**Statistical Learning and Analysis Instructor: Prof. Varun Rai Assignment 3**

**Question 1:**

Let’s talk about k-fold validation.

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold cross validation relative to:

i) The validation set approach?

ii) LOOCV?

**Cross Validation** is a very useful technique for assessing the performance of machine learning models. The cross-validation technique can be used to **compare the performance of different machine learning models on the same data set.** K-fold cross-validation is a type of cross-validation technique used as a sampling method.

K-fold cross validation is performed as per the following steps:

1. Partition the original training data set into k equal subsets. Each subset is called a fold. Let the folds be named as f1, f2, …, fk.
2. For i = 1 to i = k
   1. Keep the fold fi as Validation set and keep all the remaining *k-1* folds in the Cross-validation training set.
   2. Train your machine learning model using the cross-validation training set and calculate the accuracy of your model by validating the predicted results against the validation set.
3. Estimate the accuracy of your machine learning model by averaging the accuracies derived in all the*k* cases of cross validation.

In the k-fold cross validation method, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.

b. k-fold cross validation vs validation set approach –

In case of validation set approach estimate of test error rate can be highly variable, based on which observations we included in our training set and which are included in validation set. In contrast for k-fold cross validation estimate of test error rate most probably less variable as accuracy of machine learning model is calculated by averaging the accuracies derived in all the k cases of k-fold cross validation.

In validation set approach observations those are included in training set are basically used to fit the model rather than observations present in validation set. So as model is fit on training data set it may perform worse if there are fewer observations in training data set. As a result, validation error set might overestimate the actual test error. But in k-fold cross validation all the k-folds are used as training as well as validation set turn by turn it is very less chance to overestimate or underestimate the test error rate for the model fit on entire dataset.

c. k-fold cross validation vs LOOCV

for K-fold main advantage is its computational cost. So, if LOOCV requires fitting the statistical learning method ‘n’ times in contrast performing k-fold CV requires fitting procedure only ‘k’ times. Ex. For 10-fold CV requires fitting the learning procedure only 10 times, which is much more feasible. So, if ‘n’ is large, LOOCV may pose computational problems.

As compare to k-fold validation in training/validation set splits, performing LOOCV multiple times results in less randomness in respective splits.

K-fold validation method has an upper hand in bias-variance trade off.

**A comparison of Classification Methods – Summary**

Logistic regression and Linear Discriminant Analysis produce linear decisions, but the difference between these two approaches in coefficients for logistic regression are estimated using maximum likelihood function whereas for LDA coefficients are computed using mean and variance from a normal distribution. Logistic regression will give better results if you want to categorize the data into two categories. So, for more than two categories i.e. multiple logistic regression model can’t perform better than LDA approach for classification. LDA assumes that observations come from Gaussian distribution. The variance from the normal distribution is calculated and is used to estimate coefficients.

KNN is a non-parametric approach, meaning the decision boundary is not assumed. To make predictions for an observation X = x, it assumes K training observations closest to the observation x are identified. Then X is assigned to the class to which these observations belong. Since KNN is a completely non-parametric approach, no assumptions are made about the shape of the decision boundary. KNN computes better predictions than LDA and Logistic Regression, when the decision boundary is non-linear. But KNN doesn’t help in identifying statistically significant predictors for the response.

Quadratic Discriminant Analysis is a mid-way approach for non-parametric KNN method and the linear LDA and logistic regression. Since QDA works on quadratic decision boundary, it is useful to cover wider range of problems than the linear models. But QDA is not as flexible as other models like KNN. So, with limited number of training observations QDA will perform better as it doesn’t make any assumptions about form of a decision boundary. KNN functions on finding the nearest neighbor for every observation, and a smaller training set might give a different error estimate.