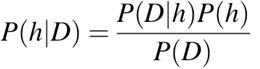
# Naïve Bayes Classifier:

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.



Advantages:

* Simple to Implement. The conditional probabilities are easy to evaluate.
* Very fast – no iterations since the probabilities can be directly computed. So this technique is useful where speed of training is important.
* If the conditional Independence assumption holds, it could give great results.
* Not sensitive to outliers.

Disadvantages

* Conditional Independence Assumption does not always hold. In most situations, the feature shows some form of dependency.
* Bad binning of continuous variables with Multinomial naive bayes: Gaussian Naive Bayes
* Not great for imbalanced data.

# Support Vector Machines (SVM)

The intention of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that separately classifies the data points.

Advantages of support vector machine:

Support vector machine works comparably well when there is an understandable margin of dissociation between classes.

It is more productive in high dimensional spaces.

It is effective in instances where the number of dimensions is larger than the number of specimens.

Support vector machine is comparably memory systematic.

Disadvantages of support vector machine:

* Support vector machine algorithm is not acceptable for large data sets.
* It does not execute very well when the data set has more sound i.e. target classes are overlapping.
* In cases where the number of properties for each data point outstrips the number of training data specimens, the support vector machine will underperform.
* As the support vector classifier works by placing data points, above and below the classifying hyperplane there is no probabilistic clarification for the classification.

# 3. Linear Regression

Regression analysis is used to estimate the relationship between a dependent variable and one or more independent variables.

Missing Values

It is sensitive to missing values.

Outliers

Linear Regression needs the relationship between the independent and dependent variables to be linear.

It is also important to check for outliers since linear regression is sensitive to outlier effects.

The outliers have an effect on different accuracy measures of a linear regression model and can further lead to errors in estimations as well.

Advantages

* Linear Regression is simple to implement and easier to interpret the output coefficients.
* When you know the independent and dependent variable have a linear relationship, this algorithm is the best to use because it’s less complex as compared to other algorithms.
* Linear Regression is prone to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages

* Sometimes lot of feature engineering is required.
* If the independent features are correlated it may affect performance.
* It is often quite prone to noise and overfitting.

# Logistic Regression

 Logistic Regression is used when the dependent variable(target) is categorical.

Advantages

* Logistic regression is easier to implement, interpret, and very efficient to train.
* It makes no assumptions about distributions of classes in feature space.
* It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions.
* It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative)
* Good accuracy for many simple data sets and it performs well when the dataset is linearly separable.
* It can interpret model coefficients as indicators of feature importance.

Disadvantages

* If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
* It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.
* Non-linear problems can’t be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios.
* It is tough to obtain complex relationships using logistic regression. More powerful and compact algorithms such as Neural Networks can easily outperform this algorithm.
* Logistic Regression models are not much impacted due to the presence of outliers because the sigmoid function tapers the outliers.
* But the presence of extreme outliers may somehow affect the performance of the model and lowering the performance.

# Decision Tree Regression or Classifier

* Not sensitive to outlier
* Decision Tree can automatically **handle missing values**.
* Decision Tree is usually **robust to outliers** and can handle them automatically.
* **Advantages of Decision Tree**  
    
  **1. Clear Visualization**: The algorithm is simple to understand, interpret and visualize as the idea is mostly used in our daily lives. Output of a Decision Tree can be easily interpreted by humans.  
    
  **2. Simple and easy to understand**: Decision Tree looks like simple **if-else statements**which are very easy to understand.  
    
  **3.** Decision Tree can be used for both **classification and regression problems**.  
    
  **4**. Decision Tree can handle both **continuous and categorical variables**.  
    
  **5.** **No feature scaling required**: No feature scaling (standardization and normalization) required in case of Decision Tree as it uses rule based approach instead of distance calculation.  
    
  **6. Handles non-linear parameters efficiently:** Non linear parameters don't affect the performance of a Decision Tree unlike curve based algorithms. So, if there is high non-linearity between the independent variables, Decision Trees may outperform as compared to other curve based algorithms.  
    
  **7**.**Less Training Period**: Training period is less as compared to Random Forest because it generates only one tree unlike forest of trees in the Random Forest.   
    
  **Disadvantages of Decision Tree**  
    
  **1. Overfitting:** This is the main problem of the Decision Tree. It generally leads to overfitting of the data which ultimately leads to wrong predictions. In order to fit the data (even noisy data), it keeps generating new nodes and ultimately the tree becomes too complex to interpret. In this way, it loses its generalization capabilities. It performs very well on the trained data but starts making a lot of mistakes on the unseen data.

**2. High variance:**As mentioned in point 1, Decision Tree generally leads to the overfitting of data. Due to the overfitting, there are very high chances of high variance in the output which leads to many errors in the final estimation and shows high inaccuracy in the results. In order to achieve zero bias (overfitting), it leads to high variance.   
  
**3. Unstable:**Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.   
  
**4. Affected by noise:** Little bit of noise can make it unstable which leads to wrong predictions.  
  
**5. Not suitable for large datasets:** If data size is large, then one single tree may grow complex and lead to overfitting. So in this case, we should use Random Forest instead of a single Decision Tree.

# Ensemble (RF, XGboost, GB)

Advantages of ensemble methods

Not sensitive to outliers

1. Ensemble methods have higher predictive accuracy, compared to the individual models.

2. Ensemble methods are very useful when there is both linear and non-linear type of data in the dataset; different models can be combined to handle this type of data.

3. With ensemble methods bias/variance can be reduced and most of the times, model is not underfitted/overfitted.

4. Ensemble of models is always less noisy and is more stable.

Disadvantages of Ensemble learning

1. Ensembling is less interpretable, the output of the ensembled model is hard to predict and explain. Hence the idea with ensemble is hard to sell and get useful business insights.

2. The art of ensembling is hard to learn and any wrong selection can lead to lower predictive accuracy than an individual model.

3. Ensembling is expensive in terms of both time and space. Hence ROI can increase with ensembling.

# K Nearest Neighbor Algoritm

Advantages:

1. **No Training Period**- KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction and because of this it is very time efficient in term of improvising for a random modeling on the available data.
2. **Easy Implementation**- KNN is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be calculated using distance formula such as- Euclidian or Manhattan
3. As there is no training period thus new data can be added at any time since it wont affect the model.
4. Not sensitive to outliers

Disadvantages:

1. **Does not work well with large dataset**as calculating distances between each data instance would be very costly.
2. **Does not work well with high dimensionality**as this will complicate the distance calculating process to calculate distance for each dimension.
3. Sensitive to noisy and missing data
4. **Feature Scaling-**Data in all the dimension should be scaled (normalized and standardized) properly.

# KMeans Algorithm

# Advantages

* It is fast
* Robust
* Easy to understand
* Comparatively efficient
* If data sets are distinct, then gives the best results
* Produce tighter clusters
* When centroids are recomputed, the cluster changes.
* Flexible
* Easy to interpret
* Better computational cost
* Enhances Accuracy
* Works better with spherical clusters

# Disadvantages

* Needs prior specification for the number of cluster centers
* If there are two highly overlapping data, then it cannot be distinguished and cannot tell that there are two clusters
* With the different representations of the data, the results achieved are also different
* Euclidean distance can unequally weigh the factors
* It gives the local optima of the squared error function
* Sometimes choosing the centroids randomly cannot give fruitful results
* It can be used only if the meaning is defined
* Cannot handle outliers and noisy data
* Do not work for the non-linear data set
* Lacks consistency
* Sensitive to scale
* If very large data sets are encountered, then the computer may crash.

# Hierarchical Algorithm

Most hierarchical clustering software does not work with values are missing in the data.

Also sensitive to outliers.

Advantages

• Hierarchical clustering outputs a hierarchy, ie a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram

. • Easy to implement

Disadvantages

• It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.

• Time complexity: not suitable for large datasets

• Initial seeds have a strong impact on the final results

• The order of the data has an impact on the final results

• Very sensitive to outliers

# Principal Component Analysis

Advantages of PCA:

* Easy to compute. PCA is based on linear algebra, which is computationally easy to solve by computers.
* Speeds up other machine learning algorithms. Machine learning algorithms converge faster when trained on principal components instead of the original dataset.
* Counteracts the issues of high-dimensional data. High-dimensional data causes regression-based algorithms to overfit easily. By using PCA beforehand to lower the dimensions of the training dataset, we prevent the predictive algorithms from overfitting.

Disadvantages of PCA:

* Low interpretability of principal components. Principal components are linear combinations of the features from the original data, but they are not as easy to interpret. For example, it is difficult to tell which are the most important features in the dataset after computing principal components.
* The trade-off between information loss and dimensionality reduction. Although dimensionality reduction is useful, it comes at a cost. Information loss is a necessary part of PCA. Balancing the trade-off between dimensionality reduction and information loss is unfortunately a necessary compromise that we have to make when using PCA.
* PCA is not robust against outliers. The algorithm will be biased in datasets with strong outliers. This is why it is recommended to remove outliers before performing PCA.

# Neural Networks

**Advantages of Artificial Neural Networks**

* Artificial neural networks have the ability to provide the data to be processed in parallel, which means they can handle more than one task at the same time.
* Artificial neural networks have been in resistance. This means that the loss of one or more cells, or neural networks, influences the performance of Artificial Neural networks.
* Artificial neural networks are used to store information on the network so that, even in the absence of a data pair, it does not mean that the network is not generating results.
* Artificial neural networks are gradually being broken down, which means that they will not suddenly stop working and these networks are gradually being broken down.
* We are able to train ANN’s that these networks learn from past events and make decisions.

### **Disadvantages of Artificial Neural Networks**

* With ANN arms hanging along with the execution of parallel processing, and so they need processors that support parallel processing, so the ANNs are dependent on the hardware.
* Since it’s similar to the functionality of the human brain, we may not be able to determine what is the proper network structure of an Artificial Neural network.
* Not only do artificial neural networks, but also the statistical models can be trained with only numeric data, so it makes it very difficult for ANN to understand the problem statement.
* When an artificial neural network that provides a solution to the problem statements that we really don’t know on what basis it will give the solution, and this time, ANN is not a reliable