6.390 Introduction to Machine Learning

Recitation Week #1

Issued February 6, 2023

1. Lisa is a machine learning engineer who is tasked with predicting the air quality index in a given location based on a set of features. She is provided with five different relatively large data sets, each corresponding to a different region in the world. Each data set that Lisa is given has a corresponding testing set that Lisa does not have access to.

   Lisa sets out to use these five different large data sets to train five different models (each data set yields one model). She separates each data set that she is given into a training set that the corresponding model has access to during training, and a "validation" set that the model does not see during training.

   With the hope of estimating how well her models perform on new feature vectors, she measures the error of her models on the training set during training, and the error on the validation set after training. For each model, the client measures the error of the model on the corresponding testing set and reports the error to Lisa. These experiments yield the following results:

	error on training set	error on validation set	error on testing set
data set 1	HIGH	HIGH	HIGH
data set 2	LOW	HIGH	HIGH
data set 3	LOW	LOW	HIGH
data set 4	MID	LOW	LOW
data set 5	HIGH	HIGH	MID

   (a) Identify which data set(s) exhibit each of the following behaviors and explain your reasoning:

      i. Lisa's model is overfitting to the training set (check all that apply):

      ◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

ii. It is likely that having access to more training data would decrease error on the testing set (check all that apply):

◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

iii. Lisa's hypothesis class might not be expressive enough (check all that apply):

◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

(b) Notice that for data sets 3 and 5, the error on the validation set is significantly different than the reported error on the testing set. What are some possible causes of this discrepancy?