

Homework 4

Due: 02/16/2012 (before class)

February 9, 2012

Problem 1 (20pt): Experiment with k NN classifier

Data sets There are two data sets, i.e., segment and heartstatlog, will be used by this problem. They can be downloaded from <http://www.cse.msu.edu/~cse847/assignments/data-1.zip>. Each data set is allocated to a separated directory under the zip file, and has the same format. Every data set includes three files: “xxx_trainSet.txt”, “xxx_trainLabels.txt”, and “xxx_test.txt”. Each row of “xxx_trainSet.txt” corresponds to a data point whose class label is provided in the same row of “xxx_trainLabels.txt”. Each row of “xxx_testSet.txt” corresponds to a data point whose class label needs to be predicted. You will train a classification model using “xxx_trainSet.txt” and “xxx_trainLabels.txt”, and use it to predict the class labels for the data points in “xxx_testSet.txt”.

1. Build a k Nearest Neighbor (k NN) classifier that uses the leave one out cross validation approach for determining the best k value. Report the results for the leave one out cross validation results for both training data sets.
2. The key hypothesis behind the k NN approach is that two data points are likely to share the same class if they are close to each other. Use the training data in the file to examine this hypothesis. You need to describe how you test this hypothesis, and report the evaluation results with the training data.

Problem 2 (10pt): KD-tree

In the class, we discussed KD-tree for efficiently finding the nearest neighbors. One key step in KD-tree is to compute the shortest distance from a query point $\mathbf{q} \in \mathbf{R}^d$ to all the data points in a hyper-rectangle Δ that is defined as

$$\Delta = \{\mathbf{x} = (x_1, \dots, x_d) \in \mathbf{R}^d : x_i \in [a_i, b_i], i = 1, \dots, d\}$$

where a_i and b_i specify the lower and upper bound for the i -th attribute, respectively. Describe your approach for efficiently computing the shortest distance from query \mathbf{q} to the hyper-rectangle Δ , i.e.,

$$d(\mathbf{q}, \Delta) = \min_{\mathbf{x} \in \Delta} \|\mathbf{x} - \mathbf{q}\|_2$$

Since $d(\mathbf{q}, \Delta)$, the shortest distance from \mathbf{q} to Δ , may be significantly smaller than the shortest distance from \mathbf{q} to the data points in Δ , discuss possible ways to reduce the gap between the two distances.

(Optional) Problem 3 (10pt): Nearest Neighbor

Consider N data points uniformly distributed in a p -dimensional unit ball centered at original. Consider the nearest neighbor estimate at the original. Prove that the median distance from the original to the closest data point is:

$$d(p, N) = \left(1 - 2^{-1/N}\right)^{1/p} \quad (1)$$

Furthermore, show that the above expression can be simplified as $d(p, N) \sim 1 - \log N/p$ when $N \gg 1, p \gg 1$, and $p \gg \log N$.