Mid Term Exam

15.062 Data Mining

Problem 1 (25 points)
For the following questions please give a *True* or *False* answer with one or two sentences in justification.

1.1 A linear regression model will be developed using a training data set. Adding variables to the model will always reduce the sum of squared residuals measured on the validation set.

1.2 Although forward selection and backward elimination are fast methods for subset selection in linear regression, only step-wise selection is guaranteed to find the best subset.

1.3 An analyst computes classification functions using discriminant analysis for a data set with three classes C1, C2 and C3. She assumes that all three classes are equally likely to arise in the application. She later learns that the probability of C1 is twice that of C2 and C3. The probabilities for C2 and C3 are equal. If she re-computes the classification functions using this information, the value of the classification function for C1 will increase for every data point.

1.4 A classification model's misclassification rate on the validation set is a better measure of the model's predictive ability on new data than its misclassification rate on the training set.

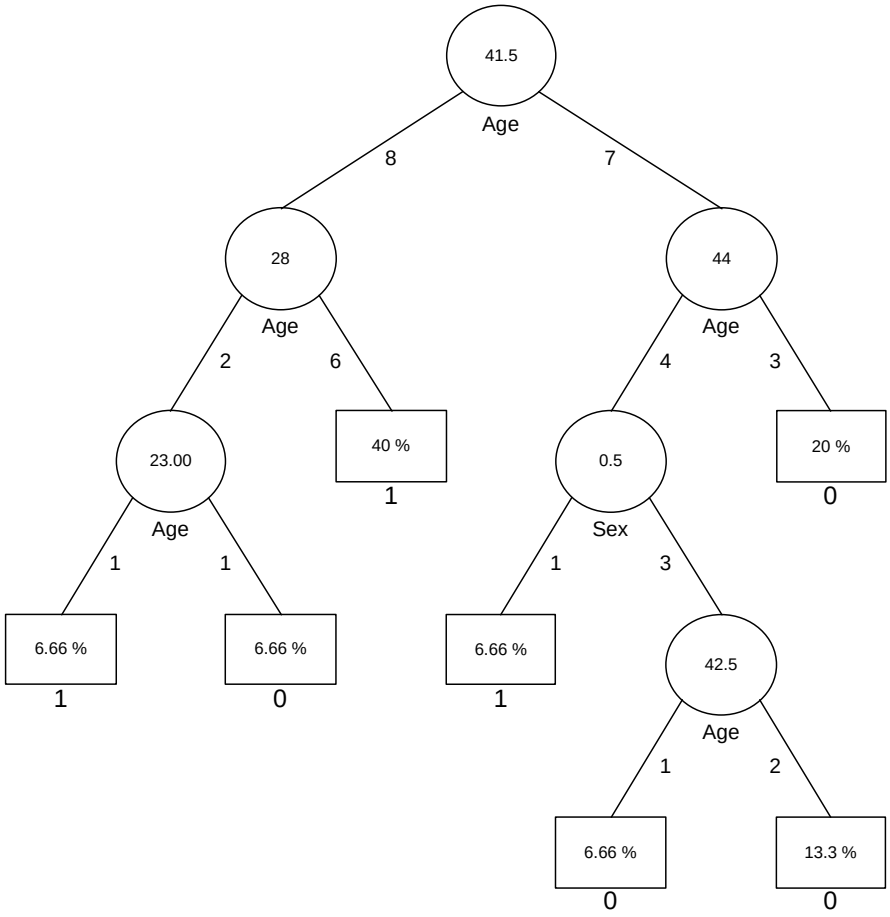
1.5 A neural net classifier for two classes constructs a separating boundary between the classes that is linear in weighted sums of the input values.

Problem 2 (10 points)
A dataset of 1000 cases was partitioned into a training set of 600 cases and a validation set of 400 cases. A k-Nearest Neighbors model with k=1 had a misclassification error rate of 8% on the validation data. It was subsequently found that the partitioning had been done incorrectly and that 100 cases from the training data set had been accidentally duplicated and had overwritten 100 cases in the validation dataset. What is the misclassification error rate for the 300 cases that were truly part of the validation data?

Problem 3 (10 points)
A Naïve Bayes classifier has been constructed with ten variables. A particular case that has to be classified has information only on eight variables. How would you use the classifier for this case?

Problem 4 (15 points)

A credit card company has created a classification tree shown below (in XLMiner format) for a promotion consisting of a special offer on credit card insurance. The class 1 corresponds to customers who responded, class 0 consists of those who did not. The company will use the tree to send promotion offers to dormant customers to persuade them to begin to use the card. The training data consists of individuals who responded to the promotion. Age, sex and income were used as input variables for these customers.



The company would like to have a few, simple, English language rules that embody the decisions represented by the tree. Write out succinctly the rules you would suggest.

Problem 5 (20 points)
When running an Artificial Feed-forward Neural Network and Logistic Regression on a data set, we obtained the following misclassification errors:

Data set A	**Parameters**	**Training set error**	**Validation set error**
Neural Network	XLMiner defaults	66.67%	75.00%
Logistic Regression	50% cutoff	6.23%	7.04%

a. What do you think is happening with the Neural Network? Explain your reasoning.
b. What parameters would you change? In what direction? Explain your reasoning.

On another data set, we had the following results:

Data set B	**Parameters**	**Training set Error**	**Validation set error**
Neural Network	XLMiner defaults	2.11%	45.24%
Logistic Regression	50% cutoff	16.23%	12.02%

c. What do you think is happening with Neural Nets? Explain your reasoning.
d. What parameters would you change? In what direction? Explain your reasoning

Problem 6 (20 points)
An insurance company has examined a random sample of 190 automobile accident claims for fraud. A logistic regression model is fitted to this data with the dependent variable being coded as one for a case that was fraudulent, and as zero otherwise. The five independent (predictor) variables included in the model are:

i. CityCode: =1 if the claimant lived in a large city, =0 otherwise;
ii. SexCode:=1 for males, =0 for females;
iii. Age in years;
iv. FaultCode:=1 if the fault in the accident was that of the policy holder, =0 otherwise;
v. Deductible Amount (in dollars).

The model estimate for the logarithm of the odds of fraud is:

53.119 - 0.081×CityCode + 0.367× SexCode + 0.060×Age - 1.738× FaultCode - 0.142×Deductible Amount

(a) Describe in words the base case claimant whose odds for fraud are $e^{53.119}$.

(b) What are the odds for fraud in an accident where the policyholder was at fault compared to one where the fault was not that of the policyholder, assuming all other variables take their base case values?

(c) Does the odds for fraud increase or decrease with age?

(d) What is the probability of fraud in a claim by a male policyholder aged 30 years, who lives in a major city, has a deductible of $400 and who was not at fault in the accident?