KTH Information and Communication Technology

Exam in ID2222 Data Mining
2017-01-12, 14:00-19:00, rooms 205, Sal B, 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

Briefly describe the stages and corresponding techniques of finding textually similar documents based on Jaccard similarity using the 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 and explain the A-priory algorithm and its pipeline used to find frequent item sets and association rules. What is *support* of an item set and how to estimate it? What is *confidence* of an association rules and how to estimate it? (8 p)

III. Mining Data Streams

a) Compare the Stream Data Management System (SDMS) with the Database Management System (DMS). (4 p)

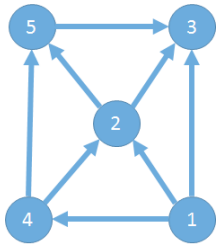
b) What is the difference between an ad-hoc query and a standing query? Give an example of each query type. Describe the state you need to store in order to answer the following queries for a stream of temperatures: (i) What is the maximum temperature ever recorded? (ii) What is the average temperature ever recorded? (iii) What is the average temperature of the last hour? (6 p)

c) 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 elements that are key-value pairs. How can you limit the sample size as the sample grows? (6 p)

d) Define what is sliding window and what is exponentially decaying window. Give a usage example for each window type. (6 p)

IV. Link Analysis

Consider the following graph, and answer the following questions.



a) Write the PageRank equations for the nodes of the above graph, assuming ß = 0.8 (teleportation probability = 1-ß). Denote the PageRank of node x by $r(x)$. Take into account effects of the dead-ends. (4 p)

b) Write the Hubs and Authorities equations for the above graph. Denote the hub score of a node x by $h(x)$ and the authority score by $a(x)$. (4 p)

c) Identify top 3 highest scoring nodes in PageRank (no need to explicitly compute scores).
Identify the node scoring highest in authority rank.
Identify the node scoring highest in hub rank. (6 p)

d) Explain why in Hubs and Authorities we do not need to use teleportation as in PageRank. (4 p)

V. Mining Social-Network Graphs. Spectral Analysis

Assume you got access to 4 undirected graphs G1, G2, G3 and G4 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=1000 nodes with average degree of 60. You order the eigenvalues of the Laplacian matrix of G1 in ascending order and notice that the second eigenvalue $\lambda_2 = 72.8$.

G2 has N=1000 nodes with average degree of 60. You order the eigenvalues of the Laplacian matrix of G2 in ascending order and notice that the second eigenvalue $\lambda_2 = 4.9$.

G3 has N=1000 nodes with average degree of 60. You order the eigenvalues of the Laplacian matrix of G3 in ascending order and notice that $\lambda_3 = 0$ and $\lambda_4 = 99$.

G4 has N=1000 nodes with average degree of 60. You order the eigenvalues of the Laplacian matrix of G4 in ascending order and notice that the second eigenvalue $\lambda_2 = 19.3$.

Assume you have also created your own undirected graphs G5 and G6 where you know the models and the parameters used to generate those graphs.

G5 you generate by the basic preferential-attachment model (adding one edge at each round) until you reach N=1000 nodes.

G6 you construct to be d-regular undirected connected random graph of N nodes, where N=1000 and d = 60.

Answer the following questions

a) Does random walk converge to a unique stationary distribution on G1, G2, G3, G4, and G6 graphs? Explain for each graph. (2 p)

b) Order G1 and G4 based on the convergence speed of random walk (from higher to lower). Explain your answer. (2 p)

c) From the set of graphs (G1, G3, G5, G6) select two with the worst expansion properties. Explain why. (2 p)

d) Which of the two graphs G2 and G6 is expected to have large diameter? Explain. (4 p)

e) What is the difference between spectral clustering approaches that use the affinity/adjacency matrix A and the Laplacian matrix L. (4 p)

VI. Sampling with Random Walks

a) We perform a random walk on a connected, undirected friendship graph from a social network and acquire the following sample set.

– 4 nodes from Australia with degree 2
– 3 nodes from Japan with degree 6
– 2 nodes from Sweden with degree 4
– 5 nodes from Australia with degree 8
– 4 nodes from Sweden with degree 16
– 5 nodes from Sweden with degree 10
– 1 node from Japan with degree 2
– 5 nodes from Japan with degree 4

– 2 nodes from Australia with degree 16

What is your estimate on the fraction of Swedish people on the graph? (6 p)

VII. Clustering

a) Let us consider a one-dimensional space, where we wish to perform a *hierarchical clustering* of 6 points with the coordinates 1, 2, 7, 8, 10, and 35. Show what happens at each step until there are two clusters, and give these two clusters. Your answer should be a table with a row for each step; the row should contain the members of the new cluster formed, and its centroid. More specifically, if you are merging a cluster C1 = {x, y} of centroid c1 with a cluster C2 = {z, q} of centroid c2, you should report {x,y,z,q} in the table, as well as the new centroid obtained with these 6 points (6 p)

b) Can k-means clustering ever give results that contain more or less than k clusters? Explain. Does the final clustering of k-means depend on the initial k points? Explain. (4 p)

VIII. Recommender Systems

a) Assume you want to design a recommendation system for an online bookshop that has been launched recently. The shop has over 10 million products, but its rating database has only 5000 ratings. Which of the following would be a better recommendation system? (i) User-user collaborative filtering; (ii) Item-item collaborative filtering; (iii) Content-based recommendation. Justify your answer (if possible in one sentence) (6 p)

b) Suppose the shop is using the recommendation system you suggested in a) above. A customer has only rated two products: “Jamie Oliver's Christmas Cookbook” and “200 Slow Cooker Recipes” and both ratings are 5 out of 5 stars. Which of the following books is less likely to be recommended? (i) “Spectral graph theory”; (ii) “Cook Now Eat Later”; (iii) It depends on other users’ ratings. (4 p)

c) After some years, the bookstore has enough ratings that it starts to use a more advanced recommendation system like the one that won the Netflix prize. Suppose the mean rating of books is 3.4 stars. Alice, a faithful customer, has rated 350 books and her average rating is 0.4 stars higher than average users’ ratings. “Animals Farm”, is a book title in the bookstore with 250,000 ratings whose average rating is 0.7 higher than global average. What would be a baseline estimate of Alice’s rating for “Animal Farm”? (4 p)

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