CSC 240/440 - 2017 Spring: Homework 5

Hand in the hard copy to CSB 614 before 4:50pm May 9

Requirement

Due to the request from some students, the homework is posted online right now, but would be updated probably every week until it is formally released. (A) or (G) indicates questions for all or just graduate students. Undergraduate students are not required to do (G) questions, but they can get bonus points from that. **Please hand in the hardcopy of your homework before the class.**

The homework must be completed individually. However, you are encouraged to discuss the general algorithms and ideas with classmates in order to help you answer the questions. If you work with one or more other people on the general discussion of the assignment questions, please record their names over every question they participated.

However, the following behaviors will receive heavy penalties (lose all points and apply the honest policy explained in syllabus)

- explicitly tell somebody else the answers;
- explicitly copy answers or code fragments from anyone or anywhere;
- allow your answers to be copied;
- get code from Web.

Please also indicate how many late days you want to apply to your submission (Check the late policy in the Syllabus). All late submission without indicating late days or running out the late days cannot be accepted. For medical reasons, if the homework in time cannot be submitted on time, you have to submit the certificate with your homework.

1 (A) 2 points

Consider the table of Question 2 in chapter 6, that is, Table 6.22. Generate a figure similar to Figure 6.4. Define frequent itemsets to be great than or equal to 4. Indicate infrequent itemsets and show the support for all frequent item sets.

2 (A) 2 points

(a) Prove that all nonempty subsets of a frequent itemset must also be frequent.
(b) Given the frequent itemset l and $s \subseteq l$, prove that the confidence of the rule $s' \implies (l - s')$ cannot be greater than $s \implies (l - s)$ where $s' \subseteq s$.
(c) One variation of the Apriori algorithm subdivides the transactions of a database D into n nonoverlapping partitions. Prove that any itemsite that is frequent in D must also be frequent in at least one partition of D.

3 (A) 2 points

Refresh your memory on the properties that are satisfied by valid distance metrics by rereading section 2.4.3. A closed pattern is defined to be a pattern p such that there is no superpattern p' with the same support as p (**Def 6.4**). In association rule mining, one pattern distance measure between closed patterns $P1$ and $p2$ used is defined to be

$$J(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|} \tag{1}$$

where $T(P)$ is the supporting transaction set of P. Prove that this is a valid distance metric.

4 (G) 2 points

Give the total number of possible rules extracted from a data set that contains d items and prove it.

5 (A) 2 point

What are outliers? List possible reasons for outliers.

6 (A) 2 points

Consider a binary classification problem. Assume that the labels of a few samples are incorrect (but you do not what they are). Design a strategy to identify all outlier samples (you can assume that all samples are linear separable if their labels are correct).

7 (A) 2 points

Give an example to show when filling missing values via mean value or median value does not make sense.

8 (A) 2 points

Read the document of XGBOOST and summarize the algorithm protocol and comment the ensemble approach.

9 (A) 2 points

List main parameters you can turn in XGBOOST. If you find XGBOOST has overfitting on your problem, what strategies you will try to avoid overfitting (give 2 at least).

10 (A) 3 points

Comment the key difference among supervised learning, unsupervised learning, and semi-supervised learning. Use one sentence to summarize the key idea in the self training algorithm and generative models respectively.