**The Center for Advanced Computer Studies**
**University of Louisiana at Lafayette**
**CMPS 566**
**Final Examination**

**Date: May 5, 2005**
**Time: 10:15 a.m. - 12:45 p.m.**

**Instructor: Dr. Vijay V. Raghavan**
**Total Marks: 75**

## PART A ( 20 marks)

**Note:** There are 5 questions. Answer *any* 4 out of the 5 questions.

Q1. Attribute-oriented Induction.

Q2. Stepwise Forward Selection.

Q3. Computing Correlation Rules based on the “a priori property”.

Q4. Measures of judging classifier accuracy.

Q5. Multi-level Association Rules using level-cross filtering.

## PART — B ( 15 marks)

**Note:** Answer *1 out of 2* questions.

Q6. Binning is an important process in dealing with noisy data and data volume. Use the following data that stands for prices of commonly sold items by *xyz* corp. The numbers in sorted order are: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Assume we want 4 bins.

a) Find the bins using a distance based methods to decide the bin boundaries. (You may remember that *MaxDiff* is such a method).

b) Use Equidepth method to determine the bins.

c) Compare the above methods in terms of advantages and disadvantages for data mining.

Q7. Suppose that a data warehouse for *Big-University* consists of the following four dimensions: *student, course, semester,* and *instructor*, and two measures *count* and *avg_grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

(a) Draw a *snowflake schema* diagram for the data warehouse.

(b) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of CS courses for each *Big-University* student.

(c) If each dimension has five levels (including **all**), such as *student* < *major* < *status* < *university* < **all**, how many cuboids will this cube contain (including the base and apex cuboids)?

(d) Explain the meaning of term “footprint” using the answer you gave to part (b) above.

## PART C (40 marks)

**Note:** Answer 2 out of 3 questions.

Q8. Suppose we want to cluster the following 6 points into 4 clusters:

$(2, 10)$ $(2, 1)$ $(3, 5)$ $(5, 9)$ $(6, 4)$ and $(4, 8)$.

a) Use first four points as starting centroids. Show understanding of *k*-means method by showing two iterations.

b) Use AGNES method to find the clusters.

c) Compare advantages and disadvantages of the above two approaches.

d) In what ways can *k*—mediod be more advantageous compared to either of the above two methods.

Q9.

| Customer ID | Height | Hair | Eyes | Credit Rating |
|---|---|---|---|---|
| $e_1$ | short | dark | blue | B |
| $e_2$ | tall | dark | blue | B |
| $e_3$ | tall | dark | brown | B |
| $e_4$ | tall | red | blue | A |
| $e_5$ | short | blond | blue | A |
| $e_6$ | tall | blond | brown | B |
| $e_7$ | tall | blond | blue | A |
| $e_8$ | short | blond | brown | B |

Using data set given in the above table. And let the credit rating be the class label attribute.

(a) What is the naive prediction with respect to credit rating A and B? Provide the prediction accuracy.

(b) Derive an optimal decision rule based on the naive Bayesian classification. Show all the concrete computation results of your derivation.

(c) Predict the credit rating for a new customer who is short, has red hair and blue eyes.

Q10.

| Credit Rating | Eye-Color | Height | Count |
|---|---|---|---|
| A | blue | tall | 2 |
|  | brown | tall | 0 |
|  | blue | short | 1 |
|  | brown | short | 0 |
| B | blue | tall | 1 |
|  | brown | tall | 2 |
|  | blue | short | 1 |
|  | brown | short | 1 |

The above table is obtained via attribute-oriented induction on the table of Q9.

a) Transform the table into a cross-tab, where rows are credit rating values and columns are the various (Eye-color, Height) value pairs. For each cell, show count, t-weight and d-weight values.

b) Map the class credit rating = “B” into a (bidirectional) quantitative descriptive rule.

(c) Construct the *context* that would be obtained from the above data, involving attributes Eye-color, Height & Credit Rating. The context should be displayed in the *flattened* form.

(d) Using the context given in part c), illustrate

i) a non-feasible set of features (attribute values), *B*, such that $(\alpha(B), B)$ is not a formal concept.

(ii) a non-feasible set of tuples, *A*, such that $(A, \beta(A))$ is not a formal concept.

(iii) Identify two formal concepts $(A_1, B_1)$ and $(A_2, B_2)$ such that one is superconcept of the other.

(iv) Identify formal concept $((A_3, B_3)$ that is not comparable to the ones in part (iii)