STAT 652 Project 1

Dhruv Patel, 301471961

Q11. Explain what is cross validation and what is bootstrap?

Answer:
Cross Validation is a method to evaluate models. The basic idea is to divide the dataset into two parts, and train the model on a large dataset and test for the model performance and accuracy on the rest of the dataset. It helps to find how well a model generalizes on a new dataset. Different methods are Holdout Method, K-Fold CV and LOOCV.

Bootstrap method is a resampling technique used to estimate statistics on a population by sampling a dataset with replacement. In other words, some subset from the population is selected upon which certain statistical estimates are calculated. This is done repetitively for the number of bootstrap samples selected. The estimated parameter by bootstrap sampling is comparable to the actual population parameter and since we need few samples for bootstrapping the computation requirement is less.

Q30. What are the hyper parameters of bagging? Discuss how to choose the hyper-parameter.

Answer:
The most important hyper parameters of bagging is (n_estimators) i.e the number of trees, (max_features) i.e the number of features to draw from X to train each base estimator. Ideally, we can increase n_estimators until no further improvement is seen in the model.

Manual hyper-parameter tuning: In this method, different combinations of hyper-parameters are set (and experimented with) manually. This is a tedious process and cannot be practical in cases where there are many hyper-parameters to try.

Automated hyper-parameter tuning: In this method, optimal hyper-parameters are found using an algorithm that automates and optimizes the process. Different methods are

- **Grid Search** - We create every possible combination of hyper-parameters possible and record the performance.
- **Random Search -** Randomly select a few hyper-parameters instead of computing for all possible combinations.
- **Bayesian Optimization**: It builds a probabilistic model of the function mapping from hyper-parameter values to the objective evaluated on a validation set. By iteratively evaluating a promising hyper-parameter configuration based on the current model, and then updating it.

Q3. What are the assumptions of a linear model? What will happen if we have correlated variables in a linear model?

Answer:
Assumptions of a linear model are:

1. The variance of residuals i.e, (predicted values of Y - actual values of Y) is same for any value of X.
2. Observations are independent of each other.
3. Relationship between X and the mean of Y is linear.
4. For any fixed value of X, Y the model is normally distributed.

When the variables are correlated in linear model,

1. The estimated regression coefficient of any one variable depends on which other predictor variables are included in the model.
2. The precision of the estimated regression coefficients decreases as more predictor variables are added to the model.
3. The marginal contribution of any one predictor variable in reducing the error sum of squares varies depending on which other variables are already in the model.
4. Hypothesis tests may yield different conclusions depending on which predictor variables are in the model. (This effect is a direct consequence of the three previous effects.)
5. High multicollinearity among predictor variables does not prevent good, precise predictions of the response within the scope of the model.