Statistical Learning - Milan, Fall2019

Homework problem 3

Flexibility of modeling approaches and its implications

In this problem we will investigate both theoretically and empirically the meaning and results of flexibility in predictive modeling.

1. For each of these methods, say which version is more flexible and which less flexible, and give a short explanation:
   (a) Least squares regression with few explanatory variables vs least squares regression with many variables and/or interactions
   (b) k-Nearest Neighbor (k-NN) with k=1 vs k-NN with large k (say, k=100)

2. Draw a schematic graph (or describe it in words) with the x-axis being the flexibility of the model, and the y-axis being squared error, and draw a line for each one of the relevant quantities, showing how it behaves when the flexibility grows:
   (a) Prediction squared bias
   (b) Prediction variance
   (c) Irreducible error
   (d) Test/Prediction squared error
   (e) Training squared error

   The important aspects of the schematic is to show for each one if it is increasing, decreasing, constant or non-monotone, and preserve the order between the graphs (for example, we know that the irreducible error is always smaller than the prediction error).

3. For the k-NN case, try it out empirically: The file kNN-HW1.r pointed from the class home page does the following:
   (a) Draws a single test set of size ntr=100, given by the matrix X.te, and vectors f.te and y.te.
   (b) Draws 200 training sets of size nte=100 (X.tr, f.tr and y.tr), and for each one applies k-NN regression to the test set with $k \in \{1, 5, 10, 20, 50, 100\}$.
   (c) Collects all the predictions in the array test.preds of test predictions $\hat{f}(x)$, whose first coordinate is the observation number, second is k (the neighbors number) and third is the set index (1-200). For example, the vector of test set predictions for the 10th training set and 5 neighbors is test.preds[,"5",10]
   (d) Also given is the array test.errs (which gives $y - \hat{f}(x)$ for test observations), with similar coordinates.
   (e) At the end it gives some examples how to use test.preds and test.errs to do some calculations.

   Your task is to calculate the four first quantities from the previous item for this example (excluding training error) and plot them on the y-axis against the value of k on the x-axis.
   Hint: Some of the quantities may simply be read from the code (such as the irreducible error), others have to be calculated from the matrices test.preds and err.preds and/or the vectors f.te and/or y.te.