-class wrongly (not the class of interest), then the outcome is called False Negative (FN).

Unit 3: Classification Accuracy

Classification Model Performance:

- To summarize:

  True Positive (TP): Correct classification of interested class

  True Negative (TN): Correct classification of a class of not interest

  False Positive (FP): Incorrect classification of class of interest (Omission)

  False Negative (FN): Incorrect classification of a class of not interest (Commission)

- A matrix of these alternate outcomes of each class is Called Confusion Matrix →

Actual /Predicted	Positive (Predicted)	Negative (Predicted)
True (Actual)	True Positive (TP)	False Positive(FP) (Omission)
False (Actual)	False Negative (FN) (Commission)	True Negative (FP)

Unit 3: Classification Accuracy

Classification Model Performance:

- The quantities derived from the Confusion Matrix are as follows:

  Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ Precision = $\frac{TP}{TP+FP}$

  Sensitivity = $\frac{TP}{TP+FN}$ Specificity = $\frac{TN}{TN+FN}$

- Sensitivity is a Evaluation Measure of Positives.
- Sensitivity is also known as Recall.
- Specificity is evaluation measure of Negatives.
- For a Good–Fit model, the accuracy should be high (~ 100% ideally)
- Sensitivity and Specificity need not be always high.

Unit 3: Classification Accuracy

Classification Model Performance:

- Types of Classification Models are characterized by:
  - Good Models: High Accuracy + High Sensitivity
  - Poor Models: High Accuracy + Low Specificity
  - Not Good Models: Low Accuracy + Low Sensitivity
  - Bad Models: Low Accuracy + High Specificity
- Both False Negatives and False Positives should be low.
- This means both the omission and commission errors should be low.
- These above criteria lead to conclusion that high Sensitivity and High Specificity models are the best.
- Based on Confusion matrix values, Kappa Coefficient is derived, which is a measure of overall performance of a classifier.

Unit 3: Classification Accuracy

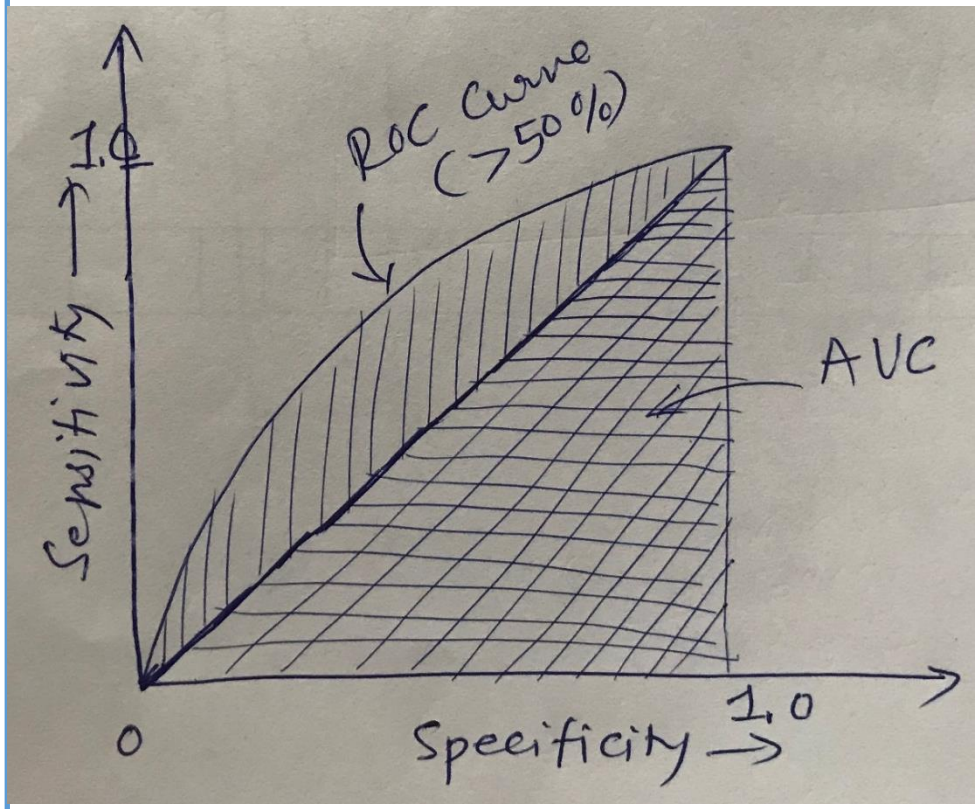
Area Under the Curve in ROC method:

- Before we deal with Area Under the Curve (AUC) method, we need to understand ROC Curve (Receiver Operator Characteristics).
- ROC is used to evaluate the performance of a Classification model.
- This method was used during WW-II (1939-1945) to analyse Radar Received Data and discriminate between Enemy Aircraft and other Radar Noise signals.
- The ROC plot shows the relationship between Sensitivity (total positivity) plotted on Y-Axis and Specificity (total negativity) plotted on X-Axis.

Unit 3: Classification Accuracy

Area Under the Curve (AUC) in ROC method:

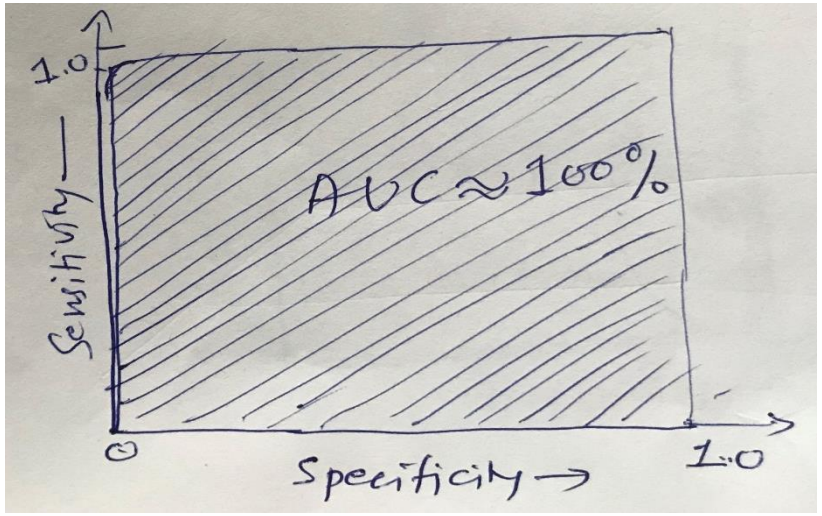
- When the specificity decreases, the sensitivity does not decrease rapidly as per specificity.
- The values of the sensitivity and the specificity range from 0 to 1.
- If the Area Under the ROC Curve is more than that under the diagonal line (i.e. closer to top), then the accuracy of the model is high.
- If the ROC curve is closer to the diagonal (~50%), the accuracy is less.
- As such, 50% is no prediction.

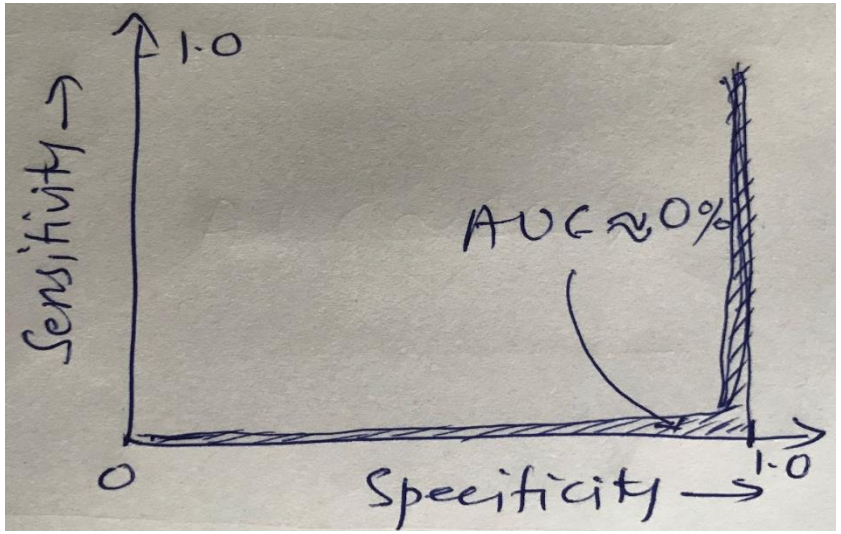


Unit 3: Classification Accuracy

Area Under the Curve (AUC) in ROC method:

- The ROC is in fact a measure of the AUC.
- The AUC indicates the degree of separability of Positivity from Negativity.
- When AUC is 100 %, the model can classify all the positive and negative classes perfectly.
- High AUC (>> 50%) is needed for the model to be well fitting.





Unit 3: Classification Accuracy

Area Under the Curve (AUC) in ROC method:

- If AUC is more than 50 %, then there is fair/good chance that the model can differentiate between positive and negative classes.
- In nutshell, high AUC is needed for the model to be considered fit.
- Higher the AUC, higher the accuracy and hence higher reliability of the classification model under consideration.
- This method needs the results of multiple application of the classification model under scrutiny to generate large number of sensitivity and specificity values and generate the ROC plot.
- Large number set can be generated by k-Fold Cross validation also.

APPLIED MACHINE LEARNING

UNIT-III: CLASSIFICATION

END