Data Mining Project 1: Data Pre-Processing

Released on September 13
Due when class starts on Oct 4

Specification

In this project, the students are to apply data pre-processing techniques to a gene expression dataset for colon cancer. The dataset contains 62 samples collected from colon-cancer patients. Among them, 40 tumor biopsies are labeled as "negative" and 22 are labeled as "positive" (which are biopsies from healthy parts of the colons of the same patients). Each tuple (row) in the data consists of the readings for the genes, and the class (which is the last column). Each gene is an attribute. The columns are separated by ",", which is a commonly used format in data mining. The dataset can be found at WebCT under "Projects" called p1data.txt. We number the genes 1 to N in the left-to-right order.

Your code should be able to work on other data with different number of rows and different number of columns. It may need a scan to find out the number of genes and the number of instances/rows.

There are two tasks:

Task 1. To discretize the data using equi-width binning with 4 intervals for each of the first k attributes.

Task 2. Use the entropy-based method to discretize all genes and to select the top-k genes, ranked by the information gain. Use 2 bins for each gene.

Students should first develop their code by working with a small number (e.g. 30 genes). The submitted code should work with all k genes for task 1 and work with all genes for task 2.

Your program should ask the user to give the value of k and the name of the file (containing the gene expression data, which is assumed to be located in the same folder as the executable). The executable should be called **binningexe**. It should produce four output files: ewidth.bins, ewidth.data, entropy.bins, and entropy.data.

In the ewidth.bins file, you should have the following information for each of the k genes: gene number, (bin_1_lb, bin_1_ub] bin_1_count, ..., (bin_4_lb, bin_4_ub] bin_4_count. Use one line for one gene's info.

In the ewidth.data file, you should have the result of the discretized data for the first k genes: Use a, b, c as the names of the bins, with a for the leftmost bin, and map the original data into discretized data. You should keep the class for each tuple but ignore genes k+1, k+2 etc.

You should produce similar information for the entropy.bins file, and the entropy.data file. In the new data, the genes should be re-ordered based on information gain.

Project 1 Submission Guidelines

Submit a) a CD containing your program files, and b) a printed report listing b1) information on the bins for the first 5 genes for Task 1 and b2) information on the bins for the top 5 genes for Task 2. The information should be identical to that produced by your program.

You should not deviate from these guidelines and requirements.

Correctness, efficiency, readability etc will be factors for marking.

In your code, you must include comments that show the major steps (including: inputting data, equi-width binning, entropy-based binning, outputting results, etc).

A demo will be scheduled for each student to demonstrate the compilation and execution of the project.

Notes: The dataset was first made available by U. Alon, et al. ("Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays", PNAS, 96:6745-6750, 1999).

The Work Must Be Your Own: You cannot use code found from the internet or from other sources in your projects (or homeworks). Your submissions must be your own work.