Stat 427/627 Statistical Machine Learning Quiz 5 Jun Lu. American University

Quiz 5. Cross-validation

Name: Lisa Liubovich Attempt (circle one): [✓ BEFORE] AFTER

1. Consider a data set with 500 observations. We want to use 10-folder cross-validation to choose among 4 proposed models with the same response variable, but different orders of the predictors.

   (a) How many CV-values will we calculate?

Dataset is divided into 10 equal parts and the validation process is repeated 10 times. Each model will be validated 10 times (once for each fold). Therefore 40 CV values will be calculated in total.

   (b) To compute each CV-value, how many times do we fit the model?

10 folds and 4 models → the model fitting process will be carried out 40 times in total

   (c) Following (b), each time a model is fit, how many "validation" observations will be predicted and used to compute the test MSE?

Training set: (k-1)n/k = 9*500/10 = 450

Validation set: n/k = 500/10 = 50

Each time a model is fit, it is done on 450 observations (training set) and the remaining 50 observations (validation set) are used to compute the test MSE. The process is repeated for each of the 10 folds.

2. The head coach of the Washington Commanders football team Ron Rivera decides that he can predict the game result at half-time, based on the passing yards by quarterback Carson Wentz by that time. Suppose that the following data are collected during the recent 4 games and the coach uses the KNN algorithm to predict the next game.

Games	1	2	3	4
Passing yards by half-time (X)	100	200	150	140
Result of the game	Loss	Win	Win	Loss

Use LOOCV (Leave-One-Out Cross-Validation) to estimate the prediction error rate when K = __ (Stat 427 only) Use KNN with K = 1.

(Stat 627 only) Use KNN with K= 3.

(You can use the following table as a worksheet to assist your calculation.)

Left-out Game. (deleted)	X	K = 3 Neighbors	Neighboring responses	Predicted response	Observed response	Correct or error
1	100	2,3,4	Win, Win, Loss	Win	Loss	Error

2	200	1,3,4	Loss, Win, Loss	Loss	Win	Error
3	150	1,2,4	Loss, Win, Loss	Loss	Win	Error
4	140	1,2,3	Loss, Win, Win	Win	Loss	Error

Caution: This is just an example. The sample size (n=4) is too small to use KNN in the real-world.

Prediction error rate is 1.0