

Homework 3 – Implementing regression

Machine Learning / Greg Hamerly

February 9, 2007

This homework is due in class on Friday, February 16, 2007.

For this homework, you will use Matlab to implement least-squares linear regression and nearest-neighbor regression for a simple dataset.

Dataset Download the “prostate” dataset at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html> and load it into Matlab. You should divide the dataset into a training set and a test set according to the last column (but you shouldn’t keep the last column in the dataset). Likewise you should remove the first column (the numbering). The value to predict is the ‘lpsa’ column, so you may want to separate that into its own variable.

Least-squares estimator Find the least-squares estimate of the linear model with the training set, and include an intercept term. Give your estimate of the least-squares parameters.

Least-squares prediction strength For the linear model you just estimated, find the sum-of-squared-error for the model when you apply it to the test set. Analyze your findings.

k -nearest neighbor model Write software to compute the k -nearest neighbors in the training set, and makes a regression prediction based on the average response of the k neighbors.

Nearest neighbor prediction strength For the nearest neighbor model, find the sum-of-squared-error for the model when you apply it to the test set. Try this with $k=1$, $k=5$, and $k=10$. Analyze your findings.

Your report should be written in L^AT_EX. You should describe how you set up your experiments, make any relevant graphs to illustrate your findings, and analyze your results. Please also comment on what you learned. Finally, please provide the code you wrote as an appendix.