Machine Learning
Fall 2014
Lab #2
Nearest Neighbors
Due September 24, 1:30 p.m.

Your task is to add functionality to your nearest neighbors program in order to speed up classification.

Do the following:
1. Implement a simple attribute selection algorithm that chooses the M attributes with the highest variance across the training set; experiment with various values on M. About 10% of the attributes in the training set have zero variance, meaning they have the same value for all examples and are thus useless. I ran quick experiments and obtained an accuracy of 97% for M=400, 92.7% for M=100 and 56.3% for M=10.
2. Implement and evaluate an attribute selection algorithm using the aggregation approach discussed in class. You may choose a subset of the training set to do attribute selection, otherwise this process will be too slow.
3. Implement and evaluate an attribute selection algorithm using the deletion approach discussed in class.
4. Implement and evaluate Hart’s algorithm for building a more compact training set
5. (Mandatory for students taking CS 5354, extra credit for students taking CS 4365) Implement the kd-tree algorithm to speed up the search for nearest neighbors. You may get better results if you put at the root the attribute with the largest variance, then the one with the second largest variance, and so on.
6. (Extra credit) Apply PCA preprocessing to the data, then apply part 5.

Write a report describing your results and include your source code as appendix. I am particularly interested in your accuracy, running time, and examples of misclassified instances.