

# CS573: Homework 5

Due date: Monday December 6, 4pm to CS mailroom

## Programming assignment

In this assignment you will implement an sequential pattern mining algorithm and apply it to the *Molecular Biology (Promoter Gene Sequences) Data Set* from the UCI ML repository. The data set is available at:

<http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Promoter+Gene+Sequences%29>.

It consists of 108 snippets of the DNA sequence for E. coli, where positive instances correspond to “promoters” (which initiate the process of gene expression) and negative instances correspond to non-promoters. In this assignment, you will consider the sequences associated with the first 10 positive instances, using the 57 characters of their sequences to look for patterns.

Again, you can use a language of your choice to implement the algorithms (e.g., Java, C, Python, R). Please hand in a hard copy of your code with the assignment.

1. Implement the Multi-Stream Dependency Detection (MSDD) algorithm to find patterns in sequential data streams. (15 pts)

The algorithm is described in the paper:

T. Oates, P. Cohen (1996). Searching for Structure in Multiple Streams of Data, ICML'96. (<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.9668>)

Download the data above and select the first set of 10 positive instances to use below. Remove the first attribute that corresponds to the class labels, treat the second attribute as the stream identifier and attributes 3-59 as a stream of tokens.

Run the MSDD algorithm with  $w_p = w_s = 2$ ,  $\delta = w_p$ ,  $k = 20$ ,  $f := G$  statistic, to construct a set of rules to characterize the transformed data. Report the rules themselves, along with their numerical scores (e.g., G statistic, probability). *Note that the G statistic is not defined if any cell count in the contingency table is zero. To address this, you can either use Laplace corrected cell counts, or use  $\log(0) = 0$  and  $x/0 = 0$  in your calculations.*

2. Evaluate the effect of pruning. (5 pts)

For  $w_p = [1, 2, 3, 4, 5]$ ,  $w_s = 1$ ,  $\delta = w_p$ , calculate the size of the pattern space (i.e., the total number of possible rules that the unpruned algorithm would consider). In addition, run your algorithm and record (i) the total number of nodes added to the open list, and (ii) the total number of nodes expanded in the algorithm. Plot all three measures as a function of  $w_p$ . Discuss the results.

3. Evaluate the stability of the constructed rule set. (5 pts)

Partition your streams into 3 temporal folds (i.e., tokens 3-21, 22-40, 41-59) and construct rules on two of the folds. Compare the rule sets and their scores, identifying similarities and differences.

4. Adjust for multiple comparisons (5 pts)

*Reference: D. Jensen and P. Cohen (2000). “Multiple Comparisons in Induction Algorithms.” Machine Learning 38: 309-338.*

- (a) Discuss how association rule algorithms can suffer from multiple comparison problems.
- (b) Test the significance of the discovered rules reported in (1). The G statistic is approximated well by  $\chi^2$ , so a  $\chi^2$  distribution (for the same size contingency table) can be used to test significance with  $\alpha = 0.05$ . State the appropriate degrees of freedom and give the resulting cutoff value for the G score. For the set of rules discovered above, report which scores are significant.
- (c) Describe a Bonferroni correction to adjust for multiple comparisons in the significance tests.
- (d) Recompute the significance of the discovered rules using Bonferroni correction and discuss any change in the significance assessments.