Kunal Mehta
50352960
kunalkam

Quiz 4

1. An unfortunate split is one of the disadvantages of using the hold-out method of dividing data into training and testing subsets. This method entails setting aside a small random set of observations from the dataset as a testing dataset and using the rest as a training data set, usually done in 60:40 to 90:10 ratios. Now this randomized method of selecting observations for the subsets can sometimes result in the testing dataset having a skewed distribution of the data that may end up not being a good enough mix to thoroughly test the model, leading to misleading reports regarding its efficiency.

   For example in the iris dataset, there are a total of 150 observations, with 50 of each class. If an 80-20 ratio is used to subset the data, it can so happen that the testing dataset has an overabundance of one particular class with very few, and in extreme cases, no observations corresponding to the other classes. Such a testing dataset is not a good base to test the model on and will give accuracy reports that do not really depict the real efficiency of the model.

   K-fold cross validation is a good way to avoid this since it divides the dataset into k folds of approximately equal size, and then with every iteration one of the folds is treated as the testing set and the metrics are calculated. This process happens k times, obviously and therefore all of the data is essentially used for both training and testing purposes. The error metrics are also averaged out over k to give a clearer picture about all the data. The only disadvantage to this approach is the computational complexity as compared to hold-out method.

2. Leave one out cross validation is similar to k-fold cross validation in the sense that it also makes folds, except that in this case, for each iteration, there is just one observation left out as a testing. So essentially, the training dataset for each iteration is the entire dataset minus one observation. This process naturally is performed n times, n being the number of observations. The error metrics are also averaged out over n. Leave one out cross validation can be thought of as a special case of k-fold cross validation where $k = n$.

3. As explained in the slides, essentially AIC and BIC differ mathematically, not in principle. They are both penalty-based metrics, and their difference lies in how they penalize the complexity of a model. AIC and BIC both can be used to choose the best model theoretically since they address the same problem but provide a different answer due to their difference of calculation and assumptions. For instance, if AIC suggests a 7 variable model and BIC suggests a 10 variable model, it would be prudent to check models of sizes in the range 7-10 and then select one from them as the best one. BIC as a measure depends on the number of observations which is not a factor that affects AIC.

   a) For this case, number of variables are less than 10. As mentioned in the problem, the number of observations is also relatively low, 50 to be precise. In such a case, if I had to choose one as a more appropriate method, I would choose to trust AIC as a metric. The reason for this is that BIC tends to penalize higher complexity more than AIC, but when the variable numbers are already so few in number, there is a higher chance that BIC would choose too simplistic a model as compared to AIC.
   b) For this case, the number of variables is quite high when compared to the number of observations. In such a case, I would definitely choose to trust BIC since it heavily penalizes high complexity in the model, whereas AIC is more liberal. With $p = 40$, a simple model is really desirable, particularly with the n we have and BIC would provide just that, a more simplistic solution.