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.

In addition to the standard k-nearest-neighbor model, implement the following two versions of approximate nearest-neighbor that attempt to speed-up the prediction process:

1. Attribute selection. Use only a subset of the features (columns) in the dataset. Choose the n columns that have the highest variance (where n is a user-selected parameter) in the training data, and use them to compute distances at prediction time.

2. Geometric hashing, as described in class.

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