Semi-supervised

- Getting your hands on labeled data can be really expensive and a tedious task. Getting a dataset with some labels or no labels and then labeling some of them is an easier task.
- Now once you have your data basically you train the model on the labeled dataset and then predict the ones that are not labeled. This reduces the manual labelling time significantly.

Unsupervised

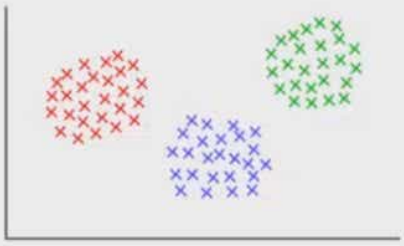
- If the target variable or the value that we are trying to predict is not available, then we perform an unsupervised task.

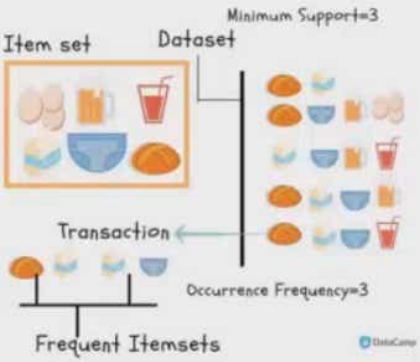
Clustering :

- Where you group similar points together.

Association:

- Where you try to find a pattern and try to recommend





Supervised

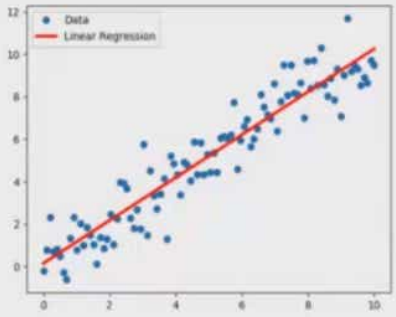
- When the data set that we are working with has labels in it which tells us which all column values represent which category or a continuous value, we perform a supervised task on it.

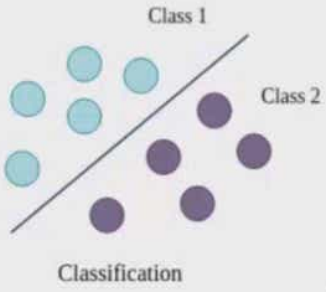
Regression:

- When the value that you want to predict is of continuous type.

Classification:

- When the value that you trying to predict is a category.





Classification

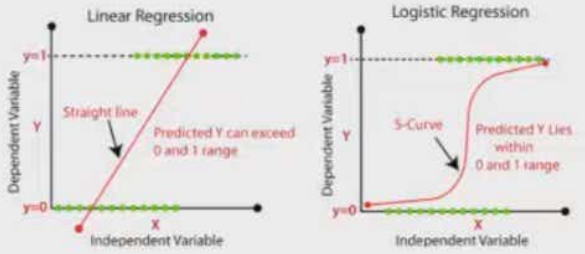
Assumptions to Linear Regression

- Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
- Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.
- Homoscedasticity: The residuals have constant variance at every level of x.
- Normality: The residuals of the model are normally distributed.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading.

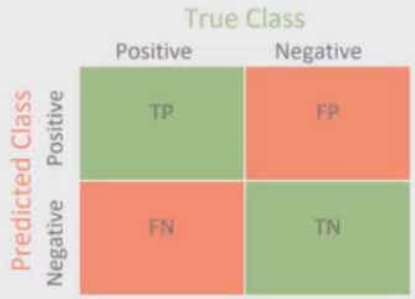
Logistic Regression

- Logistic regression is basically a supervised classification algorithm. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C or D (multinomial regression).
- It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.



Evaluate Logistic Regressions

- Confusion matrix is a good way to have a look at the correctly identified classes and misclassified classes.
- Using the values from there we can find the accuracy. The formula for that is total number of correctly classified records divided by the total number of records.

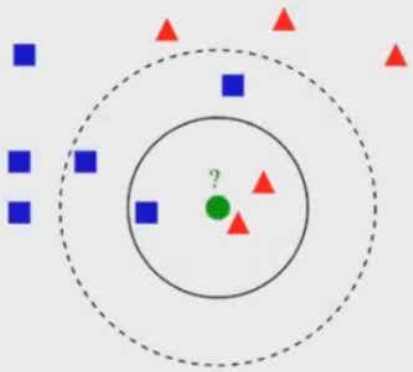


Logistic Regression

- Types of Logistic Regression:
  - Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
  - Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
  - Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".
- Assumptions for Logistic Regression:
  - The dependent variable must be categorical in nature.
  - The independent variable should not have multi-collinearity.

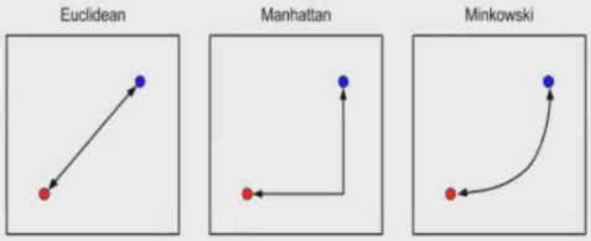
K nearest Neighbours

- KNN stands for K nearest Neighbors. Now k in nothing but a placeholder, which depicts the number of neighbors you want to take into consideration. Example k =3, I am going to take the 3 most nearest neighbors.
- How do we measure which element is close, we use some distance measure to decide that.



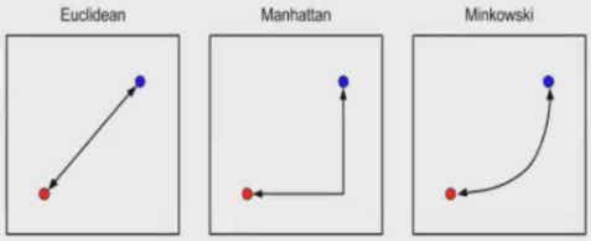
Distance Measures for KNN

- Euclidean : The distance is calculated through a straight line between two points.
- Manhattan : The distance is the summation of the perpendicular distance and horizontal distance.
- Minkowski : it's the distance between 2 points by using a curved line.



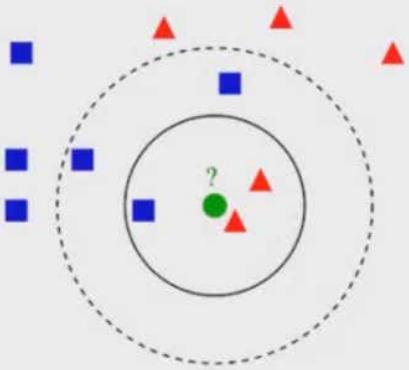
Distance Measures for KNN

- Euclidean : The distance is calculated through a straight line between two points.
- Manhattan : The distance is the summation of the perpendicular distance and horizontal distance.
- Minkowski : it's the distance between 2 points by using a curved line.



K nearest Neighbours

- KNN stands for K nearest Neighbors. Now k in nothing but a placeholder, which depicts the number of neighbors you want to take into consideration. Example k =3, I am going to take the 3 most nearest neighbors.
- How do we measure which element is close, we use some distance measure to decide that.



KNN

- Model Summary:
- Precision : What proportion of positive identifications was actually correct?
- Recall : What proportion of actual positives was identified correctly?
- F1 Score : It is calculated from the precision and recall of the test, The F1 score is the harmonic mean of the precision and recall.

$$\text{Precision} = \frac{TP}{TP + FP}$$

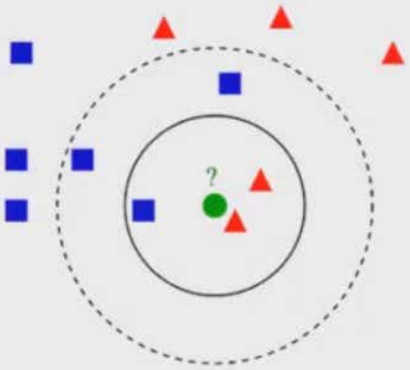
$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

2024-05-04 14:57:58

K nearest Neighbours

- KNN stands for K nearest Neighbors. Now k in nothing but a placeholder, which depicts the number of neighbors you want to take into consideration. Example k =3, I am going to take the 3 most nearest neighbors.
- How do we measure which element is close, we use some distance measure to decide that.



KNN

- K Fold and Stratified K Fold Cross Validation: In K fold we basically fold our data set k times and do the train test on that data set. This gives us K number of accuracies, now you can either look at the range in which the accuracies lie or just have an average accuracy.
- Stratified K Fold CV is just an extension of the same old K fold, the only difference is that the data over here is stratified, has an equal proportion of class distribution in both the training and the testing splits.

KNN

- Grid Search Cross Validation: This is a hyperparameter tuning method where you put in all the parameter values that you want to train and test your model with and on the basis of that you get a combination of all the values passed. You can select the best out of that.
- Random Search Cross Validation: This is similar to Grid Search but doesn't make a combination of all the values. It make a combination of that values that are most likely to give you better results. (Best for larger datasets and more number of parameters)

KNN

- Confusion Matrix:
  [TP: 1 , FP: 1
  FN: 8, TN: 90]
- Precision = 1/1+1 = 0.5
- Recall = 1/1+8 = 0.11
- F1-Score = 2 x(0.055/0.61) = 0.18
- Accuracy = 91 / 100 = 0.91

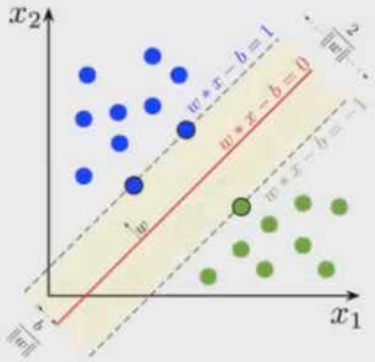
Applying Bayes Theorem to Understand Conditional Independence

Q. Let event A signify the occurrence of head and event B be the occurrence of a 1 in a roll of a dice.Find the Probability of B given that A occurs.

- P(A) = ½ and P(B) = ⅙
  P(A∩B) = P(A) x P(B) [As there is no intersection points]
- P(B|A) = P(A∩B)/P(A)
  P(B|A) = (1/12) / (1/2)
  P(B|A) = ⅙
- P(B|A) = P(B) =⅙ as they are conditionally independent.

SVM

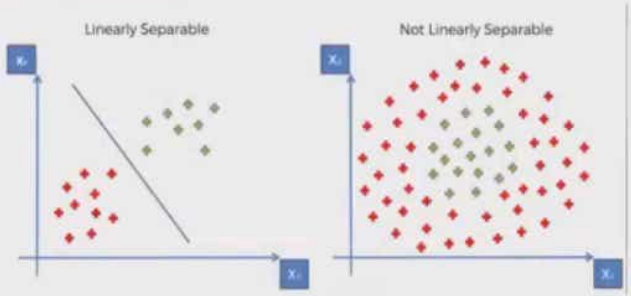
- SVM stands for Support Vector Machines, Where the main agenda of the models is to split the data into n categories, n being the number of categories present. The basic version of the SVM uses a linear hyperplane to distinguish between 2 or more classes.



- A hyperplane is nothing but a line (2D) or a plane (3D) that is used to determine which side of it will it fall into when a new point is provided to it.

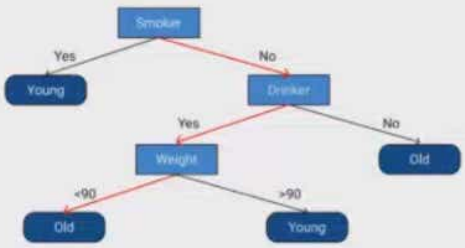
What are the steps to create a SVM model?

- The Process is simple:
- Step1 : Identify the support vectors.
- Step2 : See if the data is linearly separable.
- Step3 : If it is then make linear separator between them (exactly in the middle).
- Step4 : your model is ready.



CART

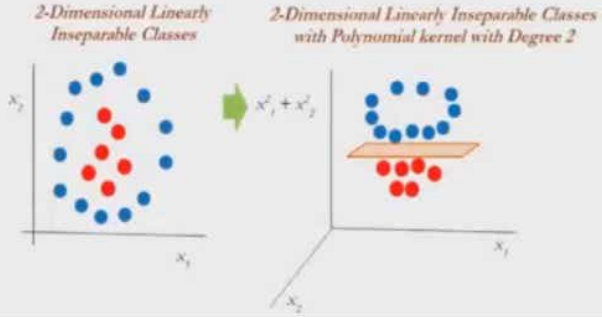
- CART stands for Classification and Regression Trees. The classification trees are also referred as decision trees. The name of the model is decision tree because it looks like an inverted tree. The decision is made on the basis of some condition and the splits are made accordingly.



- The tree goes on till the last node (referred as the leaf node), till it can't be split any further.

What are the steps for non linearly separable data points?

- The Process is simple:
- Step1: Use the kernel function to take the data points to a higher dimension.
- Step2 : Identify the support vectors.
- Step3 : Draw the hyperplane in N dimensions.
- Step4 : your model is ready.



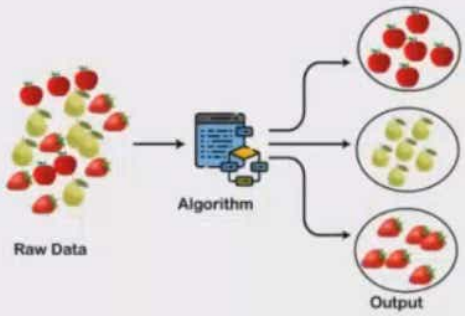
Disadvantages of Decision Trees:

Decision trees to the max depth tend to overfit. (Generally)

- Overfit: when the model takes in the training values and fit it perfectly and fails to perform well on the testing set, we say that our model has overfit. You can test by checking the training and testing accuracy. If the training accuracy is way higher than the testing accuracy, it is safe to say that the model did over fit.
- Underfit: when the data is unable to fit well, ie. the accuracy is lower on both the training and testing set we go on to say that the model underfit. This is generally the case when the model we are using is too simple for the data set.

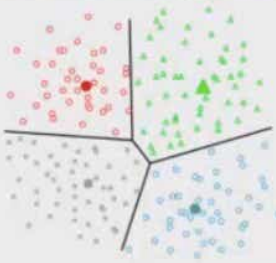
Clustering

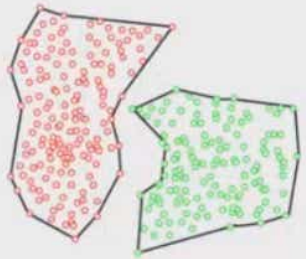
- A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.
- The clustering technique is commonly used for statistical data analysis.

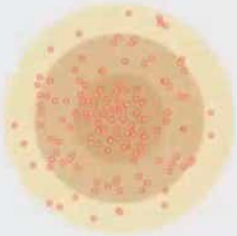


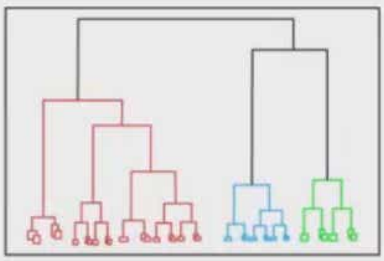
Types of Clustering Methods

- Partitioning Clustering
- Density-Based Clustering
- Distribution Model-Based Clustering
- Hierarchical Clustering
- Fuzzy Clustering



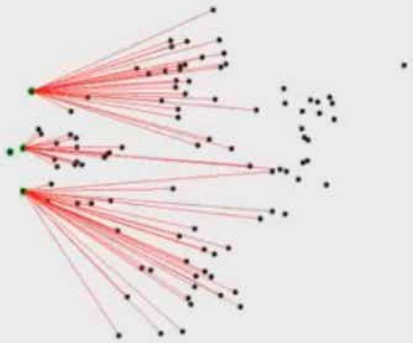






K means Clustering

- We choose an arbitrary k value which depicts the number of centroids, which in turn depicts the number of clusters.
- The distance from centroids to all the data points is calculated and the centroid with the least distance is assigned that that point.
- The cycle goes on till the the centroid values don't change significantly.

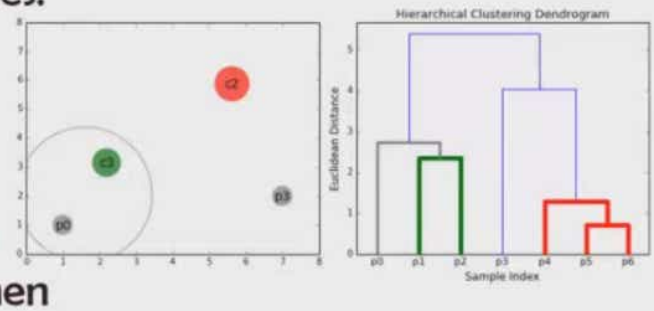


How do linkage affect the dendrogram ?

- There are 3 kind of linkage that can be used:
- Single Linkage: For two clusters C1 and C2, the single linkage returns the minimum distance between two points i and j such that i belongs to C1 and j belongs to C2.
- Complete Linkage: For two clusters C1 and C2, the complete linkage returns the maximum distance between two points i and j such that i belongs to C1 and j belongs to C2.
- Average Linkage: For two clusters C1 and C2, first for the distance between any data-point i in C1 and any data-point j in C2 and then the arithmetic mean of these distances are calculated.

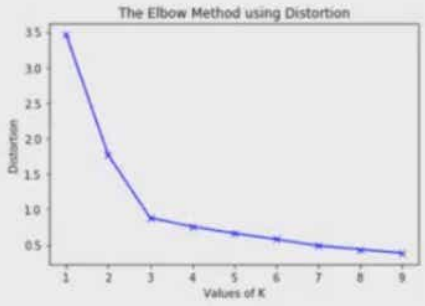
Hierarchical Clustering

- We use something called a dendrogram for this purpose. There are 2 approaches:
- Top down approach: Consider all the data points to be in one single cluster, and then go on splitting.
- Bottom up approach: Consider all the data points to be single clusters and then club them as you move up.
- Now in each case the distance is calculated and the ones with the least distance are clubbed together.



How to find the optimal value for K in K means?

- We use something called the elbow graph. The graph plots the value of k against the amount of distortion in the data points. The point after which the graph plateaus, is called the elbow point.
- In the example we can see the elbow form at k=3.

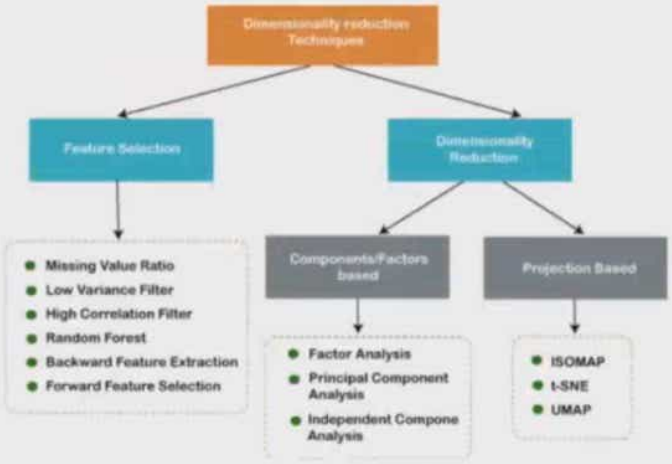


DBSCAN

- The DBSCAN algorithm uses two parameters:
- minPts: The minimum number of points (a threshold) clustered together for a region to be considered dense.
- eps (ε): A distance measure that will be used to locate the points in the neighborhood of any point.
- Reachability in terms of density establishes a point to be reachable from another if it lies within a particular distance (eps) from it.
- Connectivity, on the other hand, involves a transitivity based chaining-approach to determine whether points are located in a particular cluster.

Dimensionality Reduction

- It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information.
- It is commonly used in the fields that deal with high-dimensional data.



Dimensionality Reduction

Benefits

- By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
- Less Computation training time is required for reduced dimensions of features.
- Reduced dimensions of features of the dataset help in visualizing the data quickly.
- It removes the redundant features (if present) by taking care of multicollinearity.

Principal Component Analysis

- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- These new transformed features are called the Principal Components.
- PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.
- The PCA algorithm is based on some mathematical concepts such as:
  - Variance and Covariance
  - Eigenvalues and Eigen factors

Some common terms used in PCA algorithm

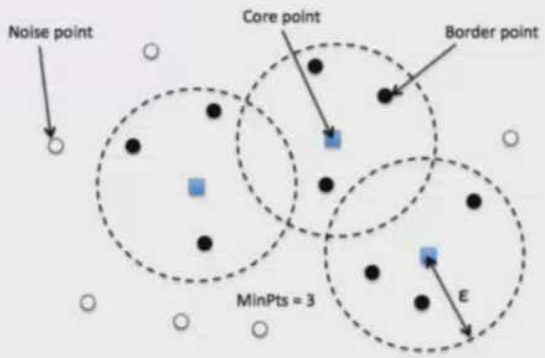
- Dimensionality
- Correlation
- Orthogonal
- Eigenvectors
- Covariance Matrix

Steps for PCA algorithm

- Getting the dataset
- Representing data into a structure
- Standardizing the data
- Calculating the Covariance of Z
- Calculating the Eigen Values and Eigen Vectors
- Sorting the Eigen Vectors
- Calculating the new features Or Principal Components
- Remove less or unimportant features from the new dataset.

DBSCAN

- Core — This is a point that has at least m points within distance n from itself.
- Border — This is a point that has at least one Core point at a distance n.
- Noise — This is a point that is neither a Core nor a Border. And it has less than m points within distance n from itself.



Linear Discriminant Analysis

- It is used as a pre-processing step in Machine Learning and applications of pattern classification.
- The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.
- Limitations of Logistic Regression
  - Two-class problems
  - Unstable with Well-Separated classes
  - Unstable with few examples

What Defines a Good Recommendation?

- The quality of a recommendation can be assessed through various tactics which measure coverage and accuracy.
- Accuracy is the fraction of correct recommendations out of total possible recommendations while coverage measures the fraction of objects in the search space the system is able to provide recommendations for.
- Recommender systems share several conceptual similarities with the classification and regression modelling problem.
- K Fold Cross Validation
- MAE (Mean Absolute Error)

Linear Discriminant Analysis

- It is used as a pre-processing step in Machine Learning and applications of pattern classification.
- The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid the curse of dimensionality and also reduce resources and dimensional costs.
- Limitations of Logistic Regression
  - Two-class problems
  - Unstable with Well-Separated classes
  - Unstable with few examples

What Defines a Good Recommendation?

- The quality of a recommendation can be assessed through various tactics which measure coverage and accuracy.
- Accuracy is the fraction of correct recommendations out of total possible recommendations while coverage measures the fraction of objects in the search space the system is able to provide recommendations for.
- Recommender systems share several conceptual similarities with the classification and regression modelling problem.

Recommendation Systems

- Recommendation engines are a subclass of machine learning which generally deal with ranking or rating products / users.
- They're used by various large name companies like Google, Instagram, Spotify, Amazon, Reddit, Netflix etc. often to increase engagement with users and the platform.
- Recommender systems are often seen as a "black box", the model created by these large companies are not very easily interpretable.

Reinforcement Learning Applications

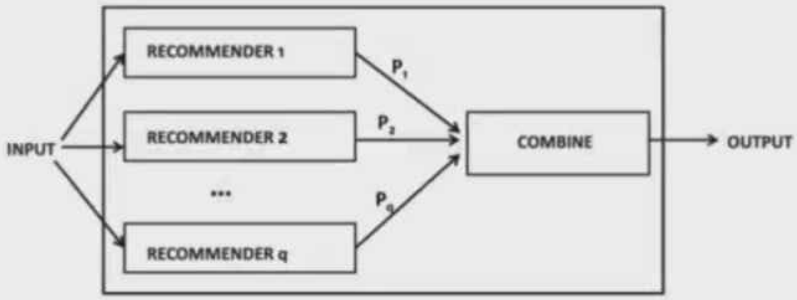
- Robotics
- Control
- Game Playing
- Chemistry
- Business
- Manufacturing
- Finance Sector

Collaborative Filtering Systems

- Collaborative filtering is the process of predicting the interests of a user by identifying preferences and information from many users.
- There are two common types of approaches in collaborative filtering, memory based and model based approach.
- Examples:
  - YouTube content recommendation to users
  - Coursera course recommendation

Hybrid Recommendation System

- Hybrid recommender systems are ones designed to use different available data sources to generate robust inferences.
- The parallel design provides the input to multiple recommendation systems, each of those recommendations are combined to generate one output.

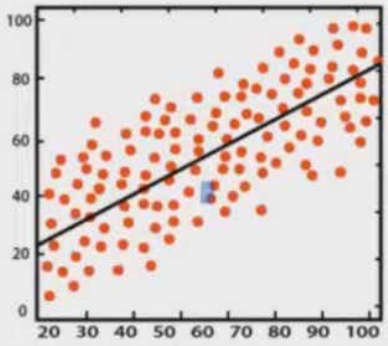


(a) Parallel design

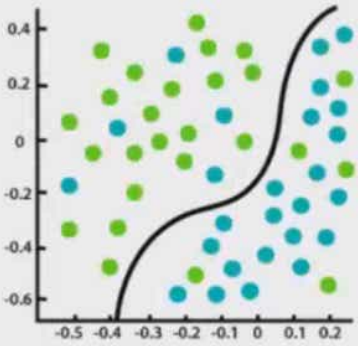
Reinforcement Learning Algorithms

- Q-Learning:
  - Q-learning is an Off policy RL algorithm, which is used for the temporal difference Learning. The temporal difference learning methods are the way of comparing temporally successive predictions.
- State Action Reward State action (SARSA):
  - SARSA stands for State Action Reward State action, which is an on-policy temporal difference learning method. The on-policy control method selects the action for each state while learning using a specific policy.
- Deep Q Neural Network (DQN):
  - As the name suggests, DQN is a Q-learning using Neural networks. For a big state space environment, it will be a challenging and complex task to define and update a Q-table.

Regression vs. Classification



Regression



Classification

Content Based Systems

- Content based systems generate recommendations based on the users preferences and profile.
- Unlike most collaborative filtering models which leverage ratings between target user and other users, content based models focus on the ratings provided by the target user themselves.
- The simplest forms of content based systems require the following sources of data
  - Item level data source
  - User level data source

Reinforcement

- Here the learning is a continuous process. Every time the model makes a wrong prediction it is instructed that it made a wrong prediction and that it needs to rectify the same. The cycle goes on till the point the machine can do its assigned task without fail.