CS 4361/5361 Machine Learning

Fall 2020

Exam 1

The code provided includes partial implementations to perform the following operations:

1. (35 points) Use the Naive Bayes algorithm to find the features that are most useful to discriminate among two given classes. A feature is useful to discriminate classes c_1 and c_2 if the mean value for that feature among examples of c_1 is very different from the mean value of that same feature among examples of c_2. See images and note at the end.

2. (35 points) Use linear regression to find the best feature to predict the target function. For a dataset with n features, we train and evaluate n models. Model 0 leaves feature 0 out for training and testing, model 1 leaves feature 1 out, and so on. Then we return the index of the feature that was left out in the worst-performing model.

3. (35 points) Find the best value of k for the k-nearest neighbors algorithm. Evaluate k-nearest neighbors for different values of k and return the one that results in the highest accuracy.

4. (Extra credit) Return the classes predicted by k-nearest neighbors and the confidence in those predictions. In k-nearest neighbors, the confidence of a prediction for a test example x is the vote the predicted class receives divided by the vote received by all classes (usually these votes are weighted by inverse distance; see lines 51-53 in the implementation). If all the neighbors of x belong to the same class, then the confidence is 1.

Complete the implementations so they work as intended.

Notes on question 1: The first figure shows the mean attribute values for each of the classes in the MNIST dataset. The second figure shows the square difference between the mean of class 4 and the mean of class 9, highlighting the attributes where the difference between the classes is largest (bright pixels).