CSCI 4380/6380 – Data Mining Fall 2023

Assignment 3 Due date: September 29th, 11:59pm, EST

---

1. (10 points) Why is Naïve Bayesian classification called "Naïve"? Briefly outline the major ideas of Naïve Bayesian classification.

2. (30 points) 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)?

(b) Use your algorithm to construct a decision tree from the given data.

(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 Naïve Bayesian classification of the status for the tuple be?

3. (20 points) Show that accuracy is a function of sensitivity and specificity, that is, prove:

$$Accuracy = Sensitivity\frac{P}{P+N} + Specificity\frac{N}{P+N}$$

where sensitivity $= \frac{TP}{P}$, and $specificity = \frac{TN}{N}$, P is the total number of positive examples and N is the total number of negative examples.

4. (20 points) The data tuples of the figure below are sorted by decreasing probability value, as returned by a classifier. For each tuple, compute the values for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Compute the true positive rate (TPR) and false positive rate (FPR). Plot the ROC curve for the data.

Tuple #	Class	Probability
1	*P*	0.95
2	*N*	0.85
3	*P*	0.78
4	*P*	0.66
5	*N*	0.60
6	*P*	0.55
7	*N*	0.53
8	*N*	0.52
9	*N*	0.51
10	*P*	0.40

Fig. Tuples sorted by decreasing score, where the score is the value returned by a probabilistic classifier.

5. (20 points) Programming

Write an algorithm for k-nearest-neighbor classification given k, the nearest number of neighbors, and n, the number of attributes describing each tuple.