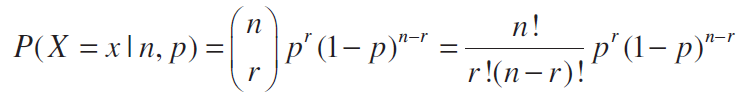
1. Please describe the central limit theorem and provide an example.

Central Limit Theorem (CLT) is one of the interesting concepts in Statistics which simplifies our analysis when we have multiple independent & identical distribution (IID). When it is performed many times then the distribution of the sum of the average will be a close approximation of a normal distribution.

For an example consider the principles of Bernoulli process of a binomial random variable. Where for a given *n* and *p*, when we vary the *r* from *0* to *n* it would approach a normal distribution.



The interesting feature of CLT is that there is no constrain that *p* value should be 0.5 or it should be a symmetrical distribution. This is surprising that for any type of distribution, we can get approximate probabilities for sums of any random variable as long as they are independent & identical distribution (IID).

Higher the *n* the approximation is closer to normal distribution. For a smaller sample size (n<30) performing error correction would result in better approximation.

For example, consider ticket sale volume for a football stadiums and check the probability of under booking or over booking and it impact on the revenue. It’s tedious to do millions of binomial calculation for it but instead a simple normal distribution would give very close value for the same.

1. Describe a classification algorithm that you have previously put into production and why it was chosen.

I have used various classification algorithms like “K-Nearest Neighbors”, “Decision Tree”, and “Logistic Regression”. My favorite one would be the Logistic Regression as I have applied it in various different scenarios.

For example – while studying customer behavior, age is one of the important independent variable and it’s not valuable to know what their exact age but we just need to know if they are minor or not. So we convert the age into 2 group isMinor=True and isMinor = False. By this we were able to convert a multi value variable into binary or logical value.

For another example let’s consider the attached New York City Airbnb data points. We have 5 different Neighborhood group – Bronx, Brooklyn, Manhattan, Queens, Staten Island, which a statistical model cannot take into account as it’s not numerical value. We cannot give sequential numerical values to it, as their relationship to price is not linearly related. So the best option would be to create 5 different logistic regression independent variable with values 0 or 1.

1. Describe the difference between bagging and boosting methods, and when to use one or the other.

The 3 main challenges with model learning are - variance, bias and noise. We always have a large bias with simple trees and large variance with complex trees. To overcome these challenges we use **Bagging** and **Boosting** ensemble methods. The main idea behind the ensemble methods is to group weak learners together to form a strong learner.

Bagging method is used to decrease the model’s **variance** by performing sampling technique called **bootstrapping**. In this we create multiple subset of observation from the original dataset with replacement. Then a weak model is created on each of the subset. Final prediction are determined by combining and running all models in **parallel** independently.

Example: The calculation of portfolio risk. The overall risk or variance can be decreased by creating subgroups based on decision tree elements - Stock, Bonds, Commodities, FX etc.

Boosting method is used to decrease the model’s **bias** by performing a sequential learning process to provide better results. The early learners fitting simple model to the data and then net errors are analyzed. Then recursively we solve for net error from the prior tree.

Example: Fourier Transform - Decomposition of periodic function into multiple sin/cos function with different frequencies. The early learner would be function with a weight and lower frequency, then each subsequent recursive learner based on net error, would have its own weight and higher frequency.

1. Describe 2 regularization techniques for a random forest model

In many situation our model performs exceptionally well with training data but it is not able to predict well in testing data. To overcome this challenge we need to avoid overfitting of training data. Regularization techniques help us to overcome these challenges. There are many Regularization techniques like L1 & L2 regularization, Dropout, Data augmentation, early stopping, etc.

L1&L2 Regularization – It is the most common type of Regularization. We add an extra Regularization term to overall cost function, due to this additional regularization term, the smaller weight matrices leads to simple models. This will also reduce overfitting of the training data.

Data Augmentation – It is the simplest way to reduce overfitting by increasing the training data size. Training data images can be increased by rotating, flipping, scaling, shifting or by few of these combinations. This will provide a huge leap in improving the training data.

Early Stopping – In this we segregate data in 2 parts for training and validation. This helps in keeping the model on track by check and balance. We immediately stop the training of the model when we notice any poor validation results.