TU Delft

CYBER DATA ANALYTICS (CS4035)

Exam, July 2 2018
9:00 - 12:00

Important: This exam consists of 5 questions split into subquestions. For each subquestion you can get a maximum of 5 points. Always give full explanations of your answers and number all steps of the asked algorithms. Do not forget to put your name and student number on every sheet of paper. Answers are required to be in English.

Question 1 - class imbalance

(a) Explain how to use oversampling and undersampling when the data set consists of 99% benign cases. Describe both the input and output of both training and testing phases of the machine learning workflow. (5 pt)

(b) Explain how oversampling and undersampling changes the model returned by a learning algorithm. (5 pt)

(c) SMOTE is a popular alternative to sampling. Give an advantage and a disadvantage of using SMOTE compared to using sampling. (5 pt)

(d) Several researchers advice to, after SMOTE-ing the data, perform additional transformations such as removing Tomek links. Think of one positive and one negative effect of removing such links. Briefly explain these effects. (5 pt)

Question 2 - time series

(a) Anomaly detection uses correlation in the training data in order to detect outliers that violate these correlations in testing data. Explain how correlation plays a role in PCA and ARMA anomaly detection techniques. (5 pt)

(b) After training an anomaly detection method, you notice several anomalies in the training data. Is it considered good or bad practice to remove these from the data before training? Expain your answer for both PCA-based and ARMA-based anomaly detection methods. (5 pt)

(c) Let the Singular Value Decomposition of the data be given by $Y = USV^\top$, where $U \in \mathbb{R}^{m \times k}$, $S \in \mathbb{R}^{k \times k}$ is a diagonal with the ordered singular values as its entries, and $V \in \mathbb{R}^{n \times k}$.

Describe in pseudo-code how a PCA-based anomaly detection method could be implemented, given the data set Y and the matrices U, S, and V. (5 pt)

Question 3 - hashing

Consider a stream W of length n. Each item i in the stream is an integer. A count-min sketch is a specific data structure for approximating the amount of times an element i is in W.

(a) What does a count-min sketch with $\log(\frac{1}{\delta})$ rows and $\frac{2}{\epsilon}$ columns guarantee for the quality of approximation of an element's freqency? (5 pt)

Suppose a count-min sketch with 4 columns uses the following three hash functions:

$$h_1(i) = 3i \mod 4$$
$$h_2(i) = 2(i-1) \mod 4$$
$$h_3(i) = 3(i+3) \mod 4$$

(b) Compute the content of the sketch after observing the following stream $W : 2, 2, 5, 4, 5, 6, 2, 4$. First compute the outcome of the different hash functions on the different unique stream elements. What does the sketch return when queried for the number of times the element 2 occurs? (5 pt)

(c) Locality sensitive hashing is a technique that can speed up nearest neighbor testing using hash functions. Explain how this technique works for nearest neighbor search and what speed-up to expect when using m hashing bins to search over n data points, assuming uniform hash functions. (5 pt)

(d) Suppose that a hashing scheme h' assigns elements i and j to the same bin if and only if the three hash functions above give the same answer for i and j, i.e.,

$$h'(i) = h'(j) \Leftrightarrow ((h_1(i) = h_1(j)) \wedge (h_2(i) = h_2(j)) \wedge (h_3(i) = h_3(j)))$$

Is h' locality sensitive? Explain why (not). (5 pt)

Question 4 - sequential data mining

Suppose you are given the following labeled sequence data:
`(a,-), (aa,+), (aaba,-), (aabbb,+), (abb,+), (ba,-), (bb,-), (bab,+), (bbb,+)`

(a) Draw the prefix tree (including labels) for this data that is used as input for learning state machines. (5 pt)

(b) Number the states and detemine which states cannot be merged with the start state, explain in each case why. (5 pt)

(c) A common heuristic for state merging counts the number of accepting states that are merged with accepting states and the number of rejecting states that are merged with rejecting states. Find a merge with value of 2 for this heuristic, perform the merge and show the result. (5 pt)

(d) State machines can also be learned from unlabeled data. Give a merge heuristic that can be used to learn state machines from unlabeled data and explain how it is used in the state merging algorithm. (5 pt)

Question 5 - privacy

(a) Privacy can be preserved in data mining using perturbative and non-perturbative methods. Briefly explain the difference between these methods. Give two example algorithms of both methods. (5pt)

Suppose you are asked to publish the following table, in which Age, Gender, State of domicile, and Religion are assumed to be quasi-identifiers.

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	29	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

(b) Give the definition of K-anonimity. Make the data 2-anonymous using generalization and suppression (use abbreviations to save space). (5pt)

(c) Another methods to anonymize the data is to synthesize new rows. Give 2 advantages and 2 disadvantages of using synthetic data compared with using generalization and suppression. (5pt)