Austin Anderson
8/31/2015
Machine Learning
Homework 1

Introduction

The main python script is titled:

knn.py

It is thoroughly commented throughout, but generally I modified functions in the Knearest class. Those functions are:

1. Majority

   This function now takes in the Knearest self structure and indices associated with nearest neighbors. It finds the labels associated with the index values and counts the frequency of the unique labels in those indices. It orders the count result and looks for maximum count values. If a maximum is found, the associated label is appended to a vector. The function ultimately returns the median as calculated by numpy of that maximum label vector.

2. Classify

   The classify function takes in feature data and uses the BallTree structure to find the k-nearest neighbors with the query call. That call returns the distance to the neighbors and their associated indices. Those index values are then passed along to the majority function that returns the majority label.

3. Confusion Matrix

   The confusion matrix runs the knn classifier on the input data set. It reads through the set, classifies it, and looks up the true label associated for that segment of data. It sums data values in a dictionary at address [true value][predicted value]. If the classifier works perfectly, only entries on the diagonal will have non-zero values as diagonal entries row and column index values are equal. The function returns this matrix.

Results

The code runs very slow if all the data is used in classification. With the --limit flag set to 500, the code runs decently fast. Accuracy at K=3 is around 83%. Increasing K does not seem to appreciably slow down runtime, but it does lower accuracy somewhat. Increasing the data limit improves performance at the cost of significantly increased computation time.

Data Limit	K	Computation Time	Accuracy
500	3	6.86 seconds	83%
500	5	6.91 seconds	80%
500	11	6.99 seconds	78%
1000	3	12.08 seconds	87%
5000	3	46.80 seconds	94%
No Limit	3	438.12 seconds	97%