MSC-BDT5002/MSC-IT 5210 Knowledge Discovery and Data Mining, Fall 2017

Assignment 1

Deadline: Oct 6th, 2017

1 Submission Guidelines

- Assigments should be submitted to mscbdt5002fall17@gmail.com (for BDT students)/mscit5210fall17@gmail.com (for IT students) as attachments.
- Attachments should be original .pdf or .docx, NOT compressed.
- Attachments should be named in the format of: Ax_itsc_stuid.xxx. E.g. for a student with itsc account: sdiaa, student id: 20171234, the 1st assignment can be named as: A1_sdiaa_20171234.pdf.
- Submissions after the deadline or not following the rules above are NOT accepted.
- Your grade will be based on the correctness, efficiency and clarity.
- The email for **Q&A**: mscbdt5002it5210@gmail.com
- **Plagiarism will lead to zero mark.**
- Updated date: Sept 26th 2017.

2 Data Preprocessing

2.1 Wavelet Transform

Haar wavelet is the simplest type of wavelet as mentioned. Now given the scaling function that is a simple rectangle function

$$\varphi(t) = \begin{cases} 1, & \text{if } 0 \leq t \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

with only two nonzero coefficients $h(0) = h(1) = 1/\sqrt{2}$ (Note: The lecture example is under the hypothesis that two nonzero coefficients $h(0) = h(1) = 1/2$).

Questions (20 marks)

- Describe the discrete wavelet transform.
- Compute the discrete Haar wavelet transform of

$$[1, 4, 2, 3, -2, -1, 2, 1]^T.$$

Please answer above questions based on the reference: Burrus C S, Gopinath R A, Guo H. Introduction to wavelets and wavelet transforms: a primer[J]. 1997.

2.2 Principal Components Analysis

Principal Components Analysis as an unsupervised linear dimensionality reduction framework achieves impressive performance many among state-of-art techniques. An original dataset has been shown in Table 1.

Questions (30 marks)

- Calculate the covariance matrix of data as shown in the Table 1.
- Calculate eigenvectors and eigenvalues of the covariance matrix.
- Calculate the proportion of total population variance explained by the first two components.

Note: Programming is allowed. If so, you are required to write a C++/ python2.7/ matlab program (internal functions are allowed). Please post your results in the assignment document and submit your source code. Note that all the codes should be compilable and well-commented (provide enough comments for each key line of code), otherwise you may lose some marks if the code is very difficult to understand.

Table 1: An Original Dataset

ID	Attributes 1	Attributes 2	Attributes 3
1	-1	-1	1
2	-2	-1	4
3	-3	-2	-2
4	1	1	1
5	2	1	2
6	3	2	1
7	1	2	4

3 Pattern Discovery

There are session records as shown in the Table 2. Please using equal-depth binning to discretize Number of Web pages, number of bins 4, Using equal-width binning to discretize Session Length, number of bins 3, Category the country attribute according to its continent (e.g. China, Japan, Korea belong to Asia, Canada and USA belong to North American). We set $min\_sup = 0.3$ for following questions.

Questions (50 marks)

- Show the major steps to find the frequent patterns using Apriori of the transactions.
- Show the major steps to find the frequent patterns using FP-Growth of the transactions.
- Based on the frequent patterns you get, which are closed frequent patterns? Which are max frequent patterns?

Table 2: Session Records

Session ID	Country	Session Length	#web pages	Buy
1	USA	1213	9	No
2	China	2017	11	Yes
3	Germany	598	35	Yes
4	France	898	45	No
5	Canada	672	9	No
6	Japan	998	14	Yes
7	Korea	1543	18	Yes
8	China	267	7	No
9	USA	1702	13	No
10	England	1345	36	Yes