161.326: Statistical Machine Learning
Assignment 2

Semester 2, 2010

This assignment is due on **Monday 27th September**, and is worth 20% of your grade. The assignment should be submitted through WebCT in the form of a zip file containing your write up (as a text file; Word or pdf file is also sufficient) and all of the programs that you used, as well as the output from them. We must be able to extract and run your code, and reproduce your results.

1. Use cross-validation on the prostate cancer training data to select the best model from the reduced regression model (predictors 1,2,4 and 5), ridge regression, and principal components regression. Compare your results to those obtained by fitting the models to the full training data and then evaluating them on the test data.

2. For the prostate cancer data, consider the the full regression model, the reduced regression model (only predictors 1,2,4 and 5), and the ridge regression model (all from Lecture Slides 5), the backward regression model from Exercise 5.5, and the best model obtained in Question 4 of Assignment 1.

   (a) For each model:

      i. Calculate the BIC, and use these to calculate the posterior probabilities

      ii. Compare the prediction errors on training and test data

   (b) Use the posterior probabilities to form an average model, $\hat{\beta} = \sum_m \Pr(M_m|Z)\hat{\beta}_m$ (see Lecture Slides 6.53) and compare its performance on the test data with that of the bagged model (Exercise 6.6).

   (c) Discuss

   NB. You will need to adjust for the centering in the ridge regression

3. When you arrive at the pub, your five friends already have their drinks on the table. Jim has a job and buys the round half of the time. Jane buys the round a quarter of the time, and Sarah and Simon buy a round one eighth of the time. John hasn't got his wallet out since you met him three years ago.

   Compute the entropy of each of them buying the round and work out how many questions you need to ask (on average) to find out who bought the round.

   Two more friends now arrive and everybody spontaneously decides that it is your turn to buy a round (for all eight of you). Your friends set you the challenge of deciding who is drinking beer and who is drinking vodka according to their gender, whether or not they are students, and whether they went to the pub last night. Use ID3 to work it out, and then see if you can prune the tree.

Drink	Gender	Student	Pub last night
Beer	T	T	T
Beer	T	F	T
Vodka	T	F	F
Vodka	T	F	F
Vodka	F	T	T
Vodka	F	F	F
Vodka	F	T	T
Vodka	F	T	T

4. The `iris.py` program in the Chapter 9 folder demonstrates the use of both the k-means and Self-Organising Map (SOM) algorithms on the iris data that we saw when we looked at supervised neural networks. When you run the code, it shows you a plot of which network neurons were used to identify the different classes as red squares, and green and blue triangles. Any neurons that did not match any of the inputs are shown as black dots. The two plots are based on the training data and test data respectively. You should use this code (and dataset) as the basis for your answer to this question. Choosing an optimal network size for the SOM is mostly just a question of trying out different values. The question is what you should optimise the value on. You don't want any neurons that match inputs from two different classes, but you don't want too many neurons, either. Devise a scoring scheme that evaluates how well each map size does based on these two criteria, plus anything else that you think is important. Use it to optimise the size of the network for the iris data.