# CS-E4650 Methods of Data mining
# Project work

**Deadline Sunday 6.12. 2020 23:59**

## Overview

The task is to cluster two data sets as well as possible, comparing at least two different clustering methods on both data sets, evaluate the goodness of clusterings with normalized mutual information, and submit optimal clusterings, program code and the project report. The project work can be done either alone or in pairs (i.e., groups of 1–2 course participants).

## Data

Each group (of 1–2 students) gets two numerical data sets to cluster. The first data set is reminiscent to typical gene expression data ("Gene data") and the second one to mass spectrometry data ("MS data"), but they have been artificially generated. The difficulty level has been checked so that all Gene data sets and all MS data sets are equally easy or difficult to cluster and it is possible to find a good clustering with the methods introduced in the course. However, due to individual differences, the clusterings will be different and also the optimal configurations of methods and parameters are likely different.

The data format is the following: The first line is a header listing column labels and then begins the data matrix, where samples are on rows and features on columns. The first two columns are special: the first is sample identifier and the second class label. The rest are features that should be used in clustering.

In the Gene data, there are five classes and in the MS data, three classes. The classes are used only in the evaluation of the clustering with normalized mutual information. Note that the optimal number of clusters is not necessarily the same as the number of classes.

See below how to load data from A+.

## Task

The task is to perform the entire clustering process on the data sets and choose a configuration of methods that produces as good clustering as possible. This includes selection of distance or similarity measures, possible

dimension reduction or feature selection, clustering methods, numbers of clusters and other parameters required by the method. Each group should try at least two clustering methods but preferably more.

The goodness of clustering is evaluated using **normalized mutual information** (NMI), the version by Strehl and Ghosh (2003), with **geometric mean in the denominator**. Given clustering $C_1, \ldots, C_k$ and classification $D_1, \ldots, D_q$,

$$NMI = \frac{I(C, D)}{\sqrt{H(C)H(D)}}$$

where $I = \sum_{C_i \in C} \sum_{D_j \in D} P(C_i, D_j) \log \frac{P(C_i, D_j)}{P(C_i)P(D_j)}$ is mutual information and $H(C) = \sum_{C_i \in C} P(C_i) \log P(C_i)$ and $H(D) = \sum_{D_i \in D} P(D_i) \log P(D_i)$ are entropies.

In the implementation, you can use Python, C, C++, Java, Matlab or Scala and available libraries (describe in the report).

## Solutions

Your solution should contain the following things:

1. Optimal clusterings for both data sets. The format is a text file where line $i$ corresponds the $i$th sample and contains only the cluster number (non-negative integer).

2. Codes of programs that yielded optimal clusterings. You can either return a separate program for both data sets or if you made a generic program give precisely the program parameters that yielded the optimal clusterings for sets. Return codes in a zip package called code.zip.

3. Report (max 3 pages + possible appendices, pdf), where you describe what you did in the data analysis:

    - Methods: All transformations, feature selection or dimension reduction, tested distance or similarity measures, tested clustering methods and parameter settings and how you determined optimal parameters. If you used visual inspection, include most important diagrams, using appendices if needed.
    - Results: Report the NMI values for all tested configurations (e.g., as a result table). Remember to include all parameter values that you used. Tell your conclusions what worked well and what failed and try to hypothesize the reasons for success or failure. If you used internal validation measures, describe their usefulness.

- Brief instructions how to run your program, including required libraries or installations.

## Loading data and submitting results

1. Register your project work group (yourself and possible partner) in MyCourses, using a questionnaire under "Project work" `https://mycourses.aalto.fi/course/view.php?id=28201&section=7`. You need to use the same group in A+.

2. The data sets and submission system are available in Aalto's A+ system. Register by enrolling to the MDM course at `plus.cs.aalto.fi/cs-e4650/2020`.

3. If you work in a group of two students, you must do the following (skip this step if you work alone):

   (a) Choose "Form a group" on the left panel and follow the instructions specified there.

   (b) Select on the top panel that you would like to submit with your partner and not alone.

   (c) Refresh the system (Hit F5 for example).

4. Go to page "Project work" and save the input data files through the Google Drive links

5. Submit your clustering solutions on the same A+ page ("Project work") and the report and code package in MyCourses (when the submission opens).

**Important notes**:

- Once you have decided how you would like to participate (alone or in a group) you should always submit accordingly, i.e., if you have already submitted alone, you should not submit in a group anymore. (Otherwise you might not be able to submit in future at all.) So if you are not sure yet whether you are going to work in a group or not, do not submit anything until you have decided.

- **Don't change your group when you have once registered**, even if the A+ interface may allow you to do it, because the checking system is designed only for one (original) pair of input data per student/group. Therefore, **only the submissions of initially registered groups**

**are evaluated**. For example, if A and B stated that they will work together, A cannot decide later that s/he would like to work with C instead. In exceptional circumstances, where your team mate has dropped the course and you have possibly found a new partner, please **inform us immediately before doing anything!** (This is to save both your and our work, since the system doesn't allow changing between group and individual submissions or changes of groups during one project without losing track what data and clustering should be evaluated.)

- You may submit your custering solution in A+ as many times as you want (maximum 1000), only the last will count though.
- You cannot submit your solutions yet. We will announce when the submission is opened.

## Evaluation

The maximum points are 35p (or scaled to 35% of the grade), where **clustering of Gene data yields max 10p, clustering of MS data max 15p and report max 10p**. The clustering performance is evaluated by comparing NMI of your clustering against optimal NMI (= target-NMI) achieved by Peter's program (using techniques introduced in the course) on your data.

In the Gene data, you will get 1p, if your NMI score is $\geq 80\%$ of the target-NMI, $\geq 82\%$ gives 2p, ..., $\geq 98\%$ gives 10p (i.e., 1p for each 2% increase).

In the MS data, you will get 1p, if your NMI score is $\geq 55\%$ of the target-NMI, $\geq 58\%$ gives 2p, ..., $\geq 97\%$ gives 15p (i.e., 1p for each 3% increase).

If you get better results than Peter's program you will get extra points (total 11p in Gene and 16p in MS).