Assignment: Validation and Regularization

Problem 1 (20 points)

One of the applications of validation set is to select appropriate model parameters. In k-fold method you'll partition data into k (10 for this assignment) randomly chosen subsets of equal size. One subset is used to validate the model trained using the remaining subsets. This process is repeated k times such that each subset is used exactly once for validation.

In this problem will use validation sets to pick parameters for k-nearest-neighbor (KNN) model. The data sets for this problem are *healthcareTrain.csv* and *healthcareTest.csv*.

Note that you can use the built-in function for KNN, but you need to write your own code for cross-validation for this problem.

1. (10 points) Create a KNN classifier model to predict the pdc-80-flag using the following continuous features "total-los","num-op","num-er","num-ndc","pre-total-cost", and "pre-CCI". Use 10-fold cross validation to determine which value of K produces the most accurate result from the range k = 31 to 101 with a step size of 2.

2. (5 points) Plot the accuracy rate from your 10-fold cross validation vs. K.

3. (5 points) Use the best value of K to predict the pdc-80-flag for the test set. How does your validation error compare to test error?

Problem 2 (20 points)

Consider a learning scenario where the goal is to learn the target function $f(x) = sin(\pi x)$ for $-1 \leq x \leq +1$ from two points in the training sets. The two training points in R^2 have a uniform distribution between -1 and +1. You will create two models in linear hypothesis set $y = mx + b$: 1) unregularized, 2) weight-decay regularized (use L_2 regularization with $\lambda = 0.1$).

1. (5 points) Generate 10,000 hypotheses for each version. Report the average hypothesis $\bar{g}(x)$ in each case.

2. (5 points) Find and report $bias^2$ for each model.

3. (5 points) Find and report variance for each model

4. (5 Points) For each case, plot $\bar{g}(x) \pm \sqrt{var(x)}$ along with $\bar{g}(x)$ and target function $f(x) = sin(\pi x)$. Which model will you choose? Why? Round your answers to 3 decimal places.