Concept Quiz
Statistical Data Mining I
Due Oct 31 at 11:59PM

1) Explain how k-fold cross validation is implemented.

Partition the training data sets in to equal subsets called folds. Loop through all the folds and for each use one fold as the validation set, and the others as the cross validation training set; then do the designated ML training with the cross validation set and calculating the result by validating with the validation set. The accuracy is the average of all the accuracy.

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

a) The hold-out method (i.e., single division of test and training)

The hold out method is dependent on only one set of train/test datas whereas the cross-fold validation can more accurately come up with an accuracy of the performance on unseen datas.

The hold out method is better when dealing with large datasets or building an initial model since it will be less time and power consuming.

The Advantage is that k-fold has lower variability in CV test error and possibly higher accuracies.

Disadvantages is that k-fold is harder to implement and requires more computational power.

b) Leave one out cross validation

The leave one out method is the k-fold method with k = n.

K-Fold :

Advantage : Variance is reduced as K goes up, every data point is tested exactly once; so K-fold will have less variance. If n is large, k-fold is more cost efficient and less cost consuming.

Disadvantage : LOOCV has less bias.

c) The bootstrap

The boostrap method is used to calculate variance of a parameter, it is used to simulate the correct distribution. The bootstrap “clones” the original data set to create multiple data sets.

Whereas the k-fold cuts up data into data sets to validate the performance of a model to make sure we are not over-fitting.

The bootstrap is not a performance validation method.