PM 591 – Machine Learning for the Health Sciences.

Assignment 3

Due 2/28/2022

Exercise 1

a. Build a KNN classifier to predict stroke using the ischemic stroke data and tune the complexity parameter $K = 1, \dots, 50$ using a single-split validation set. As features use “sex”, “age”, “CoronaryArteryDisease”, “MaxStenosisByDiameter” and “MATXVolProp”). Plot the classification error as a function of K. Which value of K do you choose? Explain.

b. Repeat a) 9 additional times with different random training/validation splits (use a loop). Plot the 10 curves, analogs to the one obtained in a. in the same graph. Do you choose the same value of K for each of the 10 splits? What does this say about the stability/variability of using a single training/validation split to perform model selection?

c. Now tune the complexity parameter $K = 1, \dots, 50$ using now 5-fold cross-validation instead of a single training/validation split. Which value of k do you choose? Explain.

d. Repeat c) 9 additional times with different cross-validation splits (use a loop). Plot the 10 curves, analogs to the one obtained in c. in the same graph. Do you choose the same value of K for each of the 10 splits? What does this say about the stability/variability of using cross/validation to perform model selection compared to a single split?

Exercise 2.

Using the ischemic stroke data with the same features than in exercise 1, train an evaluate the performance of an LDA, QDA, and logistic regression classifiers using the mlr3 package. Plot the ROC curve and report the AUC for each of the classifiers. Compare the performance of the three classifiers. Which one would you choose for predicting stroke?