*Neha Moolchandani Course: Data Mining | Professor: Dr. Chengcui Zhang Neha Moolchandani*

I Neha Moolchandani declare that I have completed this assignment completely and entirely on my own, without any consultation with others.  I understand that any breach of the UAB Academic Honor Code may result in severe penalties.

*Assignment 5*

Assignment #5

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

Bayesian Classification is naïve as it requires each conditional probability be non-zero. . It requires initial knowledge of many probabilities which may or may not be available or involve heavy computational costs. Strength is that it is easy to implement and good results are obtained in many cases. Weaknesses are that if assumes attributes are conditional independence and therefore loose accuracy as well as practically dependencies on existing variables. It makes an assumption to simplify the model while also achieve performance. It implies that each input variable is independent. Most times it is unrealistic for real data.

Because Naïve Bates ignores relationships among words it has high bias; however, because it works well in practice it has low variance.

The major idea behind naïve Bayesian classification is classifying data by maximizing the P(X|Ci)P(Ci) (where i is an index of the class) using the Bayes’ theorem of the posterior probability.

C is class

X is a specific tuple

It is efficient in computing Binary as well as Multi-class Classifications.

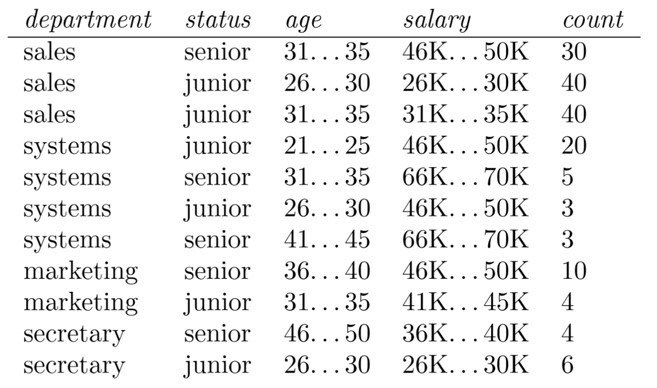
Great use in Multi-class predictions as compared to any other algorithms.

Diagram

Description automatically generated

We could do an example and compute the likelihood and then compute the post-probability using (pre-probability x likelihood.)

1. **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.**

****

**Let status be the class label attribute.**

* 1. **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)**

To modify the basic decision tree algorithm with consideration to count each generalized data tuple we can determine the common class that occurs the most among all tuples (therefore we have a variable that counts each of the occurrences and see what is the max occurrence)

* 1. **Use your algorithm or any existing decision tree implementation to construct a decision tree from the given data. (15 points)**

(salary

= 26K…..30K: Junior

= 31K…..35K: Junior

36K…..40K: Senior

=414K…..40K: Junior

=46K…..50K: (department =

= Secretary: Junior

=Sales: Senior

=Systems: Junior

=Marketing: Senior)

=66K…..75K: Senior)

* 1. **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)**

X = (Depart:Systems, Age[26-30], Salary: [46-50]

P(X | Status: Junior) = P(Department: System | Status: Junior) x P(Age:[26-30] | Status: Junior) x P(Salary = [46-50] | Status: Junior)

* P(X | Status : Junior) = (0.0157) x (113/165) = 0.0181
* P(X | Status : Senior) = P(Department : System | Status: Senior) x P(Status = Senior)
* P(X | Status : Senior) = (0.0019 x (52/156) = 0.0016
* We see that Junior>Senior
* 0.0181is greater than 0.0006
* Thus, P(Status: Junior |X) > P(Status=Senior | X)
* Thus. Naïve Bayesian Classification for this data set will provide the employee status as a Junior.