Quiz 4

- There are 5 problems total worth 12 points as shown in each question.
- You must not communicate with other students during this test.
- No books, notes allowed.
- No other electronic device except calculators are allowed. You cannot use your mobile as calculators.
- This is a 30 minute exam.
- Do not turn this page until instructed to.
- There are several different versions of this exam.

1. Fill in your information:

Full Name: _____

NetID: _____

1/1. (3 points) Below is a sequence dataset containing transaction sequences. Note that (bc) means items b and c are purchased at the same time. Let **min_sup = 4**,

SID	Sequence
S1	abcde
S2	a(bc)bd
S3	(bc)(bd)
S4	c(de)
S5	a(bd)e

Use the Generalized Sequential Patterns (GSP) mining algorithm to find L-2, C-2 frequent sequential patterns. You need to list all the steps and results from each step.

Solution. Partial Credits if support not specified.
$C_1 =< a:3 >, < b:4 >, < c:4 >, < d:5 >, < e:3 >$;
$L_1 =< b:4 >, < c:4 >, < d:5 >$ (1 point for both)
$C_2 =< bb >< bc >< bd >< cb >< cc >< cd >< db >< dc >< dd >< (bc) >< (bd) >< (cd) >$ (1 point)
$L_2 = cd:4 (1 point)$

2/1. (2 points) Below is a sequence dataset containing transaction sequences. Note that (bc) means items b and c are purchased at the same time. Let **min_sup = 2**,

SID	Sequence
S1	abcdef
S2	a(bc)bd
S3	a(bc)(bd)
S4	c(de)
S5	ab(bd)e

A. Given the above sequence database, enumerate corresponding suffix of each sequence whose $prefix =< ab >$.

B. For the above database, what would be the <a>-projected database.

Solution.

A.

SID	Suffix
S1	cdef
S2	(_c)bd
S3	(_ c)(bd)
S4	NULL
S5	(bd)e

B.

SID	Suffix
S1	bcdef
S2	(bc)bd
S3	(bc)(bd)
S4	NULL
S5	b(bd)e

2/2. (2 points) Below is a sequence dataset containing transaction sequences. Note that (bc) means items b and c are purchased at the same time. Let **min_sup = 2**,

SID	Sequence
S1	abcdef
S2	a(bc)bd
S3	(bc)(bd)
S4	bc(de)
S5	b(bd)e

A. Given the above sequence database, enumerate corresponding suffix of each sequence whose $prefix =< b >$.

B. For the above database, what would be the <a>-projected database.

Solution.

A.

SID	Suffix
S1	NULL
S2	NULL
S3	(_ c)(bd)
S4	c(de)
S5	(bd)e

B.

SID	Suffix
S1	bcdef
S2	(bc)bd
S3	NULL
S4	NULL
S5	NULL

2/3. (2 points) Below is a sequence dataset containing transaction sequences. Note that (bc) means items b and c are purchased at the same time. Let **min_sup = 2**,

SID	Sequence
S1	abcdef
S2	a(bc)bd
S3	(bc)(bd)
S4	c(de)
S5	a(bcd)e

A. Given the above sequence database, enumerate corresponding suffix of each sequence whose $prefix =< a(bc) >$.

B. For the above database, what would be the <a>-projected database.

Solution.

A.

SID	Suffix
S1	NULL
S2	bd
S3	NULL
S4	NULL
S5	(_d)e

B.

SID	Suffix
S1	bcdef
S2	(bc)bd
S3	NULL
S4	NULL
S5	(bcd)e

3/1. (3 points) Consider the construction of a Naive Bayes classifier from the following training dataset. Suppose we want to predict whether a restaurant is popular based on its price and parking availability based on the training data in Figure 4. Please answer the following questions.

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Price	Medium	High	Low	Medium	Low	Medium	Low	High
Parking	Available	Available	Available	No	Available	No	No	Available
Popularity	P	NP	P	NP	P	P	NP	P

Figure 1: Training Dataset for a Restaurant (P - popular, NP - not popular)

A. Please estimate the following terms (No Smoothing is required):

   (a) Given a random restaurant(no other given feature value), what is the probability that it would be 'Popular' or 'Not Popular' ? Comment based on the given training data.

   (b) Suppose you visit a restaurant which is 'Not Popular', what is the probability that it's an inexpensive restaurant (low price) and you will be able to park your car?

   (c) What would be the probability of (b) if you visit a 'Popular' restaurant?

B. Suppose a restaurant you visit has these values: Price = 'Low', Parking = 'Available'. Based on your previous calculations, is this restaurant classified as popular? Please show your reasons.

Solution.

A. $P(Popular) = 5/8, P(NotPopular) = 3/8$ (0.5 point)

B. $P(X|C) = P(Low|NP) * P(Available|NP) = 1/3 * 1/3$ (1 point)

C. $P(X|C) = P(Low|P) * P(Available|P) = 2/5 * 4/5$ (1 point)

D. $P(C|X) = P(X|C) * P(C); P(Popular|X) = 8/25 * 5/8 = 1/5$
$P(NotPopular|X) = 1/9 * 3/8 = 1/24; P(P|X) > P(NP|X)$ (0.5 point)

3/2. (3 points) Consider the construction of a Naive Bayes classifier from the following training dataset. Suppose we want to predict whether a restaurant is popular based on its price and parking availability based on the training data in Figure 4. Please answer the following questions.

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Price	Medium	High	Low	Medium	Low	Medium	Low	High
Parking	Available	Available	Available	No	Available	No	No	Available
Popularity	P	NP	P	NP	P	P	NP	P

Figure 2: Training Dataset for a Restaurant (P - popular, NP - not popular)

A. Please estimate the following terms (No Smoothing is required):

(a) Given a random restaurant(no other given feature value), what is the probability that it would be 'Popular' or 'Not Popular' ? Comment based on the given training data.

(b) Suppose you visit a restaurant which is 'Not Popular', what is the probability that it's an expensive restaurant (high price) and you will be able to park your car?

(c) What would be the probability of (b) if you visit a 'Popular' restaurant?

B. Suppose a restaurant you visit has these values: Price = 'High', Parking = 'Available'. Based on your previous calculations, is this restaurant classified as popular? Please show your reasons.

Solution.

A. $P(Popular) = 5/8, P(NotPopular) = 3/8$ (0.5 point)

B. $P(X|C) = P(High|NP) * P(Available|NP) = 1/3 * 1/3$ (1 point)

C. $P(X|C) = P(High|P) * P(Available|P) = 1/5 * 4/5$ (1 point)

D. $P(C|X) = P(X|C) * P(C); P(Popular|X) = 4/25 * 5/8 = 1/10$
$P(NotPopular|X) = 1/9 * 3/8 = 1/24; P(P|X) > P(NP|X)$ (0.5 point)

3/3. (3 points) Consider the construction of a Naive Bayes classifier from the following training dataset. Suppose we want to predict whether a restaurant is popular based on its price and parking availability based on the training data in Figure 4. Please answer the following questions.

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Price	Medium	High	Low	Medium	Low	Medium	Low	High
Parking	Available	Available	Available	No	Available	No	No	Available
Popularity	P	NP	P	NP	P	P	NP	P

Figure 3: Training Dataset for a Restaurant (P - popular, NP - not popular)

A. Please estimate the following terms (No Smoothing is required):

(a) Given a random restaurant(no other given feature value), what is the probability that it would be 'Popular' or 'Not Popular' ? Comment based on the given training data.

(b) Suppose you visit a restaurant which is 'Not Popular', what is the probability that it's an inexpensive restaurant (low price) and you will not be able to park your car?

(c) What would be the probability of (b) if you visit a 'Popular' restaurant?

B. Suppose a restaurant you visit has these values: Price = 'Low', Parking = 'No'. Based on your previous calculations, is this restaurant classified as popular? Please show your reasons.

Solution.

A. $P(Popular) = 5/8, P(NotPopular) = 3/8$ (0.5 point)

B. $P(X|C) = P(Low|NP) * P(NotAvailable|NP) = 1/3 * 2/3$ (1 point)

C. $P(X|C) = P(Low|P) * P(NotAvailable|P) = 2/5 * 1/5$ (1 point)

D. $P(C|X) = P(X|C) * P(C); P(Popular|X) = 2/25 * 5/8 = 1/20$ (0.5 point)
$P(NotPopular|X) = 2/9 * 3/8 = 1/12; P(P|X) < P(NP|X)$

3/4. (3 points) Consider the construction of a Naive Bayes classifier from the following training dataset. Suppose we want to predict whether a restaurant is popular based on its price and parking availability based on the training data in Figure 4. Please answer the following questions. (No need to calculate fractions)

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Price	Medium	High	Low	Medium	Low	Medium	Low	High
Parking	Available	Available	Available	No	Available	No	No	Available
Popularity	P	NP	P	NP	P	P	NP	P

Figure 4: Training Dataset for a Restaurant (P - popular, NP - not popular)

A. Please estimate the following terms (No Smoothing is required):

(a) Given a random restaurant(no other given feature value), what is the probability that it would be 'Popular' or 'Not Popular' ? Comment based on the given training data.

(b) Suppose you visit a restaurant which is 'Not Popular', what is the probability that it's moderately priced restaurant (medium price) and you will be able to park your car?

(c) What would be the probability of (b) if you visit a 'Popular' restaurant?

B. Suppose a restaurant you visit has these values: Price = 'Medium', Parking = 'Available'. Based on your previous calculations, is this restaurant classified as popular? Please show your reasons.

Solution.

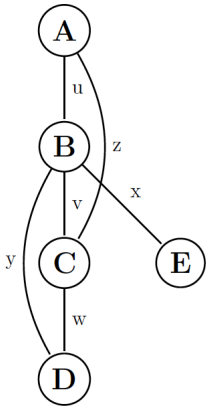
A. $P(Popular) = 5/8, P(NotPopular) = 3/8$ (0.5 point)

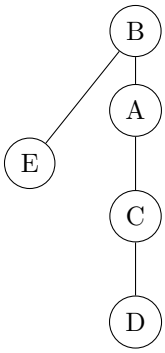
B. $P(X|C) = P(Medium|NP) * P(Available|NP) = 1/3 * 1/3$ (1 point)

C. $P(X|C) = P(Medium|P) * P(Available|P) = 2/5 * 4/5$ (1 point)

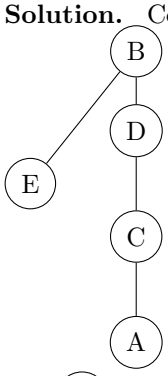
D. $P(C|X) = P(X|C) * P(C); P(Popular|X) = 8/25 * 5/8 = 1/5$
$P(NotPopular|X) = 1/9 * 3/8 = 1/24; P(P|X) > P(NP|X)$ (0.5 point)

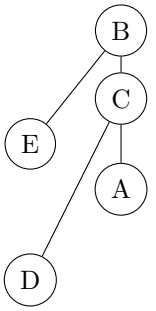
4/1. (2 points) In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x,y, z). Show three different DFS search trees rooted at node B and their right most paths respectively.



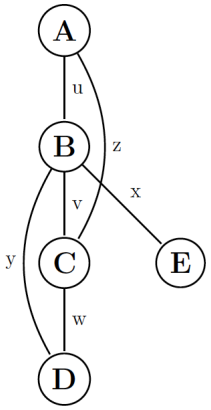


Solution. Could be more correct solutions.

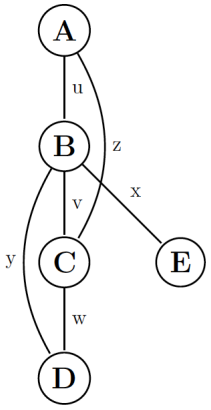




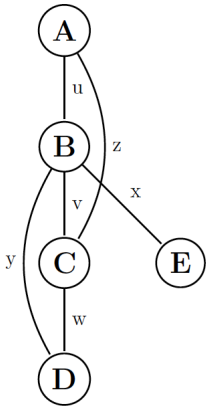
4/2. (2 points) In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x,y, z). Show three different DFS search trees rooted at node C and their right most paths respectively.



4/3. (2 points) In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x,y, z). Show three different DFS search trees rooted at node D and their right most paths respectively.



4/4. (2 points) In the graph shown above, there are five nodes (A, B, C, D, E) and six edges (u, v, w, x,y, z). Show three different DFS search trees rooted at node A and their right most paths respectively.



5/1. (2 points) Suppose we built a classifier to predict whether a patient has cancer or not. We want to evaluate the performance of the classifier, and collect its confusion matrix as in the following table. In the given confusion matrix, 'A' represents 'Actual class', and 'P' represents 'Predicted class'. Given that *Sensitivity* = 0.625, *Specificity* = 0.75 and *Precision* = 0.667, fill the blanks in the confusion matrix.
Hint: $Sensitivity = \frac{TruePositive}{TotalPositive}$, $Specificity = \frac{TrueNegative}{TotalNegative}$, $Precision = \frac{TruePositive}{TruePositive+FalsePositive}$

A / P	Yes (has cancer)	No
Yes		30
No	25	

Solution. (1 point for each)

A / P	Yes (has cancer)	No
Yes	50	30
No	25	75