15.062 – Data Mining, Spring 2003
Midterm Solutions

Problem 1 (25 points, 5 points each)

1) FALSE: Adding too many variables can result in our model over-fitting to the training set. This can adversely affect the performance on the validation set.
2) FALSE: All-subsets selection method is the method guaranteed to find the best model for a given criterion. Step-wise selection is only a heuristic – there is no guarantee of optimality.
3) TRUE: To be specific, only the constant term for C1 will increase relative to the constants for C2 and C3, while the other coefficients will remain unchanged. Intuitively, if class C1 becomes more likely, we want the corresponding classification function to more often take the highest value among the three classification functions.
4) TRUE: The misclassification rate on the training data will fail to capture the problem of over-fitting.
5) FALSE: Not necessarily, especially if the neural network has multiple hidden layers. If the activation functions at any of these nodes were nonlinear mappings of the input, then the separation boundary would not necessarily be linear.

Problem 2 (10 points)

We know there was an 8% misclassification rate in the original validation data. Thus, there was a total of 0.08*400 = 32 misclassifications in this data set. Among the duplicated cases in the validation set, we know there were no errors (since with 1-NN we fit the training set perfectly). Thus, in the true validation set we have 32 errors and 300 observations, giving a revised misclassification rate of 21/300 = 0.107 or 10.7 %.

Problem 3 (10 points)

Since the Naïve Bayes classifier assumes independence of the variables, we can simply exclude the missing variables in calculating the conditional probability of that point being in a given class.

Problem 4 (15 points)

Assume, Sex = 1 is male and Sex = 0 female. Then, according to the tree, the company should send promotions to a customer (i.e., the customer is a Class 1 customer) if:

- He/she is less than 23,
- He/she is between 28 and 41.5 years old
- Is male and is between 41.5 and 44 years old.

Problem 5 (20 points)

(a) Given that the performance of logistic regression is significantly better than the neural nets for *both* training and validation set, this seems to be a case of under-fitting. Notice that under certain conditions (single-layer, maximum likelihood, sigmoid activation function), neural nets give the same output as logistic regression. Thus, its poor performance indicates that the default parameter values are insufficient in classifying the data set.

(b) Under-fitting can be due to several parameters. Since the logistic regression seems to be successful, we can simply change the parameters of the neural net to mimic logistic regression. However, this may still be insufficient if the backprop algorithm of neural net could not terminate at a minimum. In such a case, we may want to increase the number of epochs and decrease the step length (smaller step length help guarantee convergence, but would require larger epochs). It is also possible that the backprop algorithm converged at a a poor local minimum, in which case we may want to rerun the model with a different ordering of the training set data.

(c) Given that the training set error is significantly smaller than the validation set error, this seems to be a case of over-fitting the network to the training set.

(d) There are several parameters that can be changed. Decreasing the number of nodes and layers would be the most obvious step. With fewer nodes, the complexity of the network decreases and thus the training set error will increase. Hopefully, this will be compensated by a significant decrease in the validation set error.

Problem 6 (20 points)

(a) (4 points) The base case represents a female respondent of 0 years of age who does not live in a large city. She has no deductible, and is not at fault in the accident.

(b) (6 points) We can quantify the comparison using the odds ratio as follows:

$$\frac{\text{odds of fraud for FaultCode} = 1, \text{all others} = 0}{\text{odds of fraud for base case}} = \exp(-1.738) = 0.176$$

Thus, we can say that, all other things held constant at the base case values, odds of fraud are reduced by a factor of 0.176 when the policy holder is at fault compared to when the policy holder is not at fault.

(c) (4 points) Odds of fraud increases with age, since the age coefficient = 0.06>0.

(d) (6 points) Predicted probability of fraud in this case:

$$\frac{\exp(53.119 - 0.081 + 0.367 + 0.06(30) - 0.142(400))}{1 + \exp(53.119 - 0.081 + 0.367 + 0.06(30) - 0.142(400))} = 0.169$$