Project Description
biological data science
spring 2019

Project Option 1: Clustering

Given a network of nodes and undirected edges, find a 'desirable' clustering of the nodes. Weights will be supplied but unweighted approaches are targeted.

'Desirable' features include lack of sphericity assumption, lack of linear ordering assumption, lack of Euclidean-space assumptions, lack of bias against small clusters, the ability to sift out singleton nodes, robust stability across similar trials, and scalability (time and/or memory). Of course you are not expected to achieve all of these objectives.

Types of contributions:

1. Novel algorithm and implementation
2. Novel objective function with standard search algorithm
3. Standard objective function with novel search algorithm
4. Standard approach that has been improved for scalability or to overcome an existing bias

Input format: gml file

Output format: a text file containing a list of *n* numbers, where *n* indicates the number of nodes and the i^{th} number in the list indicates the cluster number for the i^{th} node. (Cluster numbers should range from 1 to k.)

Project Option 2: SyncScore

Given a dataset representing expression levels for a set of genes or proteins, assign a score that measures the synchronous nature across the individuals. In other words, if there is a gene/expression pattern that is exhibited by a number of individuals, the SyncScore should be high. Inverse patterns should contribute to the score and not reduce it.

Input format: a text file with each row representing a gene/protein and each column representing an individual. You can assume one header row and one header column.

Output format: a real-valued number, with a higher number indicating greater synchronicity.

General Information

For both projects, we should assume non-Euclidean space (e.g. triangle inequality is not necessarily obeyed and Euclidean distance is ill-defined).

This project can be submitted by an individual or a team of students. Each team must submit an agreed-upon breakdown of contributions to the project and each team member will be scored for their own merits. Graduate students will be held to a higher standard.

The project is worth 500 points, with weights and due dates as follows:

Item	*Points*	*Date*
Project proposal	100	March 3
Code	150	March 25
Final presentation	100	varies
Summary and reflections	150	April 29

Project proposal:

A 5-minute presentation of your project idea. Slides must be submitted on Canvas no later than midnight on Sunday, March 3rd. I will randomly assign presentation times for Monday the 4th and Wednesday the 6th and will merge the slides into a single deck for smooth transitions between presentations.

You will be graded on the reasoning behind your ideas and your ability to communicate those thoughts. *You are not required to base your final project on the idea presented*. If you borrow ideas from classmates, you ***must*** include this information (along with all other outside resources) when you submit your code. You will be given extra credit if a fellow student uses your ideas.

If you have already formed a team, further development of your project idea is expected and you should prepare a 10-minute presentation. You can choose how many of you participate in the actual presentation (and this should be tracked in your member contributions).

Code:

Submitted software ***must*** compile using a makefile and run on delmar (70 points).

The core computations of your code must be written in C or C++ for efficiency, but the rest of the software package may be in another language if desired. Your code must adhere to department standards, as described in *ProgrammingStandards.pdf* (40 points).

README file should be easy to understand and include outside resources; results from benchmark trials should be included with logical names, with all results summarized into a table; and a signed academic integrity statement must be turned in (40 points).

Final presentation:

The final presentation can be made by presenting a poster at an approved meeting or by giving an oral presentation in class on May 6th or 8th. You will be graded on your ability to clearly communicate the following topics: the relevance of the problem, the approach you took, the results of your approach, and a brief discussion.

Summary and reflections:

This document should be a *concise* summary of your ideas, efforts, and reflections about what you learned in the process. You are encouraged to contrast and compare your approach with others' in the class. Each person should submit their own summary and the length is limited to ***no more than 2 double-spaced pages*** with a reasonable sized font.