KTH Information and Communication Technology

Exam in ID2222 Data Mining

2019-01-09, 08:00-13:00, rooms 308, Sal A, Sal B, Sal C, Electrum, Kista

I. Finding Similar Items

a) Briefly explain the shingling technique used to represent a document in the form of set. How one can measure similarity of two documents represented as shingle sets? What is a complexity of shingling and computing similarity of two documents of O(n) characters each using k-shingles? Assume that comparison of two k-shingles takes O(k). (6 p)

ANSWER: For the answer, see Lecture 2 "Finding Similar Items: Locality Sensitive Hashing" and Chapter 3 "Finding Similar Items" of the textbook "Mining of massive datasets"

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)

ANSWER: NO. Example: A = abab, B = baba

c) Briefly explain the minhashing technique used in finding textually similar documents. How to estimate similarly of documents represented as signature vectors? (4 p)

ANSWER: For the answer, see Lecture 2 "Finding Similar Items: Locality Sensitive Hashing" and Chapter 3 "Finding Similar Items" of the textbook "Mining of massive datasets"

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)

For the answer, see Lecture 3 "Frequent Itemsets" and Chapter 6 "Frequent Itemsets" of the textbook "Mining of massive datasets"

III. Mining Data Streams

a) What is the difference between an ad-hoc query and a standing query on a stream?
For each of the following aggregate queries on a data stream, indicate whether it is ad-hoc or standing, and describe the state you need to store in order to answer it: (i) Report each new maximum value in the stream? (ii) What is the maximum value seen so far in stream? (iii) What is the average value ever seen in the stream? (iv) What is the average value of the last hour? (6 p)

Answer: Ad-hoc queries are ordinary queries, asked one time about streams. For example, a query about a current value of some aggregate property of the stream observed so far: What is the maximum value seen so far in the stream. In contrast, standing queries are queries that are, in principle, asked about the stream at all time, continuously. For example: Report every new maximum value ever seen in the stream.

The stored states for the queries in the above question are as follows.

(i) Standing. One value for current max updated whenever a new max is observed;

(ii) Ad-hoc. One value for current max updated whenever a new max is observed;

(iii) Standing. One count and one sum;

(iii) Ad-hoc. The elements of the last 1h to compute the rolling average.

b) Describe an algorithm to maintain a fixed-size sample from an unbounded stream of elements where each element can be sampled with the same probability. What is the name of the algorithm? Prove that the algorithm maintains a sample of the stream seen so far where each element is sampled with the same probability, i.e., it doesn't skew the underlying stream's distribution. (6 p)

Answer (not explained here) Reservoir sampling: initially hash to a large number of buckets and start dropping buckets once sample grows. See the proof in Lecture 5 "Mining Data Streams. Part 1" and Chapter 4 "Mining Data Streams" of the textbook "Mining of massive datasets"

IV. Network Models

Assume a master student got access to a graph that consists of 1 billion nodes ($N = 10^9$) with an average degree of 500. She wants to figure out if this is a random (Erdos-Renyi Style) graph, or a clustered small-world graph. She measures average clustering coefficient and average path length of the graph. The results indicate that average clustering coefficient is 0.095 and average path length is O(log$N$). Is the graph Random or Small-World? Explain why in max 3 sentences. (4 p)

V. Link Analysis

For each of the following statements indicate whether it's correct or wrong. (12 = 6 x 2 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. Decreasing teleportation probability decreases the accuracy of PageRank.
3. By increasing teleportation probability one can speed up the convergence time of PageRank.
4. It is guaranteed that PageRank will always converge in 20 rounds with teleportation probability of Beta=0.15.
5. Hubs and Authorities ranking produces the same hub and authorities values for each node if the graph is bidirectional.
6. Spam Farm effectiveness increases with when PageRank teleportation probability goes towards zero.

Answer: Correct statements are 3 and 5.

VI. Sampling with Random Walks

Assume there is 1 million people in the world and the world consist of only three countries: A, B, and C. We know that on average each person in A has 300 friends, in B 200 friends and in C only 100 friends. They all become friends in Facebook. We know that the social graph that these friend relationship forms is connected and undirected. We perform a sufficiently long random walk on these nodes and find out that 60% of the nodes come from A, another 20% from B and the rest 20% from C. What is your estimate on the population sizes of these countries? (6 p)

Answer:

```
CounryAdegree = 300
CounryBdegree = 200
CounryCdegree = 100

ProportionOfRWsA = 60
ProportionOfRWsB = 20
ProportionOfRWsC = 20

PropotionSizeA= ProportionOfRWsA/CounryAdegree
PropotionSizeB= ProportionOfRWsB/CounryBdegree
PropotionSizeC= ProportionOfRWsC/CounryCdegree

sumNormalization = PropotionSizeA+PropotionSizeB+PropotionSizeC

EstimateA = PropotionSizeA/sumNormalization
EstimateB = PropotionSizeB/sumNormalization
EstimateC = PropotionSizeC/sumNormalization

EstimateA = 0.4000
EstimateB = 0.2000
EstimateC = 0.4000
```

VII. Spectral Analysis of Graphs

For each of the following statements indicate whether it's correct or wrong. (8 = 4 x 2 p)

1. In spectral analysis, the existence of large eigengap between the smallest and the second smallest eigenvalue of graph adjacency matrix A indicates that the graph is a very good expander.
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 graph Laplacian matrix L=D-A (where D is a degree matrix, and A is an adjacency matrix of G).
3. It is possible to perform graph bisection optimizing for "conductance" (i.e., assign each node to one of the two clusters) just by looking only at the eigenvalues of normalized Laplacian matrix $D^{-½}AD^{-½}$ (where D is a degree matrix, and A is an adjacency matrix of G).
4. The list of all eigenvalues of Laplacian matrix L (L=D-A, where D is diagonal degree matrix and A is an adjacency matrix of graph G) can provide us with information on which nodes belong to the largest cluster of G.

Answer: All statement are wrong.

VIII. Dimensionality reduction

a) CUR decomposition of matrix M is M=CUR. Assuming M is a sparse matrix, will C and U be sparse matrices? Explain in max 3 sentences. (2 p)

Answer (not explained here): C is sparse, U is not sparse

b) SVD decomposition of matrix M is $M=USV^T$. Select correct answer(s): (6 = 3 x 2 p)

1. U is always a square matrix.

2. The columns of V are eigenvectors of M.
3. Dot product of any two columns from U is equal to 0.

Answer: All statements are wrong.

c) 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
Alice							
Bob							
Carol							
David							
Erin							
Frank							
Grace							
Heidi							

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=

-0,545	0,029	-0,056	-0,487	-0,258	0,285
-0,062	0,012	0,542	-0,338	-0,289	-0,514
-0,59	0,027	0,015	0,72	-0,292	-0,012
-0,056	-0,626	0,004	0,14	0,47	-0,29
-0,02	0,008	0,542	0,174	-0,075	-0,245
-0,016	-0,778	0,014	-0,118	-0,382	0,233
-0,021	0,009	0,637	0,056	0,351	0,648
-0,588	0,026	-0,063	-0,252	0,518	-0,191

S=

13,737	0	0	0	0	0
0	9,064	0	0	0	0
0	0	7,797	0	0	0
0	0	0	1,502	0	0
0	0	0	0	0,856	0
0	0	0	0	0	0,595

V=

-0,589	0,043	0,001	-0,769	-0,183	-0,16
-0,545	0,039	-0,051	0,427	-0,49	0,526
-0,022	-0,706	0,011	-0,018	-0,033	0,014
-0,022	-0,706	0,011	-0,018	-0,033	0,014
-0,592	-0,027	-0,059	0,352	0,664	-0,283
-0,024	0,01	0,744	-0,177	0,365	0,531

-0,066	0,012	0,664	0,264	-0,386	-0,579

Answer the following questions. (12 = 6 x 2 p)

1. With how many main concepts (dimensions) would you identify matrix M and why?
2. Which user belongs to which concept (after dimensionality reduction)? Explain why.
3. Which movie belongs to which concept (after dimensionality reduction)? Explain why.
4. Is Heidi more similar to Grace or to Carol in terms of movie preferences?
5. Is Move1 more similar to Movie 5 or Movie 6?
6. What is the rank of Matrix M? Explain.

Answer:

1. ***3 concepts - > since the first three singular values on* S *diagonal are by far the largest.***
2. ***User to concept***
   a. ***Concept 1:* A, C, H**
   b. ***Concept 2:* D, F**
   c. ***Concept 3:* B, E, G**
3. ***Movies to concepts:***
   a. ***Concept 1:* 1, 2, 5**
   b. ***Concept 2:* 3, 4**
   c. ***Concept 3:* 6, 7**
4. **Heidi *is more similar to* Carol**
5. **Movie 1 *is more similar to* Movie 5**
6. ***The rank of matrix M is 6. Check the length of the diagonal of* S *with non-zero elements.***

IX. Recommender Systems

An online store has a database of millions of users and millions of data-items with the user ratings. It uses three recommender systems R1, R2 and R3 which are based on User-User CF, Item-Item CF and Content based respectively. Assume a completely new user A wants to do shopping in the above described online store (the shop does not have any data on A). Could any instance of the above Recommender Systems provide good recommendation to A? Explain in max 3 sentences. (6 p)

Answer: No good recommendation. Cold start problem.

X. Privacy Preserving Data Mining

For each of the following statements indicate whether it's correct or wrong. (8 = 4 x 2 p)

1. Privacy in data mining is discussed at both data publishing and function computation levels.
2. K-anonymity is a technique for privacy at data publishing level that generates anonymized versions of the original privacy sensitive datasets.
3. Secure multi-party computation is a set of protocols and cryptographic techniques that allow the computation of some functions on decentralized data in a peer-to-peer manner without having to share the data itself among the peers.
4. Achieving privacy at data publishing level often results in loss on data utility. Differential privacy provides high privacy levels all while guarantying zero loss to data utility, which makes it the de-facto standard in privacy preserving data publishing.

Answer: Statements 1, 2, and 3 are correct; Statement 4 is wrong.

XI. Label Propagation On Graphs

Figure 1 and Figure 2 show two graphs for which we want to apply the algorithm of label propagation with absorbing states. Each graph has some colored nodes that represent the labeled nodes (i.e., the nodes with absorbing states), while the rest of the nodes are unlabeled and they are assigned numbers to represent their IDs. We run the algorithm till it convergences.
For each of the following statements indicate whether it's correct or wrong. (6 = 3 x 2 p)

1. For the graph in Figure 1, node 9 can be labeled blue or red with equal probability.
2. For the graph in Figure 1, node 14 is more likely to be labeled red than blue.
3. For the graph in Figure 2, node 14 can be labeled red with higher probability than being labeled green.

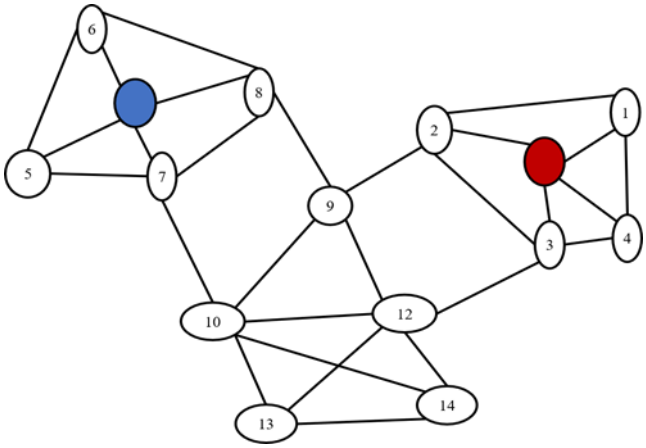

Figure 1

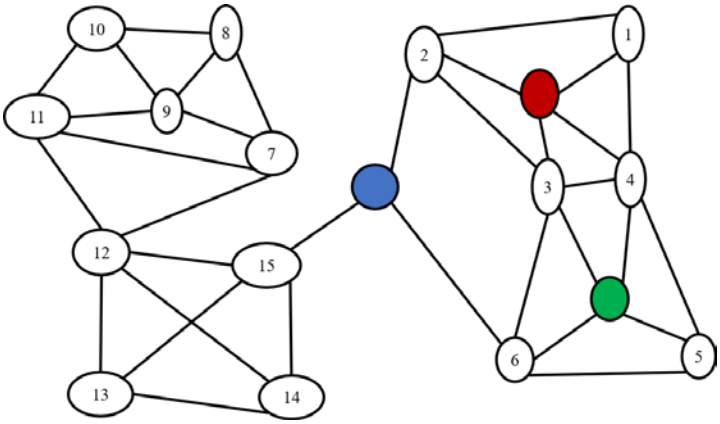

Figure 2

Answer: Statement 1 is correct; Statement 2 and 3 are wrong.

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