Introduction to Database Design, Fall 2012

Rasmus Pagh

November 20, 2012

To give an impression of what may come, the below problem is of a form that could be given at the written exam.

Data Mining (10%)

Suppose you have a data set with information on students of ITUs master's study lines. For each finished student it records admission age, institution of bachelor's degree, first-semester median grade, and completion time for the degree. The goal is to assess the probability that the completion time of a student will be 3 years or more, based on the other parameters.

A simple approach is to keep simple statistics for each combination of parameters. For example, we look at the students from KU that were admitted at age 25, got a median grade of 7, and record the percentage of them that spent 3 years or more.

a) Assume there are 10 different admission ages, 8 different institutions, 5 different possible median grades, and that students are evenly distributed over all combinations of these parameters. How much training data (i.e., how many students) are needed to predict the percentages within a deviation of 10 percent points with confidence 80%?

b) Suggest a model that could be used instead of simple statistics, reducing the amount of training data required. Briefly describe why the model would require less training data.