Homework 5 sample solutions:

1. 7.1

**Answer:** The major steps are as follows:

- The tree starts as a single root node containing all of the training samples.
- If the samples are all from the same class, then the node becomes a leaf, labeled with that class.
- Else, the node is assigned as a "test" or "decision" attribute. This is the attribute that will "best" separate the samples into individual classes. It can be determined using a heuristic or statistical measure (such as information gain or gini index). Attributes must be categorical. (Continuous-valued attributes must be discretized.)
- Next, a branch is created for each known value (or range, if discretized) of the test attribute. The samples are then partitioned into the appropriate branch.
- The algorithm recursively repeats itself to create a decision tree for the samples at each partition.

**Stopping Conditions:**

- If all samples at a given node belong to the same class, then transform that node into a leaf, labeled with that class.
- If there are no more attributes left to create more partitions, then majority voting can be used to convert the given node into a leaf, labeled with the most common class among the samples.

2. 7.2

**Answer:** The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures). This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.

The drawback of using a separate set of samples to evaluate pruning is that it may not be representative of the training samples used to create the original decision tree. If the separate set of samples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy. Furthermore, using a separate set of samples to evaluate pruning means there are less samples to use for creation and testing of the tree. However, this is considered a drawback in machine learning, but may not be so in data mining due to the larger data sets.

3. 7.6

(a) How would you modify the ID3 algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

**Answer:** ID3 should be modified as follows to take into consideration the count of each generalized data tuple.

- The count of each tuple must be integrated into the calculation of information gain.
- Take the count into consideration to determine the most common class among the samples.
(b) Use your modified version of ID3 to construct a decision tree from the given data. 

**Answer:** The resulting tree is:

\[
\begin{align*}
\text{salary} &= 26K...30K: \\
& \quad \text{junior} \\
& = 31K...35K: \\
& \quad \text{junior} \\
& = 36K...40K: \\
& \quad \text{senior} \\
& = 41K...45K: \\
& \quad \text{junior} \\
& = 46K...50K \ (\text{department} = \text{secretary}): \\
& \quad \text{junior} \\
& \quad \text{sales}: \\
& \quad \text{senior} \\
& = \text{systems}: \\
& \quad \text{junior} \\
& \quad \text{marketing}: \\
& \quad \text{senior} \\
& = 66K...70K: \\
& \quad \text{senior}
\end{align*}
\]

(c) Given a data sample with the values "systems", "junior", and "26...30" for the attributes department, status, and age, respectively, what would a naive Bayesian classification of the salary for the sample be?

**Answer:** \(P(X|\text{senior}) = 0; P(X|\text{junior}) = 0.018\). Thus, a naive Bayesian classification predicts "junior".

(d) Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers.

**Answer:** No standard answer. Every feasible solution is correct. As stated in page 304 in the text book, discrete-valued attributes may be encoded such that there is one input unit per domain value. For hidden layer units, the number should be smaller than that of input units, but larger than that of output units.

Note: the input layer should be categories, not data instances.

(e) Using the multilayer feed-forward neural network obtained above, show the weight values after one iteration of the backpropagation algorithm given the training instance "(sales, senior, 31...35, 46K...50K)". Indicate your initial weight values and biases and the learning rate used.

**Answer:** No standard answer. Every feasible solution is correct.

4. 7. 8

(a) Plot the data. Do \(X\) and \(Y\) seem to have a linear relationship?

**Answer:** Yes, from the scatter graph, it would appear that \(X\) and \(Y\) have a linear relationship.

(b) Use the method of least squares to find an equation for the prediction of a student’s final exam grade based on the student’s midterm grade in the course.

**Answer:** \(s = 12; \bar{x} = \frac{800}{12} = 72.167; \bar{y} = \frac{888}{12} = 74\). Using Equations 7.23 and 7.24, we get \(\alpha = 32.028\) and \(\beta = 0.5816\). Therefore, the equation for predicting a student’s final exam grade based on the student’s midterm grade is \(Y = 32.028 + 0.5816X\).

(c) Predict the final exam grade of a student who received an 86 on the midterm exam.

**Answer:** Using the formula from part (b), we get \(Y = 32.028 + (0.5816)(86) = 82.045\). Therefore, we would predict that a student who received an 86 on the midterm would get 82 on the final exam.