Faculty of Science

Course:	CSCI 4150U: Data Mining
Instructor:	Kourosh Davoudi
Course component:	Quiz1
Weight:	15%
Duration:	60 minutes

- Please be online on google meet till the time you submit. The link is the same as what we used for the lectures: meet.google.com/nws-agxy-qit

- The exam is an open book, and you can use the slides, your notes, and the resources that the instructor shared with you. **You are not allowed to search the internet to find the answers.**
- Any sign of academic misconduct will be followed up and can have a serious academic penalty. It is the responsibility of students to be aware of the actions that constitute academic misconduct. Please see:

  http://calendar.uoit.ca/content.php?catoid=22&navoid=879#Academic_conduct

- You have 1 hour to earn 15 points.
- **The due time is 10:50 AM** (except you have an accommodation letter), and **you should submit it before due time**. However, the exam is **available till 12:00 PM** for very special circumstances like power outages, computer crashes and etc.
- Please note that you have only one attempt.
- If there is a technical problem and you submit answers after the due time, explain the situation in the last text box after all questions. You should provide reasonable evidence.
- You need to write the answers to each question in the provided space (**please type it**). Just typing the answers is enough.
- Do not spend too much time on any problem.
- Pay close attention to the instructions for each problem and just answer what is requested.
- Good Luck!

Question 1 [1 mark]

What are the Cosine and Extended Jaccard Coefficient similarities between point **x** and **y**?

$\mathbf{x} = [1, 3, 5, 8]$

$\mathbf{y} = [2, 5, -1, 6]$

$\mathbf{x}.\mathbf{y} = (1*2 + 3*5 + 5*(-1) + 8*6 = 60$
$||\mathbf{x}|| = (1+9+25+64)^{0.5} = 9.95$
$||\mathbf{y}|| = (4+25+1+36)^{0.5} = 8.12$

$\text{Cos} = \mathbf{x}.\mathbf{y} / ||\mathbf{x}||*||\mathbf{y}|| = 60/9.95*8.12 = 0.74$

$\text{EJ} = \mathbf{x}.\mathbf{y}/||\mathbf{x}||^2 + ||\mathbf{y}||^2 - \mathbf{x}.\mathbf{y} = 0.57$

Question 2 [1 mark]

Represented two documents doc1 and doc2 as two vectors in vector space (use term frequency method and assume terms are words).

doc 1 = "*I do not fear computers*"

doc 2 = "*I fear lack of them*"

	I	do	not	fear	computers	lack	of	them
V1:	1	1	1	1	1	0	0	0
V2:	1	0	0	1	0	1	1	1

The number of dimensions should be 8. Note the order of columns might be different

Question 3 [1 mark]

What is the advantage of parallel coordinates over the scatter plot?

They can be used to visualize high-dimensional data

Question 4 [1 mark]

Given the following data for the attribute X:

25, 7, 2, 6, 12, 7, 8, 7, 3, 21, 10, 9, 8, 18, 15

A) What is the $X_{40\%}$?

15 * 0.4 = 6

2, 3, 6, 7, 7, 7, 8, 8, 9, 10, 12, 15, 18, 21, 25

Answer = 8

Question 5 [2 mark]

What is the problem with "*Information Gain*" in decision tree building? How can we resolve the issue?

It tends towards selecting the attributes with **many values**.
Using **Gain ratio** that incorporates split information could be helpful

Question 6 [2 mark]

Assume that accuracy of model M (e.g., decision tree) calculated using the *test* and *training* datasets are P_{test} and P_{train} respectively.

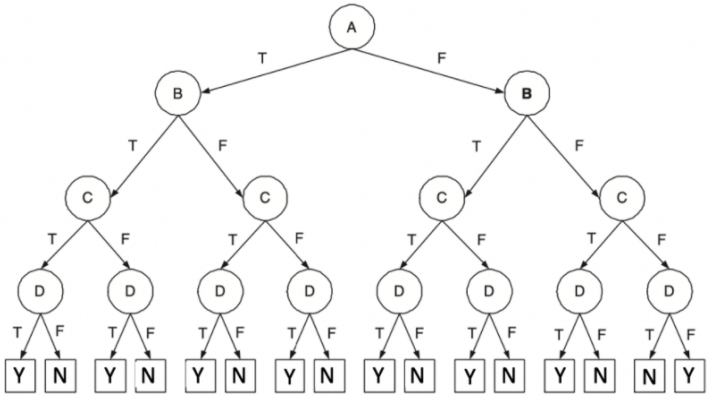
A) What is your interpretation if P_{train} is much higher than P_{test}

B) Propose potential solutions to resolve the issue.

A) it shows **overfitting** problem
B) Preparing **more training instances**, and **reducing the complexity** of the model

Question 7 [3 mark]

Given the following decision tree and the test data set (left table) with actual class labels as follows:



A	B	C	D	Class	
T	T	T	T	Y	Y
T	T	T	F	Y	N
T	T	F	T	N	Y
T	T	F	F	Y	N
T	F	T	T	N	Y
T	F	T	F	Y	N
T	F	F	T	N	Y
T	F	F	F	N	N

A) Determine the confusion matrix.

B) What is the f-measure of the model?

	Predicted	
	Y	N
Y	1	3
N	3	1

Precision = 1/1+3 = 0.25
Recall = 1/1+3 = 0.25

Fm = 2 P*R/P + R = 2 * 0.25 * 0.25/0.25+0.25 = **0.25**

Question 8 [4 mark]

The dataset shown in the following Table will be used to learn a classifier for predicting whether a wild fruit is edible or not based on its *shape*, *color*, and *odour*.

Shape	Color	Odour	Edible?
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes
C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	W	3	No
D	W	3	No

A) If we train a decision tree, which attribute would be chosen for the root of the decision tree using the Gini Index and multiway split?

NO	5
YES	5

Gini of parent P = 1- (5/10)^2-(5/10)^2 = 0.5

Shape (S):

	C	D
NO	2	3
YES	2	3

GI(S=C) = 1- (2/4)^2 - (2/4)^2 = 1/2
GI(S=D) = 1-(3/6)^2-(3/6)^2 = 1/2

GI(S) = 4/10 GI(S=C) + 6/10 GI(S=D) = 0.5
(Or Information Gain = 0)

Color (C):

	B	W	G	U
NO	1	2	1	1
YES	3	2	0	0

GI(C=B) = 1-(3/4)^2-(1/4)^2 =1-9/16-1/16= 6/16 = 3/8
GI(C=W) = 1-(2/4)^2-(2/4)2 = 1/2
GI(C= G) = 1 -(1/1)^2-(0/1)^2 = 0
GI(C=U) = 1-(1/1)^2-(0/1)^2 = 0

GI(C) = 4/10*3/8 +4/10 * 1/2 * 0 + 1/10 * 0 = 12/80 + 4/20 = 7/20 = 0.35
(Or Information Gain = 0.5-0.35 = 0.15)

Odour (O):

	1	2	3
NO	0	3	2
YES	3	2	0

GI(O=1) = 1-(3/3)^2-(0/3)^2 = 0
GI(O=2) = 1- (3/5)^2-(2/5)^2 = 1-9/25 - 4/25= 12/25
GI(O=3) = 1-(0/2)^2+(2/2)^2 = 0

GI(O) = 3/10 * 0 + 5/10* 12/25 + 2/10* 0 = 6/25 = 0.24
(Or Information Gain = 0.5-0.24 = 0.26)

0.24 < 0.35 => the first node is Odour
(or 0.26 >015)