KTH Information and Communication Technology

Exam in ID2222 Data Mining

2018-01-09, 08:00-13:00, rooms 205, 208, 301, 308, Electrum, Kista

This exam is "closed book" and you are not allowed to bring any material or equipment (such as laptops, PDAs, or mobile phones) with you. The only exceptions are dictionaries.

Instructions

- The problems are not ordered by level of difficulty.
- Answer each problem on a separate page. Write only on one side of a sheet.
- Each page must be provided with your name, your personal number, a problem number, and the page number.
- Always motivate your answer. Answers without any motivation will in general not be considered.
- Greater marks will be given for answers that are rather short and concrete than for answers written in a sketchy and rambling fashion. Information irrelevant to a question will not be considered.
- Several contradictory answers to the same question might lead to that the entire answer is disregarded even though one of the alternatives is correct.
- Small syntactical errors in program outlines (if any) will not affect your result if they do not change control flow so that the program becomes semantically incorrect.

Grading

The passing grade (E) is 60 points out of total points of 100 points.

Grade F (fail):	< 55
Grade FX (fail; eligible for completion[1]):	55-59
Grade E:	60-68
Grade D:	69-76
Grade C:	77-84
Grade B:	85-92
Grade A:	> 92

GOOD LUCK!

Vladimir Vlassov and Sarunas Girdzijauskas, ID2222 lecturers and examiners

[1] The completion can be performed as an extra individual task within one month after the exam results are reported. Contact Vladimir Vlassov if you are eligible for the completion.

I. Finding Similar Items

Describe the following three techniques used together to find textually similar documents: shingling, minhashing, and Locality Sensitive Hashing (LSH). Justify the motivation for using minhashing. Justify the motivation for using LSH. (8 p)

II. Frequent Itemsets and Association Rules

Briefly describe the A-priory algorithm to find frequent item sets. What is *support* of an item set? What is an association rule between two item sets? What is *confidence* of an association rule? (8 p)

III. Mining Data Streams

a) What is an ad-hoc query on a data stream? Give an example of an ad-hoc query and describe the state you need to store in order to answer the query in the query.
What is a standing query on a data stream? Give an example of a standing query and describe the state you need to store in order to answer the query. (6 p)

b) What is sliding window? What is exponentially decaying window? Give examples. (4 p)

c) Describe an algorithm to retrieve a fixed-size sample from a stream of elements where each of the elements can be sampled with the same probability. (4 p)

IV. Link Analysis

For each of the following statements indicate whether it's correct or wrong. (14 p)

1. It is possible to determine pageRank of the nodes of any directed graph G by looking only into the distribution of the node degrees of G.
2. It is possible to determine pageRank of the nodes of any undirected, connected non-bipartite graph G by looking only into the distribution of the node degrees of G.
3. High teleportation probability increases the accuracy of PageRank,
4. The convergence speed of pagerank does not depend on teleportation probability.
5. Hubs and Authorities ranking algorithm might not converge without teleportation.
6. Spam Farm effectiveness does not depend on teleportation probability.
7. None of the above statements is correct.

V. Mining Social-Network Graphs. Spectral Analysis

For each of the following statements indicate whether it's correct or wrong. (10 p)

1. In spectral analysis, the existence of large eigengap between the smallest and the second smallest eigenvalue of laplacian matrix L (which represents graph G) indicates the existence of many clusters in G.
2. In order to detect communities of graph G with as small as possible ratio cut (expansion) using spectral clustering, one needs to extract the eigenvectors corresponding to the largest eigenvalues of adjacency matrix A (which represents graph G).
3. It is possible to determine the number of disconnected components in graph G, by looking into the dominant eigenvector of adjacency matrix A (which represents graph G).
4. The list of all eigenvalues of of adjacency matrix A (which represents graph G) can provide us with information on which nodes belong to the largest cluster of G.
5. None of the above statements is correct.

VI. Sampling with Random Walks

Assume that the world consists of only two countries, A and B. It is known that every citizen in A has exactly d friends, and every citizen in B has exactly $2d$ friends. You perform a random walk on the friendship graph (connected, undirected) of this world; 1/3 of visited nodes come from B. What is your best guess about $|B|/N$, i.e., the fraction of population that lives in B? (8 p)

VII. Clustering

For each of the following statements indicate whether it's correct or wrong. (10 p)

1. Hierarchical clustering has $O(N \log N)$ complexity where N is number of data-points.
2. Number of iterations for k-means always converges in $O(\log N)$ number of iterations where N is number of data-points.
3. The output of k-means does not depend on the way initial k points are picked
4. The overall complexity of k-means algorithm is $O(kN)$ for N points and k clusters
5. The quality of clusters produced by BFR algorithm does not depend on the initial sampling.
6. The quality of clusters produced by BigClam algorithm does not depend on initial initialisation of matrix F.
7. BFR algorithm can not deal with clusters of arbitrary shapes
8. CURE algorithm automatically detects the number of clusters in the dataset.
9. A datapoint can be assigned to several communities by BigClam algorithm.
10. In some cases k-means can output less than k clusters.

VIII. Dimensionality reduction

a) SVD decomposition of matrix M is $M=USV^T$. Select correct answer(s):

1. The columns of U are eigenvectors of M
2. S is a dense matrix
3. Dot product of any two columns from U is equal to 1.
4. The columns of V are eigenvectors of M^TM
5. None of the above statements is correct. (10 p)

b) You have a matrix M representing ratings given by users to the products.

	Product 1	Product 2	Product 3	Product 4	Product 5
Alice					
Bob					
Carol					
David					
Erin					
Frank					
Grace					

The matrix was populated by user ratings (each cell was assigned a value from 0 to 5) and you were given a task to identify main concepts that describe Matrix M by reducing the dimensionality of M. After doing SVD on M you decided to reduce dimensionality to *d* dimensions.

The SVD decomposition of $M=USV^T$ provided you with these matrices:

U=

-0,55	-0,03	0,18	0,48	0,66	0	0
-0,02	0,65	0,45	-0,22	0,05	0,05	-0,56
-0,49	-0,02	0,04	-0,04	-0,39	-0,78	-0,07
-0,61	-0,03	0,05	-0,05	-0,49	0,62	0,06
-0,02	0,71	-0,55	0,37	-0,1	-0,02	0,23
-0,31	0,05	-0,48	-0,71	0,41	0	0
-0,01	0,26	0,48	-0,26	0,06	-0,07	0,79

S=

11,63	0	0	0	0
0	8,9	0	0	0
0	0	1,52	0	0
0	0	0	1,07	0
0	0	0	0	0,34

V^T=

-0,01	-0,7	-0,72	-0,04	-0,01
0,57	-0,02	-0,03	0,57	0,59
-0,23	-0,21	0,23	-0,54	0,75
0,5	-0,57	0,56	-0,17	-0,3
-0,62	-0,39	0,35	0,59	0,02

Answer the following questions. Explain your answers.

a. With how many main concepts (dimensions) would you identify matrix M and why? (2 p)

b. Which user and which product belong to which concept (after dimensionality reduction). Explain why. (2 p)

c. Is Carol's preferences are more similar to David's or to Erin's? (4 p)

d. What is the rank of Matrix M. Explain. (4 p)

IX. **Recommender Systems**

Consider an online shop selling movies that has a database containing information about movies: title, genre, release decade, and country. It also has information about which users have seen each movie. The rating for a user on a movie is either 0 (dislike/have not seen) or 1 (like).
A summary of the database is as follows.

Title	Genre	Release Decade	Country	Total number of ratings
A	Comedy	2000s	Sweden	50
B	Drama	2010s	USA	400
C	Thriller	2010s	Norway	20
D	Thriller	1980s	USA	0
E	Drama	1990s	Sweden	70
F	Comedy	2010s	Norway	1

Assume *Alice* is interested in comedies of the 2000s filmed in USA. Some existing recommender system *R* has recommended the movie *B* to *Alice*. The recommender system *R* could be one or more of the following options.

1) User-user collaborative filtering;
2) Item-item collaborative filtering;
3) Content-based recommender system.

Given the above dataset, which one(s) do you think *R* could be? (If more than one option is possible, you need to state them all.) Explain your answer. (6 p)

------------------------------ End of the Exam --