CSE 40647/60647 Data Science (Fall 2017)
Sample-Final Exam

(120 minutes, 100 marks, double sided reference, brief answers)

Name: NetID: Score:

1. [15] Multiple-choice questions for frequent pattern mining.

(a) [5] A database has 10 transactions. Let $min\_sup = 2$.

trans_id	items
1	{a, b }
2	{b, c, d }
3	{a, c, d, e }
4	{a, d, e }
5	{a, b, c }
6	{a, b, c, d }
7	{a }
8	{a, b, c }
9	{a, b, d }
10	{b, c, e }

Please choose frequent patterns from the following patterns.

- 1: ae
- 2: ade
- 3: abd
- 4: abcd

(b) [5] A database has 4 transactions. Let $min\_sup = 2$.

trans_id	items
1	{A, C, F, G}
2	{A, B, C, F}
3	{A, B, C, D, F}
4	{B, D, E}

Please choose closed patterns from the following patterns.

- 1: D
- 2: ABCF

- 3: BF

- 4: BD

(c) [5] A sequence database has 3 sequences. Items in the same parenthesis means they were got together in one event. Let $min\_sup = 2$.

sequence_id	sequence
1	(AB)C(FG)G
2	(AD)CB(ABF)
3	AB(FG)

Please choose sequential patterns from the following patterns.

- 1: (FG)B

- 2: (FG)

- 3: B(FG)

- 4: GF

2. [55] Classification.

Please use ID3 Decision Tree model and Naïve Bayes model to predict result of a game.
(Note: the dataset/questions in the actual final exam will be much smaller/easier than this homework/sample but the style is similar.)

Data: Each data object is a game, we have three attributes:
(1) “Is Home/Away?”, a 2-value attribute (“Home”, “Away”),
(2) “Is Opponent in AP Top 25 at Preseason?”, a 2-value attribute (“In”, “Out”),
(3) “Media”, a 3-value attribute (“1-NBC”, “2-ESPN”, “3-FOX”, “4-ABC”, “5-CBS”).
The label “Win/Lose” is binary (“Win”, “Lose”).

Training set: 24 games. Please use game ID 1–24 to build classification models.

Testing set: 6 games. Please use game ID 25–30 to evaluate the performance of classification models (blue font).

(a) [25] Construct a decision tree use **ID3** model and draw the final decision tree.

(b) [5] Use the decision tree to predict labels of game 25–30.

(c) [25] Use **Naïve Bayes** model to predict labels of game 25–30 given game 1–24.

ID	Is Home or Away	Is Opponent in AP25 at Preseason	Media	Label: Win/Lose
1	Home	Out	1-NBC	Win
2	Away	Out	4-ABC	Win
3	Home	In	1-NBC	Win
4	Home	Out	1-NBC	Win
5	Away	In	4-ABC	Lose
6	Home	Out	1-NBC	Win
7	Home	In	1-NBC	Win
8	Away	Out	4-ABC	Win
9	Away	Out	4-ABC	Win
10	Home	Out	1-NBC	Win
11	Away	Out	1-NBC	Win
12	Away	In	3-FOX	Lose
13	Away	Out	4-ABC	Lose
14	Home	Out	1-NBC	Win
15	Home	Out	1-NBC	Lose
16	Home	Out	1-NBC	Lose
17	Home	Out	2-ESPN	Win
18	Away	Out	4-ABC	Lose
19	Home	In	1-NBC	Lose
20	Home	Out	1-NBC	Win
11	Home	Out	5-CBS	Lose
22	Home	Out	1-NBC	Win
23	Home	In	1-NBC	Lose
24	Away	In	4-ABC	Lose
25	Home	Out	1-NBC	Win
26	Home	In	1-NBC	Lose
27	Away	Out	2-ESPN	Win
28	Away	Out	3-FOX	Win
29	Home	Out	1-NBC	Win
30	Away	Out	4-ABC	Win

3. [30] Clustering.

Suppose we have 10 college soccer team X1 to X10. We want to cluster them into 2 groups. For each soccer team, we have two features: One is # wins in Season 2016, and the other is # wins in Season 2017.

Team	# wins in Season 2016	# wins in Season 2017
X_1	5	7
X_2	6	7
X_3	2	8
X_4	7	8
X_5	8	4
X_6	6	4
X_7	7	3
X_8	6	3
X_9	5	2
X_{10}	4	3

(a) [10] Initialize with two centroids, (6, 4) and (6, 5). Use Manhattan distance as the distance metric. Please use K-Means to find two clusters.

(b) [10] Initialize with two centroids, (6, 4) and (6, 5). Use Euclidean distance as the distance metric. Please use K-Means to find two clusters.

(c) [10] Suppose we initialize with two medoids, (2, 8) and (8, 4). Use Euclidean distance as the distance metric. In K-Medoids clustering, given a non-medoid (5,7), do we swap the medoid (2, 8) with (5, 7)?