KTH Information and Communication Technology

Exam in ID2222 Data Mining

2020-01-08, 08:00-13:00, rooms 204, 205, 208, 301, 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.
- If not indicated otherwise (e.g., for True/False questions), 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

a) Briefly describe each of the three techniques: shingling, minhashing, and Locality Sensitive Hashing (LSH), used together to find textually similar documents. How to estimate a similarity of documents represented as shingle sets? How to estimate similarly of documents represented as signature vectors? Justify the use of minhashing and LSH. (8 p)

b) Does it mean that two documents are always identical if the similarity of their 3-shingle sets is 1? If so, prove it. If not, give a counterexample. (2 p)

c) What is a complexity of shingling and computing Jaccard similarity of two documents of O(n) characters each using k-shingles? Assume that comparison of two k-shingles takes O(k). (2 p)

II. Frequent Itemsets and Association Rules

Briefly describe the A-priory algorithm to find frequent itemsets. What is *support* of an itemset? What is an association rule between two itemsets? What is *confidence* of an association rule? (6 p)

III. Mining Data Streams

a) Explain how you can retrieve (i) a fixed-size sample from an unbounded stream of elements where each element has the same probability of being sampled;
(ii) a 5% representative sample of an unbounded stream of key-value pairs. How can you limit the sample size as the sample grows? (4 p)

b) 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. (4 p)

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

IV. Link Analysis. Page Rank. Hubs and Authorities

a) Select correct answer from the following options: Clustering Coefficient of node u is (2 p)

1. is k/N (where k is a node degree of u and N is the size of the network).
2. is a probability that any two neighbors of u are connected.
3. is k/(0.5*N*(N-1)) (where k is a node degree of u and N is the size of the network).
4. is a probability that there exists a shortest path between any two nodes through node u.

b) For each of the following statements indicate whether it's correct or wrong. (15 p = 15 x 1 p)

1. Bi-partite graphs can have many triangles.
2. Eigenvector Centrality takes global topology into account.
3. Watts-Strogatz Small-World network with E number of edges and Erdos-Renyi network with E number of edges will exhibit similar network diameter.
4. Watts-Strogatz Small-World network with E number of edges and Erdos-Renyi network with E number of edges will exhibit similar clustering coefficient.
5. Watts-Strogatz Small-World network with E number of edges and Network produced by preferential attachment with E number of edges will exhibit similar degree distribution.
6. Erdos-Renyi network has several hub nodes with very large degrees compared to other nodes.
7. A network produced by a preferential attachment model is a very good expander.
8. Random walk will always converge on a directed graph.

9. One can measure the distance between nodes using Hubs and Authorities algorithm.
10. Hubs and Authorities do not suffer from spider traps and dead ends.
11. One can measure the distance between nodes using PageRank with restarts.
12. PageRank with teleportation probability 0.001 can not converge on a weakly connected graph.
13. PageRank with restarts teleports random walker to a node with highest degree.
14. Link Farms are more effective where PageRank teleportation probability is lower.
15. Trust Rank is a variant of topic-specific page rank.

V. Sampling with Random Walks

A student performed a random walk on a connected undirected friendship graph from an international social network and collected the following statistics.

100 Swedish nodes had degree 50

50 Norwegian nodes had degree 50

150 German nodes had degree 50

35 Norwegian nodes had degree 70

210 German nodes had degree 70

30 Swedish nodes had degree 100

50 Norwegian nodes had degree 100

90 German nodes had degree 100

70 Swedish nodes had degree 140

35 German nodes had degree 140

10 German nodes had degree 200

What is your estimate on the fraction of Norwegians on the graph? (4 p)

VI. Spectral Graph Analysis

a) Assume you got access to an undirected non-bipartite graph G that you analyze by performing spectral analysis, i.e., extracting eigenvalues and associated eigenvectors of the adjacency matrix A representing graph G, as well as extracting eigenvalues and associated eigenvectors of the Laplacian matrix L representing graph G and extracting eigenvalues and associated eigenvectors of normalized Laplacian matrix NL ($NL= D^{-½}AD^{-½}$, where D is a degree matrix, and A is the adjacency matrix of G).

1. You want to bisect graph G into two clusters by optimizing on conductance measure. Can you do it be looking at the eigenvector associated with the second smallest eigenvalue of matrix NL? Explain in max 3 sentences. (3 p)
2. You want to bisect graph G into two clusters by optimizing on ratio-cut measure. Can you do it be looking at all the eigenvalues of matrix L? Explain in max 3 sentences. (3 p)
3. You want to inspect if graph G has disconnected components. Can you do it by looking at all the eigenvector associated with the second smallest eigenvalue of matrix L? Explain in max 3 sentences. (3 p)

b) You construct graph G2 by applying Configuration Model based on the degree sequence of G. You also perform spectral analysis of graph G2 by extracting eigenvalues and associated eigenvectors of the Laplacian matrix L2 representing graph G2.
   1. Which graph has larger eigengap, G or G2? Explain in max 3 sentences. (3 p)
   2. You learn that the second smallest eigenvalue of the Laplacian matrix of G is not a zero. Will random-walk converge to a unique stationary distribution on G? Explain in max 3 sentences. (3 p)

VII. **Dimensionality reduction**

a) You have performed CUR decomposition of a sparse matrix M into (M=CUR) and SVD decomposition of the same matrix M ($M=USV^T$) such that both middle matrices (U from CUR and S from SVD) have the same size d x d, where d << #rows and d << #columns of M. Answer the following questions (max 3 sentence explanation each).
   1. Will CUR decomposition be able to more accurately reconstruct the initial matrix M than SVD? Yes/No/Same? (3 p)
   2. Which decomposition would be more space efficient to store decomposed matrices: CUR or SVD? (3 p)
   3. You are asked to interpret the low dimensional latent space of decomposed matrix M in both CUR and SVD cases. For which case it will be easier to provide the interpretation CUR or SVD? (3 p)

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

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	Movie 6	Movie 7	Movie 8
Alice								
Bob								
Carol								
David								
Erin								
Frank								

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

U=

-0.022	0.005	-0.706	-0.028	-0.027	0.707
-0.046	-0.704	-0.004	0.509	-0.494	0
-0.7	0.069	0.061	-0.476	-0.524	0
-0.022	0.005	-0.706	-0.028	-0.027	-0.707
-0.093	-0.704	0	-0.505	0.491	0
-0.706	0.07	-0.017	0.507	0.489	0

S=

12.299	0	0	0	0	0
0	10	0	0	0	0
0	0	7.353	0	0	0
0	0	0	0.586	0	0
0	0	0	0	0.566	0
0	0	0	0	0	0

V=

-0.572	0.069	0.03	0.269	-0.309	0.705
-0.572	0.069	0.03	0.269	-0.309	-0.705
-0.579	-0.001	0.03	-0.594	0.558	0
-0.011	0.003	-0.576	-0.291	-0.29	0.055
-0.011	0.003	-0.576	-0.291	-0.29	-0.055
-0.068	0.01	-0.578	0.576	0.574	0
-0.056	-0.704	-0.003	0.03	-0.028	0
-0.056	-0.704	-0.003	0.03	-0.028	0

Answer the following questions. Give max 3 sentence explanations for every question. (18 p = 6 x 3p)

1. With how many main concepts (dimensions) would you identify matrix M?
2. What is the rank of M?
3. Which user belongs to which concept (after dimensionality reduction)?
4. Which movie belongs to which concept (after dimensionality reduction)?
5. Is Move 6 more similar to Movie 4 or Movie 8? Why?
6. You were told that some users ranked the movies in exactly the same way (gave the same scores). Identify these users.

VIII. Latent Factor Recommender Systems

Student got access to a sparse matrix R representing user rankings of products. The student got a task to build a latent factor recommender system and decomposed the matrix R using gradient descent with regularization into two matrices Q and P ($R \approx Q \cdot P^T$), where Q represents mapping of products to the d-dimensional latent space and P represents mapping of users to d-dimensional latent space.
For each of the following statements indicate whether it's correct or wrong. (7 p = 7 x 1 p)

1. Columns in P are orthonormal
2. Rows of Q are orthonormal.
3. Matrices Q and P are sparse (since R was sparse)
4. Row *i* in P represents the coordinates of a user *i* in the latent space
5. Columns of P are eigenvectors of matrix $R*R^T$
6. If *d* is very large, then it is not needed to perform regularization while doing gradient descent.
7. The student can predict the ranking of product *i* by user *j* by performing a dot product of row *i* of Q with row *j* of P.

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