CSE P546 Data Mining Homework 1

Due Date: 11th April for Part A, and 18th April for Part B. We would prefer that you turn in a hard copy of your solutions at the start of class. However, you can also email them to bhushan@cs. Your file must contain your name at the top, and can be in any of pdf, Word or plaintext formats.

Part A: Data Warehousing & OLAP

1. (Han & Kamber 3.4) 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 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)?

2. (Han & Kamber 3.5 a, b) Suppose that a data warehouse consists of the four dimensions, *date*, *spectator*, *location*, and *game*, and the two measures, *count* and *charge*, where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

   (a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboid [*date*, *spectator*, *location*, *game*], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM_place in 2004?

Part B: Decision Trees

1. Mitchell 3.1

2. Mitchell 3.2

3. Suppose training set S has p positive and n negative instances. Suppose attribute A splits S into the sets $\{S_1, S_2\}$ such that S_i has p_i positive and n_i negative instances. Show that the information gain for A, as defined in Equation 3.4 in Mitchell, is always non-negative. Under what conditions will it be zero.

4. Suppose there are n boolean attributes and 1 boolean class, and the training set is composed of m distinct examples drawn uniformly from the set of 2^{n+1} possible examples, what is the probability of finding a contradiction (i.e., the same example with different classes) in the data? Explain briefly how you came up with your answer. There is no need to expand any binomial coefficients that may occur in your answer.

5. Suppose that in reduced-error pruning of a decision tree, the validation-set error rate of the subtree rooted at node N is greater than the error rate obtained by converting N into a leaf. Is it possible that N is not a leaf in the optimal pruned tree? Explain.