**CSE 4701, Spring 2017**
**Project 2**

**Part II – Show you know how to use Information Gain (Due: 4/19/17 (Wed), 11:59pm, WebCT, 10% penalty for a day late)**

The class version IS_READY will be available for Part II. One difference of this class version IS_READY from your Part I IS_READY is that in the class IS_READY, the STUTUS column is set "1" if DECEASED.

Your goal is to determine which particular mutations (among the 10 you have included in IS_READY) are better predictors of the subject's "Diseased" Status. You do this using Information Gain (IG) you learned in class. How to compute IG is illustrated in the separate document CDT (Conditional Distribution Table). As illustrated in CDT, you can do this second part in two steps.

1. First, create the **conditional distribution tables** for each mutation of the ten included genes. In producing the conditional distribution table, you are required to **maximally** use SQL (e.g., INTERSECT, MINUS, etc.), meaning avoid using host language for set operations unless absolutely needed.

2. Implement Information Gain (IG) procedure in your host language and rank the IGs.

**A. Data Mining Basic (Total 80 points)**

| Gene ID | IG |
|---|---|
| APC | 0.0034 |
| KRAS | 0.0012 |
| SYNE1 | 0.0011 |
| MUC4 | 0.0009 |
| ATM | 0.0001 |

Provide the top 5 highest information gain (IG) ones. The report format is a two column table in the descending order by IG as illustrated in the table. Your report should include this data in a tabular format. The look and feel of the table can be different, but you are required to generate this table programmatically.

Note that the values in this table are for illustration purpose and they are not accurate.

**B. Discussion (Total 20 points)**

| Gene ID | IG | O CNT |
|---|---|---|
| APC | 0.0034 | ? |
| KRAS | 0.0012 | ? |
| SYNE1 | 0.0011 | ? |
| MUC4 | 0.0009 | ? |
| ATM | 0.0001 | ? |

This time, compute Overlap CNT (att1, att2) for your top five IG Genes, where Overlap CNT (att1, att2) is merely counting how many subjects have 1's in both att1 and att2. You are required to create a new table, this time adding this O CNT (Overlap CNT) as the third column as illustrated, while maintaining the descending order of the list based on IG. This table should be created programmatically and included in your report.

Next, examine both IG and O CNT and check if the ordering based on IG (higher gain the better) and O CNT (more overlapping the better) coincide. Note that IG is to check if "the higher gain the better" and O CNT is to check if "the more overlapping 1's, the better", as far as predictable power is concerned. If they do not coincide, can you explain why? Your write-up should be less than 50 words.

**C. Optional Extra Points (Up to Total 10 points)**

| Gene-Pair | IG | O CNT |
|---|---|---|
| APC-ATM | 0.0044 | ? |
| KRAS-ATM | 0.0039 | ? |
| SYNE1-ATM | 0.0038 | ? |
| MUC4-APC | 0.0038 | ? |
| KRAS-MUC4 | 0.0034 | ? |

Your goal is to try which "pairwise" combination could provide a more predictable power. That is, examine the cases in which a subject has mutations in both $G_i$ and $G_j$ genes where $G_i$ and $G_j$ are any two from the ten genes you have been examining.

Lastly, compare the analysis outcomes of B and C in less than 50 words – how you interpret the difference if there is.

**Report Format:** (i) Source code, (ii) Table for A, (iii) Table for B, (iv) Write-up of 50 words for B, and (v) Table for C with a write-up less than 50 words (optional, if challenging extra points).