**COSC 4570/5010 Data Mining**
**Spring 2020**
**University of Wyoming**

# Homework #4

**Due: April 21, 2020, 11:59 p.m. (No Extension!)**

***Submission guideline*** You need to submit only one *.zip* file. Please name the file as “Your Net id_Homework4.zip”.

## 1. Problems from the book (Introduction to Data Mining 2nd Edition by Tan, Steinbach et al.)

Solve the following:

Chapter 7: Problems 18 and 29.

OR

## Problems from the book (Introduction to Data Mining 1st Edition by Tan, Steinbach et al.)

Solve the following:

Chapter 8: Problems 18 and 29.

## 2. Clustering

Normalized Mutual Information (NMI) is used to evaluate clustering results when the actual clustering of the data (the number of clusters and the clustering assignments) is known beforehand. NMI can be calculated using Equation 1, where $l$ and $h$ are clusters from two different clusterings, $n_h$ and $n_l$ are the number of data points in the clusters $h$ and $l$ respectively, $n_{h,l}$ is the number of data points in cluster $h$ as well as cluster $l$, and $n$ is the size of the dataset.

$$NMI = \frac{\sum_{h,l} n_{h,l} \log \frac{n\, n_{h,l}}{n_h n_l}}{\sqrt{\left(\sum_h n_h \log \frac{n_h}{n}\right)} \sqrt{\left(\sum_l n_l \log \frac{n_l}{n}\right)}} \qquad (1)$$

what are the maximum and minimum values for the NMI? Provide details.

## 3. Community Detection

Download and read the Karate Club network (you can get from the Github repository of [DSCN][1] or from here http://www-personal.umich.edu/~mejn/netdata/karate.zip). The story behind the data set is quite simple: There was a Karate Club that had an administrator “John A” and an instructor “Mr. Hi” (both pseudonyms). Then a conflict arose between them, causing the students (Nodes) to split into two groups. One that followed John and one that followed Mr. Hi.

a) Compute the following statistics describing the datasets:
- number of nodes $n$
- number of edges $m$

Present your results.

b) Preprocess the data.
- store the ground truth i.e. the club each student joined.
  club ‘John A’ members were the students with following ids. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 20, 22 whereas the rest of the students joined the club ‘Mr. Hi’.
- transform the data graph into adjacency matrix.

c) Write a program that computes spectral clustering (try different types of affinity). *Comment your code.* Submit your code. Report the following.
- compare the clustering result with the stored ground truth.
- runtime of the algorithm (HINT: to get this number re-execute your computation 5-10 times and take the mode runtime).

[1] https://github.com/datascienceandcomplexnetworks/book_code/tree/master/Notebook_Chapter_IV_WWW/data