CS4361/5361 Machine Learning

Fall 2019

Lab 1 - k-nearest-neighbors

Due Friday, September 6, 2019

For this lab you will implement the k-nearest-neighbor algorithm and some of its variants. You will also test it on a classification task, the MNIST digit classification, and a regression task, the prediction of solar particle flux.

The program *zeroR.py* available in the class web page implements an extremely simple model for classification and regression, which always predicts the majority class in the training data, in the case of classification, and the mean target value in the training data, in the case of regression. It also shows the structure we will use for all supervised learning models, using a *fit* function to train the model and a *predict* function to make predictions.

Your task is to implement the k-nearest-neighbor algorithm and apply it to the same data sets. Experiment with different values of k and compare the weighted and unweighted versions of the algorithm.

Since the k-nearest-neighbor algorithm is very slow, performing experiments with the full training and/or test set might be unfeasible, especially during the development stage. Feel free to use a subset of either set for your experiments (for example, do x_test = x_test[::50], which will select every 50th example in the set). Of course, your results will be worse when you use a smaller training set.

Extra credit for 4361, mandatory for 5361. Implement an optimization to the algorithm that either improves results or reduces training time. Options include but are not limited to:

1. Attribute selection
2. Principal component analysis
3. k-d tree
4. Clustering training data
5. Graph-based approaches
6. Alternate distance metrics
7. Hashing

Write a report including (at least) the following items:

1. Problem description
2. Algorithms implemented
3. Experimental results, including accuracies or mean squared errors and running times.
4. Discussion of results
5. Conclusions
6. Appendix: Source code