| | |
|---|---|
| **Machine Learning** | **Homework 1** |
| **CSCI 5622 Fall 2017** | **Due Time Sep 15, 2017** |
| **Name: Chi Chen** | **CU identitykey: chch4713** |

# 1 K-nearest Neighbor (40pts)

**Solution.** *1. What is the role of the number of training instances to accuracy (hint: try different –limit and plot accuracy vs. number of training instances)? 2. What numbers get confused with each other most easily? 3. What is the role of k to training accuracy? 4. In general, does a small value for k cause overfitting or underfitting?*

# 2 Cross Validation (30pts)

**Solution.** *1. What is the best k chosen from 5-fold cross validation with –limit 500? 2. What is the best k chosen from 5-fold cross validation –limit 5000? 3. Is the best k consistent with the best performance k in problem 1?*

# 3 Bias-variance tradeoff (20pts)

**Solution.** *Derive the bias-variance decomposition for k-NN regression in class. Specifically, assuming the training set is fixed S = (x1, y1), . . . , (xn, yn), where the data are generated from the following process y = f(x) + ,E() = 0,Var() = 2. k-NN regression algorithm predict the value for x0 as hS(x0) = 1 k k l=1 E((y0 hS(x0))2). Prove that y(l), where x(l) is the lth nearest neighbor to x0. Err(x0) is defined as Err(x0) = 2 + f(x0) f(x(l)) 2 + .*