KTH Information and
Communication Technology

Exam in ID2222 Data Mining

2018-04-04, 14:00-19:00, room 205, 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 (phone: 08-7904115) and Sarunas Girdzijauskas (phone: 08-7904175)
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

Briefly describe each of the three techniques: shingling, minhashing, and Locality Sensitive Hashing (LSH), used together to find textually similar documents. (6 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 and how to find it? What is *confidence* of an association rule? (6 p)

III. Mining Data Streams

a) Suppose that you are designing a web crawler that indexes web pages. Given a starting page, the crawler extracts all the links from the page and adds them to a queue. Then, it picks the pages from the queue one-by-one and repeats the process. Upon discovering a new link the crawler checks if the link has been already crawled; if so, the duplicate link is not considered.

i). Suppose that you cannot afford to store all visited links in memory. Can you use an approximate stream technique to check if a web page has been visited before? Describe the technique and how you can apply it to your web crawler. (6 p)

ii). Suppose that you are building an additional service on top of your web crawler that estimates the PageRank of different pages. You can get a good enough estimation by counting how many different pages link to each page. How can you get these counts estimates for each page provided that you have limited memory? (6 p)

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

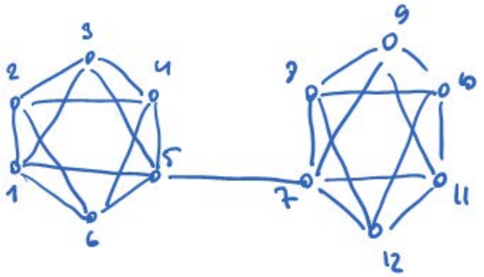
IV. Link Analysis

a) Explain why PageRank algorithm has teleportation (max 4 sentences). (2 p)

b) What happens if teleportation probability in PageRank is set to 1. Explain the consequences to the ranking (max 4 sentences). (2 p)

c) Does Hubs and Authorities algorithm require teleportation? If “yes” explain why. If “not” explain why (max 4 sentences). (4 p)

d) Assume you want to determine distances from Node 5 to all the other nodes in the graph below, using the *SimRank* approach. Answer the following questions. (no need for explicit calculation)



(i) Compare the distance from Node 5 to Node 1 with the distance from Node 5 to Node 7 (the same/smaller/larger). Explain. (4 p)

(ii) Compare the distance from Node 5 to Node 2 with the distance from Node 5 to Node 8 (the same/smaller/larger). Explain. (4 p)

V. Mining Social-Network Graphs. Spectral Analysis

Assume you got access to 3 undirected graphs G1, G2, and G3 that you analyze by performing spectral analysis, i.e., extracting eigenvalues and associated eigenvectors of each Laplacian matrix representing these graphs. In the end you arrive at the following facts about each of the graphs.

- **G1** has N=2000 nodes with average degree of 100. You order the eigenvalues of the Laplacian matrix of G1 in ascending order and notice that λ4 = 0 and λ5 = 97.1
- **G2** has N=2000 nodes with average degree of 100. You order the eigenvalues of the Laplacian matrix of G2 in ascending order and notice that the second eigenvalue λ2 = 102.9
- **G3** has *N*=2000 nodes with average degree of 100. You order the eigenvalues of the Laplacian matrix of G3 in ascending order and notice that the second eigenvalue $\lambda_2 = 5.7$

Assume you have also created your own undirected graph G4, where you know the model and the parameters used to generate that graph.

- **G4** you generate by the basic preferential-attachment model (adding one edge at each round) until you reach *N*=2000 nodes.

Answer the following questions. Explain your answers.

a) Does random walk converge to a unique stationary distribution on G1, G2, and G3? (2 p)

b) Order G2 and G3 based on the convergence speed of random walk (from higher to lower). (2 p)

c) From the set of graphs {G1, G2, G3, G4} select two with the worst expansion properties. (2 p)

d) Which of the two graphs G2 and G3 is expected to have large diameter? (4 p)

e) You analyze a new graph G5 with 1000 nodes and average degree of 30. You order the eigenvalues of the Laplacian matrix of G5 in ascending order and notice that
$\lambda_2 = 0.915$, $\lambda_3 = 0.968$, $\lambda_4 = 22.8$, $\lambda_5 = 24.1$, $\lambda_6 = 38.8$, $\lambda_7 = 46.2$.
How many communities are there in G5? (6 p)

f) Can you tell what are the sizes of the communities from the Laplacian eigenvalues of G5?
If yes, give an approximate number (within 5%) and explain how you arrived there.
If not, explain what you need to figure it out? (6 p)

VI. Sampling with Random Walks

Assume that the world consists of only two countries, A and B, where on average every citizen in country A has "dA" friends and in country B has "dB" friends. You perform a random walk on the friendship graph (connected, undirected) of this world; you have observed that 1/3 of visited nodes come from B. What is the ratio dB/dA if:

a) you know that countries A and B are the same in size (i.e., have the same number of citizens)? (3 p)

b) you know that country A is 3 times larger than country B? (3 p)

VII. Clustering

a) Assume that we have a data set of 900 points in 2D space that consists of three clear clusters A, B and C.

The cluster A is centered on (0,0), with 300 points uniformly distributed in a circle of radius 3.
The cluster B is centered on (0,30), with 300 points uniformly distributed in a circle of radius 3.
The cluster C is centered on (30,0), with 300 points uniformly distributed in a circle of radius 3.

We want to do *k*-means clustering on the data, and we choose *k* = 3. Three random existing data points are chosen and their coordinates are set for the initial centroids *x*, *y*, and *z*. The result after k-means converges is going to be three apparent clusters, which may or may not coincide with the true clusters A, B, and C. We say that one of the true clusters is correct if there is an apparent cluster that consists of all and only the points in that true cluster.

Is it possible that k-means will converge after **only one round** and will identify A, B and C as correct clusters? If "yes" give the probability of that happening. If "not", explain why. (4 p)

b) What is the reason in CURE clustering algorithm to move each of the representative points a fixed fraction of the distance between its location and the centroid of its cluster? Give max three-sentence explanation. (4 p)

c) Can BFR algorithm deal with clusters of arbitrary shapes? Give max three-sentence explanation. (4 p)

VIII. Dimensionality Reduction and Recommender Systems

a) SVD decomposition of matrix M is $M=USV^T$. Explain the relation of the matrix M^TM and MM^T with U, V and S. (4 p)

b) CUR decomposition of matrix M is M=CUR. Assuming M is a sparse matrix, will R and U be sparse matrices? Explain (4 p)

c) An online bookstore has three recommender systems R1, R2 and R3 which are Content-based, User-User Collaborating filtering based, Item-Item Collaborative filtering based respectively. Which of the recommender systems (R1, R2 or R3) will perform the best if the bookstore has very small number of ratings compared to the number of users and books? Explain (4 p)

d) Assume a **completely new** user A wants to buy a book in the above described bookstore (the shop does not have any data on A). Could any instance of R recommend a book to A? Explain. (4 p)

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