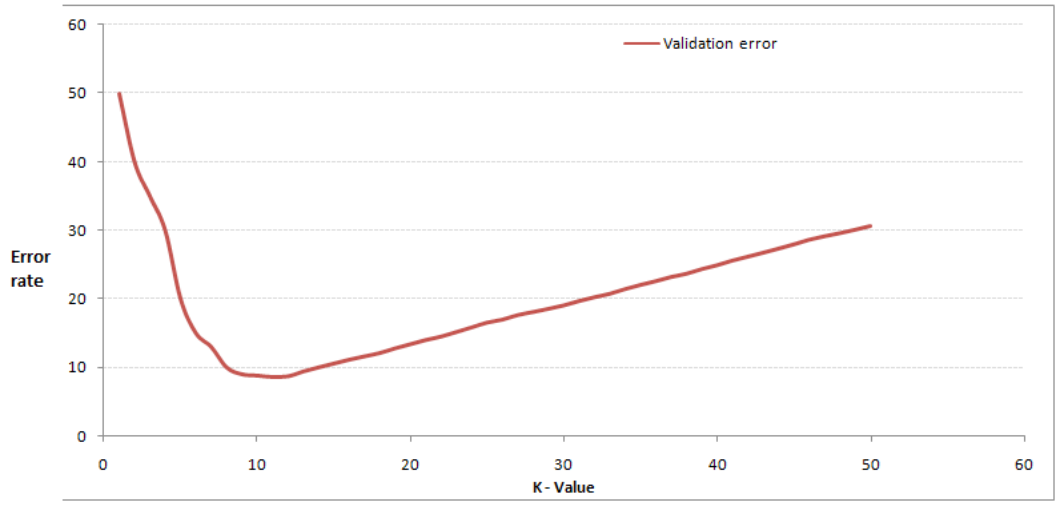
CEE 690: Health and Environmental Data Science, Spring 2019

Homework 2: 100 Total Points

Due: February 18th by 11:59 PM

Submissions required electronically through Sakai

Problem 1: Thinking about Validation (15)



Suppose we wanted to choose the number of neighbors to use in a k-Nearest Neighbor classifier. Someone had previously ran an analysis and reported the number of classification errors made on the validation set. What level of *K* would you choose? Also, the developer had chosen to use the error rate instead of AUC. What assumption do you need to check to make sure that this value is helpful? (I.E. when will the error rate be robust and give similar results as AUC?)

Consider a classification problem with a large number of predictors, as may arise, for example, in genomic or proteomic applications. A typical strategy for analysis might be as follows:

1. Screen the predictors: find a subset of "good" predictors that show fairly strong (univariate) correlation with the class labels
2. Split the data into a training and validation set
3. Using just this subset of predictors, build a multivariate classifier.
4. Use validation error to estimate the unknown tuning parameters and to estimate the prediction error of the final model.

In this approach, the validation performance estimate would not generalize to the population. What is wrong about this approach? How would you correct it?

Problem 2: **How does scaling affect penalized regression? (20)**

When we rescale the data, it affects many algorithms, potentially in many different ways. In traditional linear regression, if we rescale the data, it does act affect our ability to fit the data in any way. However, this is not true when we consider *penalized* regression methods (i.e. Ridge Regression, Lasso). Consider the ridge regression form with a single feature and a single data point:

$$\underset{\mathrm{b}}{\operatorname{argmin}}(y - xb)^2 + \lambda b^2$$

Now suppose we scale both our feature and our output, so that $y^* = ay$ and $x^* = ax$, meaning that the true relationship stays the same. What value should λ^* be to give the same answer as above?

$$\underset{\mathrm{b}}{\operatorname{argmin}}(y^* - x^* b)^2 + \lambda^* b^2$$

Qualitatively, describe scaling affects the ridge regression penalty.

Problem 3: Model selection in the Diabetes Data Set (20)

In Lecture 8, we went through code to apply Ridge Regression and Lasso to a dataset on diabetes, but not to perform model selection. Using this same dataset, with both Ridge and Lasso:

(a) Plotting the training and validation mean square error as a function of the strength of the penalty
(b) Determine a good setting of α.
Note: We're not looking for everyone to get the same answer here, because this is dependent on what settings of α you tried and the random validation split. What is more important is that you *explain* the rational for your choice.
(c) Do you prefer Ridge or Lasso here?

Next, consider using the Elastic Net here, which requires tuning over 2 settings simultaneously ("alpha" and "l1_ratio").

(d) Visualize the validation mean square error as a function of these settings
(e) Determine a good setting for the parameters and *defend* your choice.

Problem 4: Model Selection with Penalized Regression (20)

We have a synthetic dataset to show model selection in regression, and the data is stored in three files:
"hw2_problem_4_X.csv"
"hw2_problem_4_y.csv"
"hw2_problem_4_X_blind.csv"
The first two CSVs include the features *X* and a continuous outcome/target on *y*. Using a validation dataset, learn/select a model to predict the outcome *y* to minimize the predictive mean squared error.

To do this, create a procedure that:

1. Applies a scaling to the data.
2. Applies the Lasso or Ridge regression technique
3. Using validation to choose the scaling technique and model
4. Provides a predicted value on the "blind" data

Upload a saved excel file of your predictions on the blind data with your submission files. We will release the blind data's outcomes for the next assignment, and you will get to evaluate how well you actually did predicting future data.

Problem 5: Model Selection with K-Nearest Neighbors (20)

I have constructed a data problem, and the data is stored in three files:
"hw2_problem_5_X.csv"
"hw2_problem_5_y.csv"
"hw2_problem_5_X_blind.csv"
The first two CSVs include the features *X* and the binary outcome/target on *y*. Using a validation dataset, learn/select a model (i.e. how many features to keep, how many neighbors) to predict the outcome *y*.

To do this, create a procedure that:

1. Applies K-best feature selection using univariate statistics
2. An K-nearest neighbors on the training data
3. Provides a decision statistic on the "blind" data

Upload a saved excel file of your predictions on the blind data with your submission files. We will release the blind data's outcomes for the next assignment, and you will get to evaluate how well you actually did predicting future data.

Pseudo-Problem 6: Administrative: (5)

(a) (5) How many hours did this assignment take you? (There is **NO** correct answer here, this is just an information gathering exercise)
(b) (5) Verify that you adhered to the Duke Community Standard in this assignment (https://studentaffairs.duke.edu/conduct/about-us/duke-community-standard). (I.E. write "I adhered to the Duke Community Standard in the completion of this assignment"