NAME:

Data Mining Techniques: Quiz 2

1. Missing values are often a problem for classification and prediction.
a) State ONE important reason why a missing input attribute value in a data record could be a problem when training a tree. Briefly explain ONE good way for dealing with this problem. (15 points)

Problem 1: When considering the attribute as a split attribute for a node, we need to compute Gini or information gain etc. If the attribute value is missing, we do not know which child to assign the tuple to, and hence cannot compute the Gini, entropy etc for the child nodes.

Problem 2: If the training algorithm selected the attribute as the split attribute for the node, we have to send each training tuple down one of the child branches for the next recursive call of the training algorithm. If the attribute value is missing, we do not know where the tuple should go.

Solution 1 (for both problems): Fill in missing values during pre-processing.

Solution 2 (for problem 1): Ignore the tuple when computing the Gini or entropy etc for the child nodes.

Solution 3 (for problem 2): Send a fraction of the tuple down each child branch. Choose the fraction according to the fraction of the training tuples at the node whose split attribute values are not missing.

b) State ONE important reason why a missing input attribute value in a data record could be a problem when trying to use the tree to predict the class for this record. Briefly explain ONE good way for dealing with this problem. (15 points)

Problem: When the tuple reaches a node that splits on the attribute whose value is missing, we do not know which child branch to choose for continuing the tree traversal.

Solution 1: Fill in missing values before running the tuple through the tree.

Solution 2: Send a fraction of the tuple down each child branch, then compute the final prediction as the weighted average of the individual predictions from each branch. Choose fractions according to the distribution of training tuples at the node.

2. In class we talked about pre-pruning and post-pruning techniques for trees.
   a) Why do we prune a decision tree? (10 points)

The tree might overfit to the training data when it is too big. Some tree branches might reflect anomalies or idiosyncrasies of the training data. Pruning therefore can help address overfitting.

   b) Briefly explain ONE way to determine when to stop post-pruning a tree. (10 points)

Solution 1: We use a validation set of data tuples that were not used for training. We prune the tree as long as its performance on the validation data improves or stays about the same.

Solution 2: We use the MDL (minimum description length principle), finding the tree that minimizes cost(Model) + cost(Data | Model) for some appropriately defined cost functions.

3. In what sense is the Naïve Bayes classifier “naïve”? (10 points)

Naïve Bayes assumes class conditional independence, i.e., the effect of an attribute value on a given class is independent of the values of the other attributes. In practice this assumption is usually not correct. Nevertheless, Naïve Bayes often performs surprisingly well.

4. Compute the following probabilities for the data set given below: (24 points)

age	income	student	creditRating	buysComputer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

P(buysComputer = yes) = 9/14

P(age <= 30 | buysComputer = yes) = 2/9

P(income = medium | buysComputer = yes) = 4/9

P(student = yes | buysComputer = yes) = 6/9

P(creditRating = fair | buysComputer = yes) = 6/9

P(age<=30 $\wedge$ income=medium $\wedge$ student=yes $\wedge$ creditRating=fair | buysComputer=yes) =
(Hint: use the Naïve Bayes assumption here.)

(2/9) * (4/9) * (6/9) * (6/9)

5. a) Explain in one sentence what each of the following formulas tell you. (Assume the obvious semantics for attribute names and their possible values.): (12 points)

P(creditCardTransaction = “fraudulent”) = 0.1

10% of all credit card transactions are fraudulent.

P(creditCardTransaction = “fraudulent” | purchaseLocation = “Joe’s gas station”) = 0.4

40% of all credit card transactions at Joe’s gas station are fraudulent.

P(creditCardTransaction = “fraudulent” | purchaseLocation = “Jim’s gas station”) = 0.05

5% of all credit card transactions at Jim’s gas station are fraudulent.

b) Knowing the above facts, where should a driver with common sense purchase gasoline, assuming everything else about the two gas stations is identical? Briefly explain why. (4 points)

If you pay with a credit card, it would be better to buy gas at Jim’s, because one is less likely to experience a fraudulent transaction.