CPSC-340 Machine Learning and Data Mining 2008

Homework # 4
Due Friday, Nov 7 1pm.

NAME:_____

Signature:_____

STD. NUM: _____

General guidelines for homeworks:

You are encouraged to discuss the problems with others in the class, but all write-ups are to be done on your own.
Homework grades will be based not only on getting the "correct answer," but also on good writing style and clear presentation of your solution. It is your responsibility to make sure that the graders can easily follow your line of reasoning.
Try every problem. Even if you can't solve the problem, you will receive partial credit for explaining why you got stuck on a promising line of attack. More importantly, you will get valuable feedback that will help you learn the material.
Please acknowledge the people with whom you discussed the problems and what sources you used to help you solve the problem (e.g. books from the library). This won't affect your grade but is important as academic honesty.
When dealing with python exercises, please attach a printout with all your code and show your results clearly.

1. **(Ridge regression)**

   (i) Download the diabetes dataset from the LARS website by clicking on the link 'diabetes data' at:

   `http://www-stat.stanford.edu/~hastie/Papers/LARS/`

   In this diabetes study, 442 patients were measured on 10 baseline variables: age, sex, body mass index, average blood pressure, and six blood serum measurements. These features correspond to the 10 first columns of the file *diabetes.data*. The output variable (the 11th column of the file *diabetes.data*), is a quantitative measure of disease progression one year after baseline. Your goal is to constract a model using ridge regression to predict this output (response) variable. To do this, follow the following steps:

   (a) Load the data.

   (b) Write a file to normalize the data, so that it looks like the data in file

   `http://www-stat.stanford.edu/~hastie/Papers/LARS/diabetes.sdata.txt`

   (c) Choose the first 100 patients as the training data. The remaining patients will be the test data.

   (d) Using a range of regularizers (δ's), plot the 10 θ's in the y-axis against δ in the x-axis.

   (e) For each computed value of θ, compute the train and test error. Choose a value of δ that gives you the lowest minmax crossvalidation error. What is this value? Plot the train and test errors as a function of δ.

   (f) For the best theta, plot separetely, using the command subplot, the train and test error as a function of the patient number. That is, for each patient show the actual response and the prediction.

   You must hand in the code properly documented and all the plots. In this question, you need to make several choices. You will be marked according to how you make these choices and how well you are able to present the data. Think of yourself as doing a consulting job for a clinic downtown.