Assignment 2: Linear Regression

CS489/698 – Winter 2010

Out: February 11, 2010
Due: February 25, 2010

Be sure to include your name and student number with your assignment.

1. **[50 pts]** In class, we discussed several loss functions for linear regression. However all the loss functions that we discussed assume that the error contributed by each data point have the same importance. Consider a scenario where we would like to give more weight to some data points. Our goal is to fit the data points (x_i, y_i) in proportion to their weights r_i by minimizing the following objective:

$$L(w, b) = \sum_{i=1}^{m} r_i (y_i - w x_i + b)^2$$

   where w and b are the model parameters, the training data pairs are (x_i, y_i). To simplify things, feel free to consider 1D data (i.e., x_i and w are scalars).

   (a) **[25 pts]** Derive a closed-form expression for the estimates of w and b that minimize the objective. Show the steps along the way, not just the final estimates.

   (b) **[25 pts]** Show that this objective is equivalent to the negative log-likelihood for linear regression where each data point may have a different Gaussian measurement noise. What is the variance of measurement i in this model?

2. **[50 pts]** On the course webpage you will find 2 datasets for experimentation with regularized linear regression. Each dataset comes in 4 files with the training set in `trainN.csv`, the test set in `testN.csv` and the corresponding regression values in `trainN_val.csv` and `testN_val.csv`. In each file, there is one row per data record and one column per attribute. Your task is to explore the effect of the parameter λ that determines the importance of the regularization term. Vary lambda from 0 to 50 by increments of 1. For each λ compute the best w that minimizes regularized mean squared error of the training set and then estimate the mean squared error of the test set. For each dataset, plot the mean squared error as a function of λ and discuss your findings: how does lambda affect mean squared error? which λ is best and why?

   What to hand in:

   (a) printout of your code

   (b) printout of the mean squared error for each λ that shows that your code is working well

   (c) two graphs (one for each dataset)

   (d) discussion of the results

   Suggestion: while you are free to use any programming language, it will be easier to program in an environment with good matrix support and a library to solve linear systems. For instance, Matlab provides an excellent environment for matrix computation and to solve linear system $Ax = b$, it suffices to type `x = A\b`.