Advanced Machine Learning Algorithms

Naive Bayes Classifier

Naïve Bayes Classifier is a family of classifiers that uses Bayes Theorem of Probability to build machine learning models. This classifier is particularly powerful for disease prediction and document classification. The basic assumption for the Naive Bayes algorithm is that all the features are considered to be independent of each other.

Bayes theorem gives a way to calculate posterior probability P(A|B) from P(A), P(B), and P(B|A).

The formula is given by:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

where P(A|B) is the posterior probability of A given B, P(A) is the prior probability, P(B|A) is the likelihood which is the probability of B given A, and P(B) is the prior probability of B. In simple English this can be written as:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

When to use the Naive Bayes Classifier algorithm?

1. Naive Bayes is best in cases with a moderate or large training dataset.
2. It works well for dataset instances that have several attributes.
3. Given the classification parameter, attributes that describe the instances should be conditionally independent.

Applications of Naive Bayes Classifier

1. Sentiment Analysis: It is used by Facebook to analyse status updates expressing positive or negative emotions.
2. Document Categorization: Google PageRank uses document classification to index documents and finds relevancy scores. PageRank mechanism considers the pages marked as important in the databases parsed and classified using a document classification technique.
3. This algorithm is also used for classifying news articles about Technology, Entertainment, Sports, Politics, etc.
4. Email Spam Filtering: GMail uses the Naive Bayes algorithm to classify your emails as Spam or Not Spam.

Advantages of the Naive Bayes Classifier Algorithm

1. Simple and easy to implement.
2. Does not require as much training data.
3. Handles both continuous and discrete data.
4. Highly scalable with the number of predictors and data points.
5. Fast to train and classify, good for real-time predictions.

Advantages of the Naive Bayes Classifier Algorithm

1. Assumes independence of features

K Means Clustering Algorithm

K-means is a popularly used unsupervised ML algorithm for cluster analysis. K-Means is a non-deterministic and iterative method. The algorithm operates on a given data set through a pre-defined number of clusters, k. The output of the K Means algorithm is k clusters with input data partitioned among the clusters. For instance, let's consider K-Means Clustering for Wikipedia Search results. The search term "Jaguar" on Wikipedia will return all pages containing the word Jaguar which can refer to Jaguar as a Car, Jaguar as Mac OS version, and Jaguar as an Animal. K Means clustering algorithm can be applied to group the web pages that talk about similar concepts. So, the algorithm will group all the web pages that refer to Jaguar as an Animal into one cluster, Jaguar as a Car into another cluster, and so on.

For any new incoming data point, the data point is classified according to its proximity to the nearby classes. Datapoints inside a cluster will exhibit similar characteristics while the other clusters will have different properties. The primary example of clustering would be grouping the same customers in a particular class for any marketing campaign, and it is also a practical algorithm for document clustering.

Let's say we have x_1, x_2, x_3…… $x_{(n)}$ as our inputs, and we want to split this into k clusters.

The steps to form clusters are:

1. Choose k random points as cluster centers called centroids.
2. Assign each $x_{(i)}$ to the closest cluster by implementing euclidean distance (i.e., calculating its distance to each centroid)
3. Identify new centroids by taking the average of the assigned points.
4. Keep repeating step 2 and step 3 until convergence is achieved.

Advantages of using K-Means Clustering

1. Relatively simple to implement.
2. Scales to large data sets.
3. Guarantees convergence.
4. Can warm-start the positions of centroids.
5. Easily adapts to new examples.

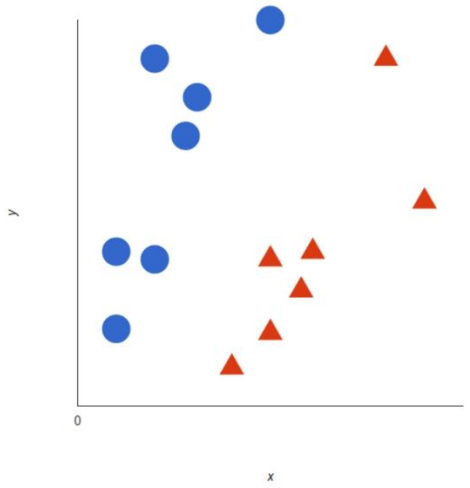
Applications of K-Means Clustering

K-Means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression. Most search engines like Yahoo and Google use the K Means Clustering algorithm to cluster web pages by similarity and identify the 'relevance rate' of search results. This helps search engines reduce the computational time for the users.

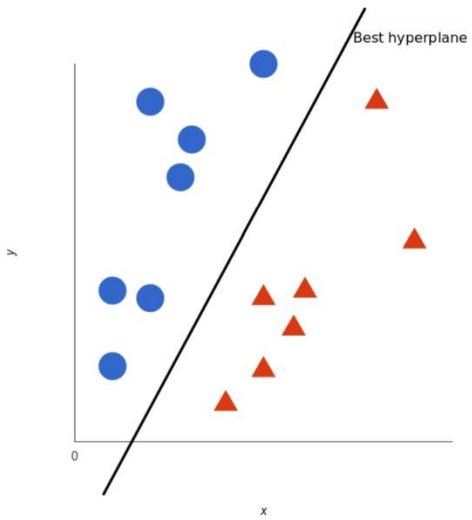
Support Vector Machines

Support Vector Machines are a set of supervised learning methods used for classification, regression and outliers detection. It organizes the data into different categories by finding a line (hyperplane) separating the training data set into classes. As there are many such linear hyperplanes, the SVM algorithm tries to maximize the distance between the various classes involved, referred to as margin maximization. If the line that maximizes the distance between the classes is identified, the probability of generalizing well to unseen data is increased.

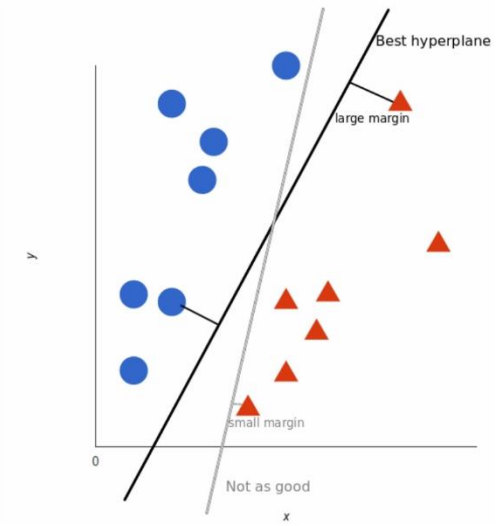
For a simplest case let us imagine we have two tags: red and blue, and our data has two features: x and y. We want a classifier that, given a pair of (x, y) coordinates, outputs if it's either red or blue. We plot our already labelled training data on a plane:



SVM takes these data points and outputs the hyperplane (which in two dimensions is simply a line) that best separates the tags. This line is the decision boundary: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red.



What exactly is the best hyperplane? It is the one that maximizes the margins from both tags. In other words, the hyperplane whose distance to the nearest element of each tag is the largest.



Advantages of Using SVM

1. SVM offers the best classification performance (accuracy) on the training dataset.
2. SVM renders more efficiency for the correct classification of future data.
3. The best thing about SVM is that it does not make strong assumptions about data.
4. It does not overfit the data.

Applications of Support Vector Machine

SVM is commonly used for stock market forecasting by various financial institutions. For instance, one can use it to compare the relative performance of the stocks to those of other stocks in the same sector. The close comparison of stocks helps manage investment-making decisions based on the classifications made by the SVM learning algorithm. Other applications include image classification, image segmentation, encryption, sentiment analysis, speech recognition.

Apriori Algorithm

Apriori algorithm is an unsupervised ML algorithm that generates association rules from a given data set. The Association rule implies that if item A occurs, then item B also occurs with a certain probability. Most of the association rules generated are in the IF_THEN format. For example, IF people buy an iPad, they also buy an iPad Case to protect it. For the algorithm to derive such conclusions, it first observes the number of people who bought an iPad case while purchasing an iPad. This way a ratio is derived like out of the 100 people who purchased an iPad, 85 people also purchased an iPad case.

Principle on which Apriori Algorithm works

1. If an item set frequently occurs, then all the subsets of the item set also happen often.
2. If an item set occurs infrequently, then all the supersets of the item set have infrequent occurrences.

Advantages of Apriori Algorithm

1. It is easy to implement and can be parallelized easily.
2. Apriori implementation makes use of large item set properties.

Applications of Apriori Algorithm

1. Detecting Adverse Drug Reactions: Apriori algorithm is used for association analysis on healthcare data like the drugs taken by patients, characteristics of each patient, adverse ill-effects patients experience, initial diagnosis, etc. This analysis produces association rules that help identify the combination of patient characteristics and medications that lead to adverse side effects of the drugs
2. Market Basket Analysis: For example, a retailer might use Apriori to predict that people who buy sugar and flour will likely buy eggs to bake a cake
3. Auto-Complete Applications: when the user types a word, the search engine looks for other associated words that people usually type after a specific word.

Logistic Regression

Logistic regression algorithm is used to estimate discrete values in classification tasks and not regression problems. The word 'regression' here implies that a linear model is fit into the feature space. This algorithm applies a logistic function to a linear combination of features to predict the outcome of a categorical dependent variable based on predictor variables. The odds or probabilities that describe the result of a single trial are modelled as a function of explanatory variables. This algorithm helps estimate the likelihood of falling into a specific level of the categorical dependent variable based on the given predictor variables.

Suppose you want to predict if there will be a rainfall tomorrow in Thyolo. Here the prediction outcome is not a continuous number because there will either be rainfall or no rainfall, so simple linear regression cannot be applied. Here the outcome variable is one of the several categories, and logistic regression helps.

When to Use Logistic Regression?

1. When there is a requirement to model the probabilities of the response variable as a function of some other explanatory variable. For example, the probability of buying a product X as a function of gender
2. When there is a need to predict probabilities that categorical dependent variables will fall into two categories of the binary response as a function of some explanatory variables. For example, what is the probability that a customer will buy a perfume given that the customer is a female?
3. When the need is to classify elements into two categories based on the explanatory variable. For example-classify females into 'young' or 'old' group based on their age.

Advantages of Logistic Regression

1. Easier to inspect and less complex.
2. Robust algorithm as the independent variables need not have equal variance or normal distribution.
3. These algorithms do not assume a linear relationship between the dependent and independent variables and hence can also handle non-linear effects.
4. Controls confounding and tests interaction.

Drawbacks of Using Logistic Regression

1. May overfit the training dataset when the training dataset is sparse and high dimensional.
2. Requires more data to achieve stability and meaningful results. These algorithms require a minimum of 50 data points per predictor to achieve stable outcomes.
3. Predicts outcomes depending on a group of independent variables and if a data scientist or a machine learning expert goes wrong in identifying the independent variables then the developed model will have minimal or no predictive value.
4. Not robust to outliers and missing values.

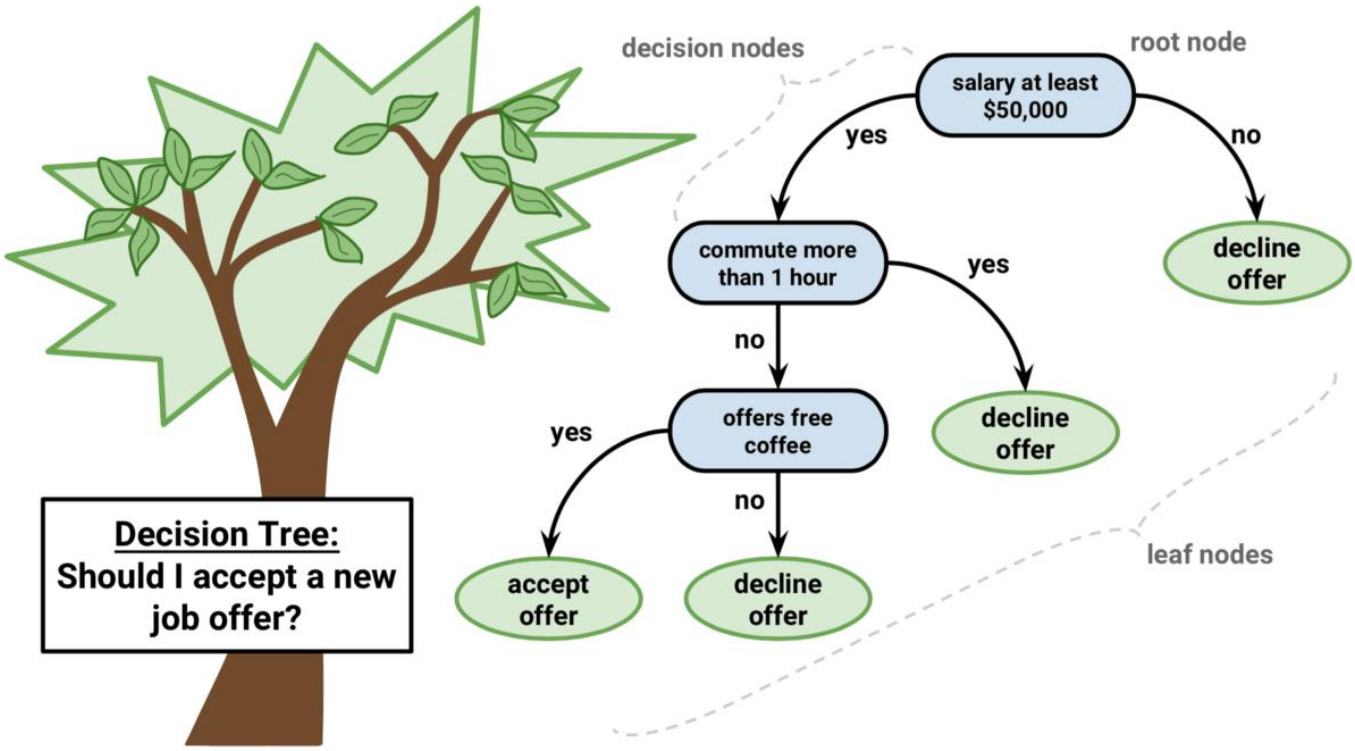
Applications of Logistic Regression

- In epidemiology to identify risk factors for diseases and plan accordingly for preventive measures.
- To predict whether a candidate will win or lose a political election.
- To classify a set of words as nouns, pronouns, verbs, adjectives.
- In weather forecasting to predict the probability of rainfall.
- In credit scoring systems for risk management to predict the defaulting of an account.

Decision Tree

A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes. The classification rules are represented through the path from root to the leaf node.

Here is an example:



When to use Decision Tree?

1. Decision trees are robust to errors and if, the training dataset contains errors- decision tree algorithms will be best suited to address such problems.
2. They are best suited for problems where instances are represented by attribute value pairs.
3. If the training dataset has missing value then decision trees can be used, as they can handle missing values nicely by looking at the data in other columns.
4. They are best suited when the target function has discrete output values.

Advantages of Using Decision Tree

1. Decision trees are very instinctual and can be explained to anyone with ease. People from a non-technical background can also decipher the hypothesis drawn from a decision tree, as they are self-explanatory.
2. When using this algorithm, data type is not a constraint as they can handle both categorical and numerical variables.
3. Decision tree machine learning algorithms do not require making any assumption on the linearity in the data and hence can be used in circumstances where the parameters are non-linearly related. These machine learning algorithms do not make any assumptions on the classifier structure and space distribution.
4. These algorithms are useful in data exploration. Decision trees implicitly perform feature selection which is very important in predictive analytics. When a decision tree is fit to a

training dataset, the nodes at the top on which the tree is split, are considered as important variables within a given dataset and feature selection is completed by default.
5. Decision trees help save data preparation time, as they are not sensitive to missing values and outliers. Missing values will not stop you from splitting the data for building a decision tree. Outliers will also not affect the decision trees as data splitting happens based on some samples within the split range and not on exact absolute values.

Drawbacks of Using Decision Tree

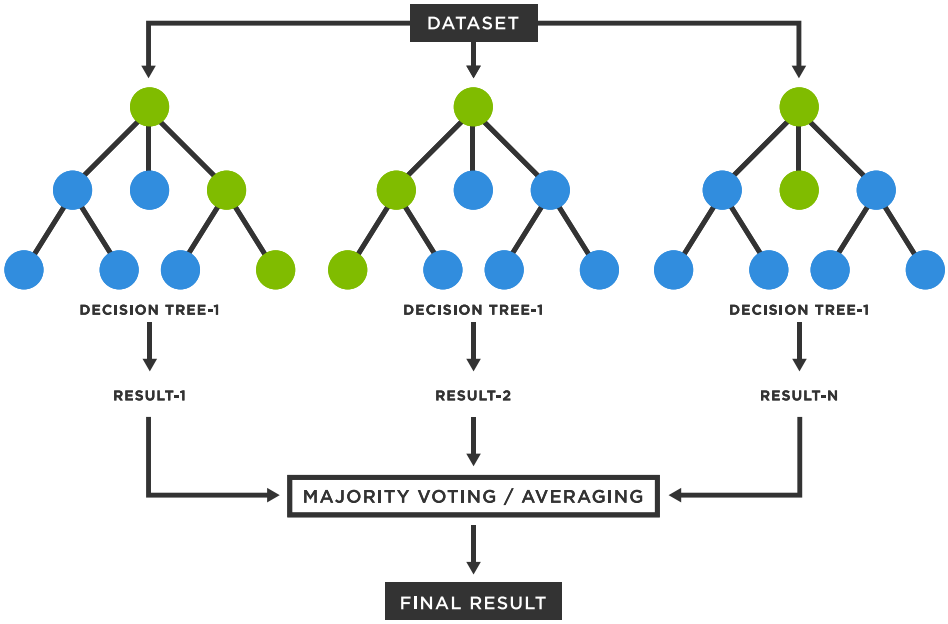
1. The more the number of decisions in a tree, less is the accuracy of any expected outcome.
2. A major drawback of this machine learning algorithm is that the outcomes may be based on expectations. When decisions are made in real-time, the payoffs and resulting outcomes might not be the same as expected or planned. There are chances that this could lead to unrealistic decision trees leading to bad decision making. Any irrational expectations could lead to major errors and flaws in decision tree analysis, as it is not always possible to plan for all eventualities that can arise from a decision.
3. Decision Trees do not fit well for continuous variables and result in instability and classification plateaus.
4. Decision trees are easy to use when compared to other decision-making models but creating large decision trees that contain several branches is a complex and time consuming task.
5. Decision tree considers only one attribute at a time and might not be best suited for actual data in the decision space.
6. Large sized decision trees with multiple branches are not comprehensible and pose several presentation difficulties.

Applications of Decision Tree

- In finance for option pricing.
- Remote sensing is an application area for pattern recognition based on decision trees.
- By banks to classify loan applicants by their probability of defaulting payments.
- Health facilities to identify at-risk patients and disease trends.

Random Forest

Random Forest is the go-to algorithm that uses a bagging approach to create a bunch of decision trees with random subset of the data. A model is trained several times on random sample of the dataset to achieve good prediction performance from the random forest algorithm. In this ensemble learning method, the output of all the decision trees in the random forest is combined to make the final prediction. The final prediction of the random forest algorithm is derived by polling the results of each decision tree or just by going with a prediction that appears the most times in the decision trees.



Why use Random Forest Algorithm?

1. It maintains accuracy when there is missing data and is also resistant to outliers.
2. Simple to use as the basic random forest algorithm can be implemented with just a few lines of code.
3. Random Forest machine learning algorithms help data scientists save data preparation time, as they do not require any input preparation and can handle numerical, binary and categorical features, without scaling, transformation or modification.
4. Implicit feature selection as it gives estimates on what variables are important in the classification.

Advantages of Using Random Forest

1. Overfitting is less of an issue with Random Forests. Unlike decision tree machine learning algorithms, there is no need of pruning the random forest.
2. These algorithms are fast but not in all cases. A random forest algorithm, when run on an 800 MHz machine with a dataset of 100 variables and 50,000 cases produced 100 decision trees in 11 minutes.
3. Random Forest is one of the most influential and versatile algorithm for wide variety of classification and regression tasks, as they are more robust to noise.
4. It is difficult to build a bad random forest. In the implementation of Random Forest Machine Learning algorithms, it is easy to determine which parameters to use because they are not

sensitive to the parameters that are used to run the algorithm. One can easily build a decent model without much tuning.
5. Random Forest machine learning algorithms can be grown in parallel.
6. This algorithm runs efficiently on large databases.
7. Has higher classification accuracy.

Drawbacks of Using Random Forest

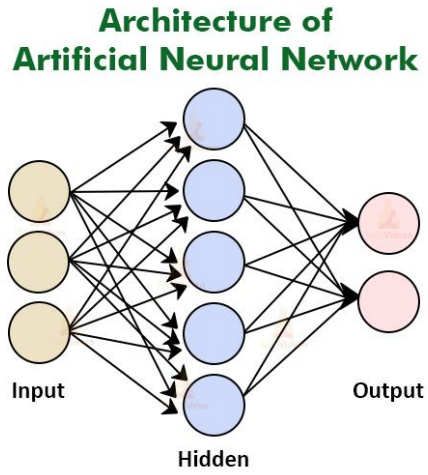
1. They might be easy to use but analysing them theoretically, is difficult.
2. Large number of decision trees in the random forest can slow down the algorithm in making real-time predictions.
3. If the data consists of categorical variables with different number of levels, then the algorithm gets biased in favour of those attributes that have more levels. In such situations, variable importance scores do not seem to be reliable.
4. When using RandomForest algorithm for regression tasks, it does not predict beyond the range of the response values in the training data.

Applications of Random Forest

- Random Forest algorithms are used by banks to predict if a loan applicant is a likely high risk.
- They are used in the automobile industry to predict the failure or breakdown of a mechanical part.
- These algorithms are used in the healthcare industry to predict if a patient is likely to develop a chronic disease or not.
- They can also be used for regression tasks like predicting the average number of social media shares and performance scores.
- Predicting patterns in speech recognition software and classifying images and texts.

Artificial Neural Networks

ANNs consist of input, hidden, and output layers with connected neurons (nodes) to simulate the human brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives signals then processes them and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.



Advantages of Using ANNs

1. Parallel processing abilities mean the network can perform more than one job at a time.
2. Information is stored on an entire network, not just a database.
3. The ability to learn and model nonlinear, complex relationships helps model the real-life relationships between input and output.
4. Fault tolerance; the corruption of one or more cells of the ANN will not stop the generation of output.
5. Gradual corruption; the network will slowly degrade over time, instead of a problem destroying the network instantly.
6. The ability to produce output with incomplete knowledge with the loss of performance being based on how important the missing information is.
7. No restrictions are placed on the input variables, such as how they should be distributed.
8. The ability to learn hidden relationships in the data without commanding any fixed relationship means an ANN can better model highly volatile data and non-constant variance.
9. The ability to generalize and infer unseen relationships on unseen data means ANNs can predict the output of unseen data.

Disadvantages of Using ANNs

1. The lack of rules for determining the proper network structure means the appropriate artificial neural network architecture can only be found through trial and error and experience.
2. The requirement of processors with parallel processing abilities makes neural networks hardware-dependent.
3. The network works with numerical information; therefore, all problems must be translated into numerical values before they can be presented to the ANN.

4. The lack of explanation behind probing solutions is one of the biggest disadvantages in ANNs. The inability to explain the why or how behind the solution generates a lack of trust in the network.

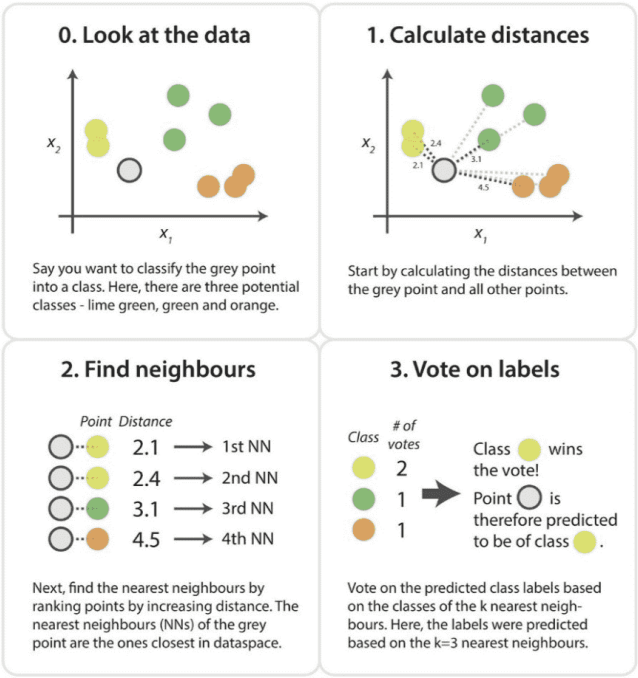
Applications of ANNs

- Image recognition
- Chatbots
- Natural language processing, translation and language generation
- Stock market prediction
- Delivery driver route planning and optimization
- Drug discovery and development.

K-Nearest Neighbors

KNN is the most straightforward classification algorithm. It is also used for the prediction of continuous values like regression. Distance-based measures are used in K Nearest Neighbors to get the correct prediction. The final prediction value is chosen based on the k neighbors. The various distance measures used are Euclidean, Manhattan, Minkowski, and Hamming distances. The first three are continuous functions, while Hamming distance is used for categorical variables. Choosing the value of K is the most essential task in this algorithm. It is often referred to as the lazy learner algorithm. A lazy learning algorithm is simply an algorithm where the algorithm generalizes the data after a query is made.

Steps in KNN:



Advantages of Using K-Nearest Neighbors

1. High accuracy but better algorithms exist.
2. It's very useful for non-linear data as there are no assumptions here.

Disadvantages of Using K-Nearest Neighbors

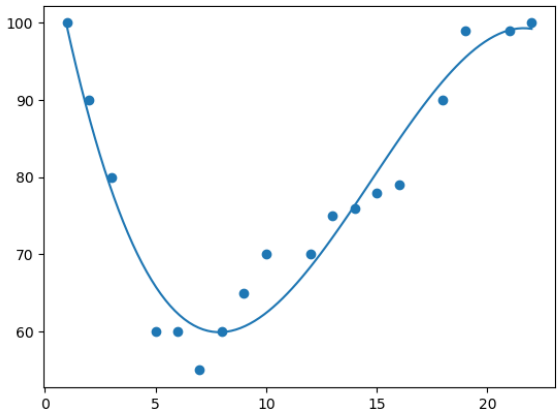
1. Computationally expensive requires high memory storage.
2. Sensitive to scaling of data.

Polynomial Regression

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Instead of assuming a linear relation between feature variables X and the target variable y, it uses a polynomial expression to describe the relationship. The polynomial regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \varepsilon_i \ (i = 1, 2, \ldots, n)$$

where ε_i is unobserved random error with mean zero conditioned on a variable x_i. β_0, β_1,... are unknown parameters/coefficients.



Advantages of Polynomial Regression

1. It offers a simple method to fit non-linear data.
2. It is easy to implement and is not computationally expensive
3. It can fit a varied range of curvatures.
4. It makes the pattern in the dataset more interpretable.

Disadvantages of Polynomial Regression

1. Using higher values for the degree of the polynomial supports overly flexible predictions and overfitting.
2. It has a high sensitivity for outliers.
3. It is difficult to predict what degree of the polynomial should be chosen for fitting a given dataset.