CS-E4650 Methods of Data mining
Home assignments 4

Deadline Sun 28.11.2021 23:30. Maximum 70 points + 6 extra points.
Submit your solutions early, since MyCourses can get stuck, if many people submit simultaneously. If your solutions are in before 24:00, there is no penalty, but after that **-10% penalty**. The ultimate deadline (with -10% penalty) is Tue 30.11. 23:30.

For each task, write a **pdf report** where you present everything **clearly and compactly**. The idea is that your solution is understandable based on the report alone, including all middle steps. **Note that notebooks are not accepted as reports.** By default, the source code is not studied or executed to see if you had actually done something correctly, if it is missing or incorrect in the report. **On this round, you can get +2p bonus in each task if your report is clear**. There is an optional latex template with examples on the course page.

Remember to put **source code** in a **zip** package that is attached (you can include the pdf also in zip, it doesn't harm, but **remember to submit it also separately**). The main reason for submitting the source code is checking that you have done the work independently and using allowed functions (if asked to implement something on your own, without library functions). The submission cannot be graded if it is incomplete.

Include your name and student number in the report and code header.

---

1. **(max 21p)** In this task, you should experiment with the Gephi tool `https://gephi.org/users/download/` and analyze a social network among school children. Install Gephi, open it and install one extra plugin ('tools → plugins'), Newman-Girvan clustering. You can also install all updates ("check for updates").

   Data `schoolclass5day1.csv` contains a matrix presentation of the interaction graph among all class 5 pupils during one day. The edge weight reflects the strength of interaction (total duration). In file `schoolclass5meta.txt` you can find some background information on pupils: the class (5A or 5B) and gender (F or M).

   Analyze `schoolclass5day1.csv` with Gephi. When you open data, set graph type as 'undirected'. You can find instructions in Gephi guide `https://gephi.org/users/quick-start/` but the appearance has changed a bit in the newest version, so check also `https://github.com/gephi/gephi/issues/1258`. The functions are available under

'statistics' and 'data table' shows values of nodes for all calculated measures. After running a function, you can also visualize results under 'appearance → ranking'.

a) (6p) Identify the most central and influential nodes with the following measures. Report at least two nodes with each measure (more if many nodes with the same value) and explain what the high value means.
   i) Node degree (select network overview – average degree)
   ii) Weighted degree (network overview – average weighted degree)
   iii) Closeness centrality (network overview – network diameter)
   iv) Betweenness centrality (already calculated in iv)
   v) Why the top-two nodes with highest weighted degree are not among top degree nodes? (Hint: look at their edge weights.)
   vi) Which are the most critical nodes for the information flow?

b) (10p) Identify dense subgraphs (communities) with the Modularity function (select network overview – modularity). What kind of communities do you find? Try to explain the communities with background variables (class and gender)!

c) (5p) Make some small experiment of your own choice with gephi and report the results! Explain briefly what you analyzed. You can e.g., visualize some aspects of the network, calculate more measures or test other community detection methods.

2. **(max 24p)** Let us consider the task of analyzing and mining subgraph patterns. Figure 1 shows an example of six molecular graph structures. The node labels correspond to atoms (carbon, oxygen or nitrogen)[1]. Three of the molecules, graphs G_1 (serotonin), G_3 (dopamine) and G_6 (melatonin) belong to **class** M (monoamines), while G_2 (acetaminophen = paracetamol), G_4 (ibuprofen) and G_5 (caffeine) belong to class $\neg M$.

   a) (14p) Determine nearest neighbours of class M molecules using two MCG-based[2] distances: i) Udist and ii) Mdist (Equations 17.2 and 17.3 in Aggarwal's book). It suffices to identify two nearest neighbours to each class M graph, unless there are many equally

[1]For simplicity hydrogen atoms and double bonds between atoms are not presented.
[2]MCG = Maximum common graph

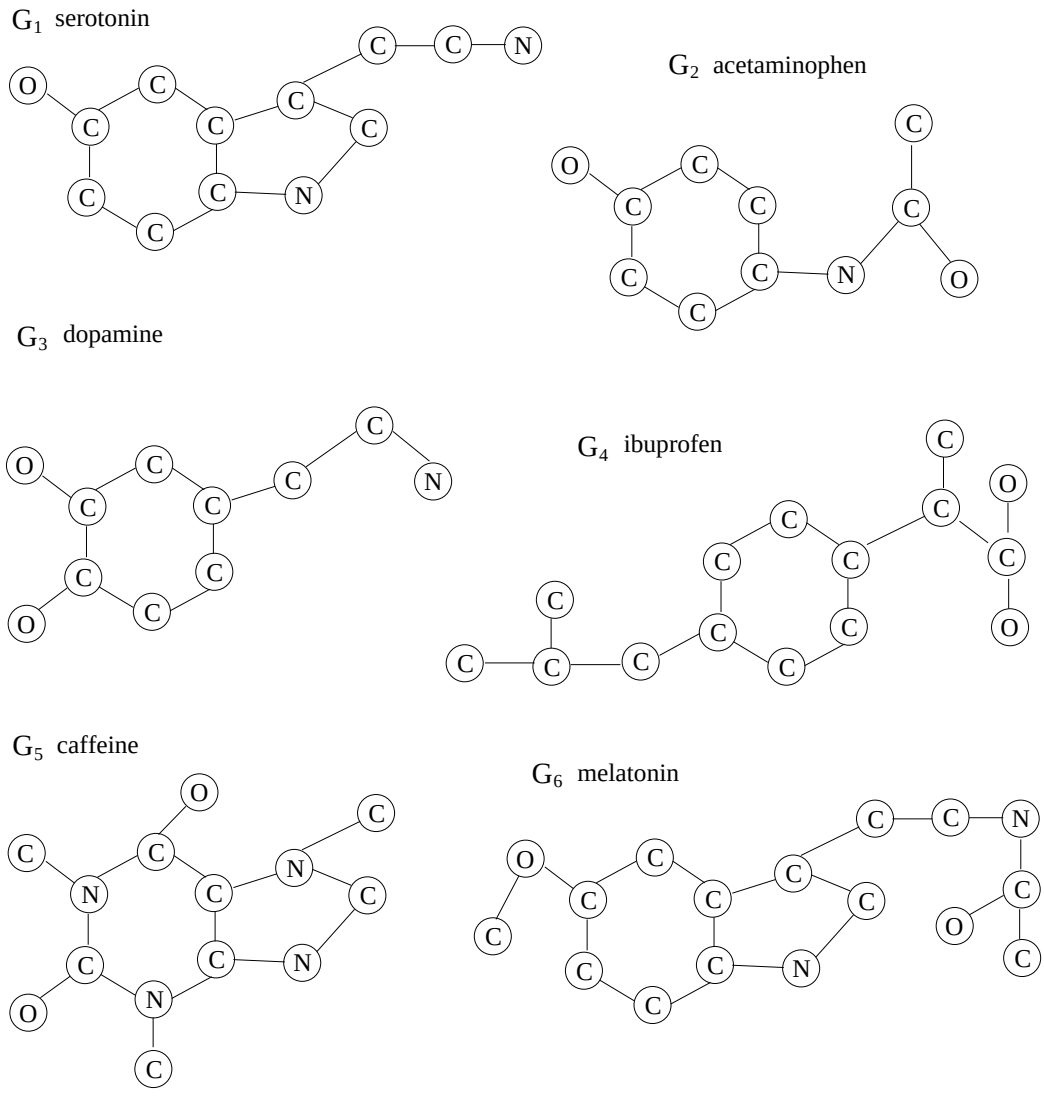

Figure 1: Six graphs corresponding to molecule structures.

close neighbours. Which distance separates class M molecules better from other molecules?

b) (5p) Identify maximum common subgraph G for all class M molecules. You can draw it by hand if you want. How well does it predict class M? I.e., calculate precision ("confidence") of graph–class rules $G \to M$ and $\neg G \to \neg M$, where rule condition G means that subgraph G occurs in the graph and $\neg G$ means that it doesn't.

d) (5p) Assume that you were given a database of compounds with

molecular structure graphs and attributes of compounds like toxic, drug, amino-acid etc. Describe a general method how you could find most significant statistical associations between subgraphs and attributes using GraphApriori and postprocessing. Is this a good approach? Consider such aspects as the effect of minimum frequency threshold, problems of missed or redundant patterns and computational efficiency.

3. **(max 25p)** Table 1 presents movie ratings by 6 users on 6 movies. The latex source of the table is available on the course page (mratingstable.tex). The ratings are between 1 (didn't like at all) to 5 (fantastic movie) and 0 means a missing rating (the user hasn't watched the movie). The users are notated $u1, \ldots, u6$ and movies $m1, \ldots, m6$. The task is to apply recommender systems for rating prediction using neighbourhood-based collaborative filtering (Aggarwal 18.5.2 and an example on lecture 10).

   a) (2p) Calculate mean ratings per user. Use all non-missing ratings in the calculation. These are needed in parts b) and c).

   b) (7p) Calculate required pairwise similarities between users[3] using a modified Pearson correlation r ("Pearson" in Aggarwal Equation 18.2). Use the mean values calculated in part a. Remember that the correlation is calculated only over co-rated movies.

   c) (11p) Predict missing ratings using two nearest neighbours ($K = 2$) and an extra requirement that the similarity is $r \geq 0.5$. Tell if the movie is recommended to the user (if the user would like it more than average).

      Report if some prediction cannot be made (not enough sufficiently similar neighbours with required ratings).

   d) (5p) Consider the item-based way of predicting the missing ratings of movies $m3$ and $m4$ with adjusted cosine similarity, as suggested in Aggarwal 18.5.2.2. Why it is not a good solution here? Suggest an alternative item-based solution that could be used instead (no need to calculate the actual predictions).

[3]Note: similarity between $u2$ and $u3$ is not needed, so 14 similarities.

Table 1: Movie ratings (scale 1–5) by 6 users ($u1$–$u6$) on 6 movies ($m1$–$m6$). Special value 0 means a missing rating.

	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
u1	3	1	2	2	0	2
u2	4	2	3	3	4	2
u3	4	1	3	3	2	5
u4	0	3	4	4	5	0
u5	2	5	5	0	3	3
u6	1	4	0	5	0	0