Assignment #5

1. Why is naive Bayesian classification called “naive”? Briefly outline the major ideas of naive Bayesian classification. (25 points)

2. The following table consists of training data from an employee database. The data have been generalized. For example, “31 ... 35” for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

| *department* | *status* | *age* | *salary* | *count* |
|---|---|---|---|---|
| sales | senior | 31...35 | 46K...50K | 30 |
| sales | junior | 26...30 | 26K...30K | 40 |
| sales | junior | 31...35 | 31K...35K | 40 |
| systems | junior | 21...25 | 46K...50K | 20 |
| systems | senior | 31...35 | 66K...70K | 5 |
| systems | junior | 26...30 | 46K...50K | 3 |
| systems | senior | 41...45 | 66K...70K | 3 |
| marketing | senior | 36...40 | 46K...50K | 10 |
| marketing | junior | 31...35 | 41K...45K | 4 |
| secretary | senior | 46...50 | 36K...40K | 4 |
| secretary | junior | 26...30 | 26K...30K | 6 |

Let status be the class label attribute.

(a) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)? (15 points)
(b) Use your algorithm or any existing decision tree implementation to construct a decision tree from the given data. (15 points)
(c) Given a data tuple having the values “systems”, “26...30”, and “46–50K” for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be? Show details of your calculation. (15 points)