What is Machine Learning?

- It is branch of computer science which gives computer the ability to learn without being explicitly programmed.

What is Supervised Learning?

- It is a process of making predictions by using labeled data.

What is Unsupervised Learning?

- It is a process of extracting patterns of structure using unlabeled data.

What is the difference between supervised and unsupervised machine learning?

- Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you'll need to first label the data you'll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

What is Regression?

- Regression is a process of establishing relationship between dependent and independent variable.
- It is a supervised machine learning technique which is used to predict continuous values.
- The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data

What is Classification?

- Classification is a process of categorizing a given set of data into classes.
- It can be performed on both structured or unstructured data.
- The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories

What is the difference between supervised and unsupervised machine learning?

- Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you'll need to first label the data you'll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

Explain Linear Regression?

- The ultimate goal of linear regression is to find a line that best fits the data.
- The goal of multiple linear regression and polynomial regression to find the place that best fits the data in n-dimension.
- It is used to predict a continuous dependent variable based on of independent variable.
- It uses Least Square criterion for estimation.
- In Linear Regression, linear relationship between variables are mandatory.

Explain Logistic Regression?

- Logistic regression is used to describe data and to explain the relationship between one dependent binary and one or more independent variables.
- Logistic equation is created in such a way that the output a probability value that can be mapped to classes and values can only be between 0 & 1.
- It is used to predict a categorical dependent variable based on od independent variable.
- It uses Maximum likely-hood estimation
- In Logistic Regression, linear relationship is not mandatory.

What is Regularization?

- Regularization is a techinique used to address over-fitting and feature selection.
- There is two types of regularization
  - L1 Regularization (Lasso Regression)
  - L2 Regularization (Ridge Regression)
- The key difference between these two is the penalty terms.
- In Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds " absolute value of magnitude" of coefficient as penality term to the loss function.
- In Ridge Regression, it adds "squared magnitude" of the coefficient as penalty term to the loss function.

Explain Support Vector Machine?

- Support Vector Machine is a supervised machine learning algorithm that can be used for both classification or regression challenges.
- However, it is mostly used in classification problems.
- We plot each data items as points in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular co-ordinate.
- Then we perform classification by finding the hyperplane that differentiates the two classes very well.
- Support Vectors are simpy the co-ordinates of individual observations, which helps us to create the boundary lines.
- There are two lines other than hyperplane which creates a margin known as Boundary Lines.
- Hyperplane is basically the separation line which divides two classes.
- Also, there is a Kernel function which is used to map a lower dimensional data into higher dimensional data.
- The SVM classifies is a frontier that best segregates the two classes(hyper-plane/line).

Explain KNN?

- K nestest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measures.
- It is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

How KNN works?

- K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.
- Its working with the help of following steps −
  - Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.
  - Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.
  - Step 3 − For each point in the test data do the following −
    - 3.1 − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
    - 3.2 − Now, based on the distance value, sort them in ascending order.
    - 3.3 − Next, it will choose the top K rows from the sorted array.
    - 3.4 − Now, it will assign a class to the test point based on most frequent class of these rows.
  - Step 4 − End

Explain Naive Bayes?

- The Naive Bayes Algorithm is based on the Bayes Theorem.
- Bayes theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event.
- It is "Naive" because it makes assumption that may or may not turns out to be correct.
- The goal is to find the class with the maximum proportional probability.
- It answers the following question. "What is the probability of A given B? and because of the naive assumption that variables are independent given the class"

Explain Decision Trees?

- The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria are different for classification and regression trees.
- Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes.
- The creation of sub-nodes increases the homogeneity of resultant sub-nodes.
- In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

What are the algorithm that uses Decision Tree?

- The algorithm selection is also based on the type of target variables.
  - ID3 → (extension of D3)
  - C4.5 → (successor of ID3)
  - CART → (Classification And Regression Tree)
  - CHAID → (Chi-square automatic interaction detection Performs multi-level splits when computing classification trees)
  - MARS → (multivariate adaptive regression splines)

Explain how ID3 works?

- The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking.
- A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment.
- Steps in ID3 algorithm:
  1. It begins with the original set S as the root node.
  2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain(IG) of this attribute.
  3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
  4. The set S is then split by the selected attribute to produce a subset of the data.
  5. The algorithm continues to recur on each subset, considering only attributes never selected before.

What is Entropy and Information gain in Decision tree algorithm ?

- The core algorithm for building decision tree is called ID3 . ID3 uses Enteropy and Information Gain to construct a decision tree.
  - Entropy:
    - A decision tree is built top-down from a root node and involve partitioning of data into homogenious subsets. ID3 uses enteropy to check the homogeneity of a sample.
    - If the sample is completely homogenious then entropy is zero and if the sample is an equally divided it has entropy of one.
  - Information Gain:
    - The Information Gain is based on the decrease in entropy after a dataset is split on an attribute.
    - Constructing a decision tree is all about finding attributes that returns the highest information gain.

What is entropy?

- Entropy is a measure of the randomness in the information being processed.
- The higher the entropy, the harder is is to draw any conclusion from that information.

What is Gini Index?

- It is a matrix for classification task in CART.
- It stores sum of squared probability of each class

How to avoid/counter Overfitting in Decision Trees?

- The common problem with Decision trees, especially having a table full of columns, they fit a lot. Sometimes it looks like the tree memorized the training data set.
- If there is no limit set on a decision tree, it will give you 100% accuracy on the training data set because in the worse case it will end up making 1 leaf for each observation.
- Thus this affects the accuracy when predicting samples that are not part of the training set.
- Here are two ways to remove overfitting:
  1. Pruning Decision Trees.
  2. Random Forest

What is Pruning in Decision?

- The splitting process results in fully grown trees until the stopping criteria are reached. But, the fully grown tree is likely to overfit the data, leading to poor accuracy on unseen data.
- In pruning, you trim off the branches of the tree, i.e., remove the decision nodes starting from the leaf node such that the overall accuracy is not disturbed.
- This is done by segregating the actual training set into two sets: training data set, D and validation data set, V.
- Prepare the decision tree using the segregated training data set, D. Then continue trimming the tree accordingly to optimize the accuracy of the validation data set, V.

title

In the above diagram, the 'Age' attribute in the left-hand side of the tree has been pruned as it has more importance on the right-hand side of the tree, hence removing overfitting.

How is a decision tree pruned?

- Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.
- Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

What is Random Forest?

Random Forest is an example of ensemble learning, in which we combine multiple machine learning algorithms to obtain better predictive performance.

Why the name "Random"?

- Two key concepts that give it the name random:
  1. A random sampling of training data set when building trees.
  2. Random subsets of features considered when splitting nodes.
- A technique known as bagging is used to create an ensemble of trees where multiple training sets are generated with replacement.
- In the bagging technique, a data set is divided into N samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging in parallel.

Which is better Linear or tree-based models?

- It depends on the kind of problem you are solving.
  1. If the relationship between dependent & independent variables is well approximated by a linear model, linear regression will outperform the tree-based model.
  2. If there is a high non-linearity & complex relationship between dependent & independent variables, a tree model will outperform a classical regression method.
  3. If you need to build a model that is easy to explain to people, a decision tree model will always do better than a linear model. Decision tree models are even simpler to interpret than linear regression!

Explain Random Forest?

- Random Forest is considered to be panacea of all data science problems.
- Random Forest is a versatile machine learning method capable of performing both regression & classification tasks.
- It creates multiple decision trees using bootstrapped datasets of the orignal data and randomly selecting a subset of variables at each step of the decision tree.
- The model then selects the mode of all of the predictions of each decision tree.
- By multiple trees it reduces the risk of error from an individual tree.
- It is type of ensemble learning method, where group of weak models combine to form a powerfull model(strong model).

How Random Forest works?

- Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- In order to classify a new object based on attribute each tree gives a classification and we say the tree "votes" for that class.
- The forest chooses the classification having the most votes cover all the trees in the forest and in case of regression, it takes the average of outputs by different trees.

OR

1. Pick at random K datapoints from the training set.
2. Building the Decision Tree associated to these k data points.
3. Choose the N number tree you want to build & repeat step 1 & 2.
4. For a new data point, make each one of your N trees predict, the category to which data points belongs and assign the new data point to the category that wins the majority, in case of regression, it takes the average of the output by different trees.

Explain Clustering?

- Clustering is an unsupervised techniques that involves the grouping or clustering of data points.
- The overall goal is to divide data into distinct groups such that ibservations within each group are similar.
- It's frequently used for customer segmentation, fraud detection and document classification.
- Common clustering techniques include k-means clustering, hierarchical clustering, mean shift clustering and density-based clustering.
- While each technique has different method in fining clusters, they all aim to achieve the same thing.

Explain Kmeans Clustering.

- It is one of the simplest and popular unsupervised machine learning algorithm.
- In other words, the k-means algorithm indetifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroid as small as possible.

How Kmeans works?

1. Choose the number of k clusters
2. Select at random k points, the centroids (not necessarily from our dataset).
3. Assign each data point to the closest centroid that form k clusters.
4. Compute and place the new centroid of each cluster
5. Reassign each data point to the new closest centroid, if any re assignment took place, go to step 4, otherwise got to finish.

How to handle Randomization in clustering?

- The k-means problem is to find cluster centers that minimise the intra-class variance i.e the squared distance from each data point being clustered to its cluster center (the center that is closest to it).
- Although finding an exact solution to the k-mean problem from arbitary input in Np-hard i.e(closest point becomes part & longest point forms another cluster).
  1. Choose first cluster center uniformly at random from data points
  2. For each observation x, compute the distance d(x) to nearest cluster center.
  3. Choose new cluster center from amongest data points, with probability of x being choosen proportional to d(X) square.
  4. Repeat step 3 until k centers been choosen.

Explain Hierarchical Clustering.

- Hierarchical cluster analysis is an algorithm that groups similar object into groups called cluster.
- The endpoint is a set of cluster, where each cluster is distint from each other cluser and the object within each cluster are broadly similar to each other.
- Also known as bottom-up approach or hierarchical agglomerative clustering.
- A structure that is more informative than the unstructured set of clusters return by flat clustering.
- This clustering algorithm does not require us to pre-specify the number of clusters.
- Bottom-up algorithms trat each data as a singleton cluster untill all cluster have been merged into single cluster that contains all data.

How Hierarchical Clustering works?

1. Make each data point a single point cluster means N cluster.
2. Take tge two closest data points and make them one cluster N-1.
3. Take the two closest cluster and make them one cluster
4. Repeat step3 untill there is only one cluster.

Explain Ensemble Learning?

- Ensemble Learning is the art of combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the model.
- It has many types but two more popular ensemble learning techniques.

Explain Bagging?

- Bagging stands for Bootstrap Aggragation.
- Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions.
- In generalised bagging, we can use different learners on different population.
- As we expect this helps us to reduce the variance error.

Explain Boosting?

- Boosting is an iterative technique which adjust the wieght of an observation based on the last classification.
- If an observation was classified incorrectly, it tries to increase the weight of this observation and viz.
- Boosting in general decreases the bias error and builds strong predictive models.
- However, they may overfit on the training data.

Explain AdaBoost (Adaptive Boosting)?

- AdaBoost (Adaptive Boosting) is a very popular boosting technique that aims at combining multiple weak classifiers to build one strong classifier.
- A single classifier may not be able to accurately predict the class of an object, but when we group multiple weak classifiers with each one progressively learning from the others' wrongly classified objects, we can build one such strong model.
- The classifier mentioned here could be any of your basic classifiers, from Decision Trees (often the default) to Logistic Regression, etc.
- Rather than being a model in itself, AdaBoost can be applied on top of any classifier to learn from its shortcomings and propose a more accurate model.
- It is usually called the "best out-of-the-box classifier" for this reason.

What is "weak" classifier?

- A weak classifier is one that performs better than random guessing, but still performs poorly at designating classes to objects.

How do ADA Boost classifier works?

- The basic concept behind Adaboost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.
- Any machine learning algorithm can be used as base classifier if it accepts weights on the training set.
- Adaboost should meet two conditions
  1. The classifier should be trained interactively on various weighed training examples.
  2. In each iteration, it tries to provide an excellent fit for these examples by minimizing training error.

title

What is Dimensionality Reduction?

- In machine learning classification problem, there are often too many factors on the basis of which the final classification is done.
- These factors are basically variables called features. The higher the number of features, the harder it gets to visualise the training set and then work on it.
- Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- It can be divided into feature selection and feature extraction.

What are the Component of Dimensionality Reduction?

- There are two components of dimensionality reduction:
  - Feature Selection -:
    - In this, we try to find a subset of the original set of variables or features to get a smaller subset which can be used to model the problem.
    - It usually involves three ways:
      - Filter
      - Wrapper
      - Embedded
  - Feature extractions:
    - This reduces the data in a high dimension space to lower dimension space, i.e a space with lesses no of dimension.

What are the methods of Dimensionality Reduction?

- The various methods used for dimensionality reduction include:
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Generalized Discriminant Analysis (GDA)
- Dimensionality reduction may be both linear or non-linear, depending upon the mothod used.

What are the Advantages of Dimensionality Reduction?

- It helps in data compression and hence reduced storage space.
- It reduces compuatation time
- It also helps remove redundant features, if any.

What are the disadvantages of dimensionality reduction?

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is something undersiable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep-in practice, some thumb rules are applied.

How is KNN different from k-means clustering?

- K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part).
- K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.
- The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn't—and is thus unsupervised learning.

Explain how a ROC curve works.

- The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds.
- It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

Define precision and recall.

- Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data.
- Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims.
- It can be easier to think of recall and precision in the context of a case where you've predicted that there were 10 apples and 5 oranges in a case of 10 apples.
- You'd have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

Explain the difference between L1 and L2 regularization.

- L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting.
- L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior.

What's the difference between Type I and Type II error?

- Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn't, while Type II error means that you claim nothing is happening when in fact something is.
- Forexample-: Type I error as telling a man he is pregnant, while Type II error means you tell a pregnant woman she isn't carrying a baby.

What's the F1 score? How would you use it?

- The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don't matter much.

How would you handle an imbalanced dataset?

- An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! SO few tactics to get over the hump are,
  - Collect more data to even the imbalances in the dataset.
  - Resample the dataset to correct for imbalances.
  - Try a different algorithm altogether on your dataset.

When should you use classification over regression?

- Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points.
- We would use classification over regression if we wanted our results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

How do you ensure you're not overfitting with a model?

- This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.
- There are three main methods to avoid overfitting:
  - Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.
  - Use cross-validation techniques such as k-folds cross-validation.
  - Use regularization techniques such as LASSO that penalize certain model parameters if they're likely to cause overfitting.

What's the "kernel trick" and how is it useful?

- The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space.
- This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates.
- Many algorithms can be expressed in terms of inner products. Using the kernel trick enables us effectively run algorithms in a high-dimensional space with lower-dimensional data.

What is the difference between R-Squared & Adjusted R-Squared?

- Basically everytime we add a independent variable to a model, the R-squared increases even if the independent variable is insignificant, it never declines.
- Whereas, in Adjusted R-squared increases only when independent variable is significant & affect dependent variable.

What is Gradient Decent?

- Gradient descent is an algorithm for finding the minimum of a differentiable function.
- We take steps proportional to the approximate gradient of the function at the current point.

How Gradient Decent works?

The steps in gradient descent

1. Start with any random w value.
2. Calculate gradient G of the function f(w) for that w value. A gradient is basically the slope which is either positive (moving upward) or negative (going downward).
3. Update the w value. $w = w - nG$

- Where n is the learning rate which depicts the magnitude of the step size. A
- A small learning rate will take more time to reach the minimum than a larger learning rate.
- But one should be careful that a too-large learning rate might skip the minimum.
- The above steps will repeat for i iterations.
- Both learning rate and no. of iterations are hyper-parameters.
- These hyperparameters must be set by the developer to fine-tune the entire process of finding the minimum.

What is bias, variance trade off ?

- Bias:
  - "Bias is error introduced in our model due to over simplification of machine learning algorithm."
  - It can lead to under fitting. When we train our model at that time model makes simplified assumptions to make the target function easier to understand.
  - Low bias machine learning algorithms — Decision Trees, k-NN and SVM
  - High bias machine learning algorithms — Linear Regression, Logistic Regression
- Variance:
  - "Variance is error introduced in our model due to complex machine learning algorithm, our model learns noise also from the training data set and performs bad on test dataset." It can lead high sensitivity and over fitting.
  - Normally, as we increase the complexity of our model, we will see a reduction in error due to lower bias in the model.
  - However, this only happens till a particular point. As we continue to make our model more complex, we end up over-fitting our model and hence our model will start suffering from high variance.

title

- Bias, Variance trade off:
  - The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.
    1. The k-nearest neighbours algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.
    2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.
  - There is no escaping the relationship between bias and variance in machine learning.
  - Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

- by Mrityunjay Kumar Pandey