Started on	Sunday, 27 April 2025, 11:58 AM
State	Finished
Completed on	Sunday, 27 April 2025, 12:09 PM
Time taken	11 mins 21 secs
Grade	**15.00** out of 100.00
Feedback	You have submitted your attempt successfully!

Information

Name: Le CHEN
Email: lchen546-c@my.cityu.edu.hk

CITY UNIVERSITY OF HONG KONG

Course code & title: CS5483 Data Warehousing and Data Mining
Session : Semester B 2024/25
Time allowed : Two hours

- This is a **computerized examination** that requires **internet access**.
- There are **NO printings** of question papers or answer books.

- Answer ALL questions using your designated Lab PC.
- For numerical answers that are not integer, they must be accurate at least up to 3 significant figures. Equivalent algebraic forms without simplification are acceptable.

This is a **closed-book** examination.

Students are allowed to use the following materials/aids:

- blank scratch papers;
- university approved calculators; and
- online python tutor: https://deep.cs.cityu.edu.hk/optmentor/opt-mentor/live.html

Materials/aids other than those stated above are not permitted. Students will be subject to disciplinary action if any unauthorized materials or aids are found on them.

Question 1
Correct
Mark 1.67 out of 5.00

(5 points) Which of the following statements are true for ensemble methods:

☐ a. The base classifiers can be trained in parallel for Adaboost.
☐ b. Random forest is nothing but bootstrap aggregation of decision trees.
☑ c. It is beneficial to diversify the base classifiers so that they can capture different pieces of information. ✓
☑ d. The original Adaboost algorithm does not apply to multi-class classification. ✓
☐ e. Accuracy will drop as we increasing the number of base classifiers due to overfitting.

Your answer is correct.

The correct answers are: It is beneficial to diversify the base classifiers so that they can capture different pieces of information., The original Adaboost algorithm does not apply to multi-class classification.

Correct

Marks for this submission: 5.00/5.00. Accounting for previous tries, this gives **1.67/5.00**.

Question 2
Correct
Mark 2.67 out of 4.00

(4 points) Match each ensemble method with exactly one statement that describes it. (Note that a statement may describe more than one method but there is only one perfect matching.)

Forest-RI [Consider only a randomly selected subsets of features for training each base classifier. ⬍] ✓

Voting [Allow different classification algorithms to be used to train the base classifiers. ⬍] ✓

Stacking [Allow a custom classification algorithm for training the combined classifier. ⬍] ✓

Adaboost [Use a weighted sum rule for the combined classifier. ⬍] ✓

Your answer is correct.

The correct answer is: Forest-RI → Consider only a randomly selected subsets of features for training each base classifier., Voting → Allow different classification algorithms to be used to train the base classifiers., Stacking → Allow a custom classification algorithm for training the combined classifier., Adaboost → Use a weighted sum rule for the combined classifier.

Correct

Marks for this submission: 4.00/4.00. Accounting for previous tries, this gives **2.67/4.00**.

Question 3
Correct
Mark 6.00 out of 8.00

(8 points) Complete the following paragraph by dragging the correct phrases into the blanks:

[kMeans] ✓ is a centroid-based clustering method, which identifies clusters by computing [cluster centers] ✓ . However, it can only identify [spherical clusters] ✓ , just like [complete linkage method] ✓ . [single linkage method] ✓ can identify non-spherical clusters, but it can mistakenly combine desired clusters together due to [chaining phenominon] ✓ . A density-based method called [DBSCAN] ✓ can grow non-spherical clusters from pillars of dense regions. To identical clusters of varying density levels, one can use [OPTICS] ✓ , which is implemented using a priority queue that flavors closely reachable points.

Your answer is correct.

The correct answer is: (8 points) Complete the following paragraph by dragging the correct phrases into the blanks:

[kMeans] is a centroid-based clustering method, which identifies clusters by computing [cluster centers]. However, it can only identify [spherical clusters], just like [complete linkage method]. [single linkage method] can identify non-spherical clusters, but it can mistakenly combine desired clusters together due to [chaining phenominon]. A density-based method called [DBSCAN] can grow non-spherical clusters from pillars of dense regions. To identical clusters of varying density levels, one can use [OPTICS], which is implemented using a priority queue that flavors closely reachable points.

Correct

Marks for this submission: 8.00/8.00. Accounting for previous tries, this gives **6.00/8.00**.

Question 4

Correct

Mark 2.00 out of 4.00

(4 points) Give the pseudocode for DBSCAN by reordering the following items:

Input: A set of points to be clustered.
Output: Cluster assignment.

✓ Repeatedly find a new point until all data points are processed.

✓ If the point is a core point, grow it into a cluster by a breadth-first search for density reachable points.

✓ If the point is not already assigned to another cluster, assign it to the current cluster.

✓ Label points not assigned to any clusters as noise.

Your answer is correct.

Correct

Marks for this submission: 4.00/4.00. Accounting for previous tries, this gives **2.00/4.00**.

Question 5

Correct

Mark 2.67 out of 4.00

(4 points) Check all the correct statements:

☑ a. A data cube helps archive transactional data for analysis. ✓
☐ b. The snowflake schema is better than the star schema for faster data analysis.
☐ c. Online transactional processing (OLTP) includes rolling-up, drilling-down, slicing, and dicing.
☑ d. Full materialization of the data cube is not efficient in terms of storage. ✓

Your answer is correct.

The correct answers are: A data cube helps archive transactional data for analysis., Full materialization of the data cube is not efficient in terms of storage.

Correct

Marks for this submission: 4.00/4.00. Accounting for previous tries, this gives **2.67/4.00**.

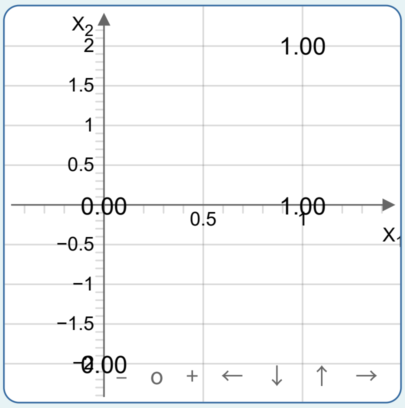
Question **6**

Not answered

Marked out of 8.00

Consider the following dataset D

Index	X_1	X_2	Y
1.	0	0	0
2.	0	-2	0
3.	1	0	1
4.	1	2	1



where X_1 and X_2 are numeric input attributes, and $Y \in \{0, 1\}$ is the class attribute.

(8 points) Using IBk classifier with Euclidean distance, compute the accuracy for the following cases. If random splitting of the dataset is needed, any valid split is acceptable. However, the split should be **stratified** whenever possible.

Give your answers as fractions (NOT percentages) to at least 3 significant figures. You may also enter your answer in an algebraic form such as $1/2$ without further simplifying.

a. Neighborhood size $k = 1$:

i. With minmax normalization:

Use training set:

Percentage split with 50% for testing:

Percentage split with 25% for testing:

2-Fold cross validation:

4-Fold cross validation:

ii. Without normalization:

4-Fold cross validation:

b. Neighborhood size $k = 3$:

i. With minmax normalization:

Use training set:

4-Fold cross validation:

The answer 1, which can be typed as `1`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer $\frac{1}{2}$, which can be typed as `1/2`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 0, which can be typed as `0`, would be correct.

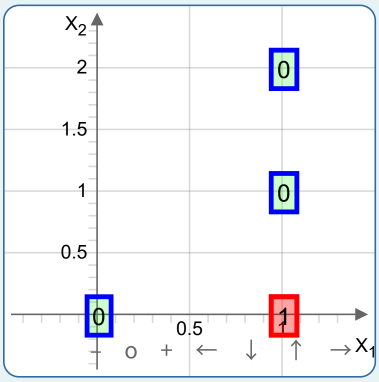
Question 7

Incorrect

Mark 0.00 out of 22.00

Consider the following dataset D

index	X_1	X_2	Y
1.	0	0	0
2.	1	0	1
3.	1	1	0
4.	1	2	0
5.	1	0	1



where $X_1 \in [0,1]$ and $X_2 \in [0,2]$ are numeric input attributes, and $Y \in \{0,1\}$ is the class attribute. Use the entire dataset for training.

(10 points) Calculate the following information quantities and determine the best first splitting attribute for C4.5.

Note that logarithm base 2 should be written as "log2" instead of "log" and you may express your answer in terms of the entropy function

$$h(p_1, p_2, \ldots) := \sum_k p_k \log_2 \frac{1}{p_k}$$

such as $1/3 * h(1/3, 2/3, 2/3) + 2/3$ without further simplifying.

$\mathrm{Info}(D) =$

$\mathrm{Info}_{X_1}(D) =$

$\mathrm{SplitInfo}_{X_1}(D) =$

$\mathrm{Gain}_{X_1}(D) =$

$\mathrm{GainRatio}_{X_1}(D) =$

$\mathrm{Info}_{X_2}(D) =$

$\mathrm{SplitInfo}_{X_2}(D) =$

$\mathrm{Gain}_{X_2}(D) =$

$\mathrm{GainRatio}_{X_2}(D) =$

Is X_1 a strictly better splitting attribute than X_2 for the first split? (Clear my choice)

(12 points) Apply the PART algorithm to build the decision list from partial C4.5 decision trees (without pruning) by maximizing the coverage. Give the coverage of each rule as a fraction.

If there are no conjuncts/consequences, choose the corresponding options INSTEAD of leaving the options as "No answer given".

✗ **Incorrect answer.**

Marks for this submission: 0.00/3.00. This submission attracted a penalty of 0.30.

✗ **Incorrect answer.**

Marks for this submission: 0.00/3.00. This submission attracted a penalty of 0.30.

✗ **Incorrect answer.**

Marks for this submission: 0.00/3.00. This submission attracted a penalty of 0.30.

Rule

Coverage

X_1>0.5 X_2>0.5 $\Longrightarrow$

Y=0

$X_1 \leq 0.5$ ⬍ $X_2 > 0.5$ ⬍ $\Longrightarrow$

$Y=0$ ⬍

$X_1 > 0.5$ ⬍ $X_2 \leq 0.5$ ⬍ $\Longrightarrow$

$Y=1$ ⬍

The information quantities can be computed from the following distributions:

$$P_Y = \left[\frac{3}{5}, \frac{2}{5}\right]$$

$$P_{X_1} = \left[\frac{1}{5}, \frac{4}{5}\right]$$

$$P_{X_2} = \left[\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right]$$

$$P_{Y|X_1} = \left[[1, 0], \left[\frac{1}{2}, \frac{1}{2}\right]\right]$$

$$P_{Y|X_2} = \left[\left[\frac{1}{3}, \frac{2}{3}\right], [1, 0], [1, 0]\right]$$

The answer $\frac{2\cdot\ln\left(\frac{5}{2}\right)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)}$, which can be typed as `(2*log(5/2))/(5*log(2))+(3*log(5/3))/(5*log(2))`, would be correct.

The answer $\frac{4}{5}$, which can be typed as `4/5`, would be correct.

The answer $\frac{\ln(5)}{5\cdot\ln(2)} + \frac{4\cdot\ln\left(\frac{5}{4}\right)}{5\cdot\ln(2)}$, which can be typed as `log(5)/(5*log(2))+(4*log(5/4))/(5*log(2))`, would be correct.

The answer $\frac{2\cdot\ln\left(\frac{5}{2}\right)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)} - \frac{4}{5}$, which can be typed as `(2*log(5/2))/(5*log(2))+(3*log(5/3))/(5*log(2))-4/5`, would be correct.

The answer $\frac{\frac{2\cdot\ln\left(\frac{5}{2}\right)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)} - \frac{4}{5}}{\frac{\ln(5)}{5\cdot\ln(2)} + \frac{4\cdot\ln\left(\frac{5}{4}\right)}{5\cdot\ln(2)}}$, which can be typed as `((2*log(5/2))/(5*log(2))+(3*log(5/3))/(5*log(2))-4/5)/(log(5)/(5*log(2))+(4*log(5/4))/(5*log(2)))`, would be correct.

The answer $\frac{3\cdot\left(\frac{\ln(3)}{3\cdot\ln(2)} + \frac{2\cdot\ln\left(\frac{3}{2}\right)}{3\cdot\ln(2)}\right)}{5}$, which can be typed as `(3*(log(3)/(3*log(2))+(2*log(3/2))/(3*log(2))))/5`, would be correct.

The answer $\frac{2\cdot\ln(5)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)}$, which can be typed as `(2*log(5))/(5*log(2))+(3*log(5/3))/(5*log(2))`, would be correct.

The answer $-\frac{3\cdot\left(\frac{\ln(3)}{3\cdot\ln(2)} + \frac{2\cdot\ln\left(\frac{3}{2}\right)}{3\cdot\ln(2)}\right)}{5} + \frac{2\cdot\ln\left(\frac{5}{2}\right)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)}$, which can be typed as `-((3*(log(3)/(3*log(2))+(2*log(3/2))/(3*log(2))))/5)+(2*log(5/2))/(5*log(2))+(3*log(5/3))/(5*log(2))`, would be correct.

The answer $\frac{-\frac{3\cdot\left(\frac{\ln(3)}{3\cdot\ln(2)} + \frac{2\cdot\ln\left(\frac{3}{2}\right)}{3\cdot\ln(2)}\right)}{5} + \frac{2\cdot\ln\left(\frac{5}{2}\right)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)}}{\frac{2\cdot\ln(5)}{5\cdot\ln(2)} + \frac{3\cdot\ln\left(\frac{5}{3}\right)}{5\cdot\ln(2)}}$, which can be typed as `(-((3*(log(3)/(3*log(2))+(2*log(3/2))/(3*log(2))))/5)+(2*log(5/2))/(5*log(2))+(3*log(5/3))/(5*log(2)))/((2*log(5))/(5*log(2))+(3*log(5/3))/(5*log(2)))`, would be correct.

The answer **False** would be correct.

A correct answer is: "(No conjunct on X_1)"

A correct answer is: "X_2>0.5"

A correct answer is: "Y=0"

The answer $\frac{2}{5}$, which can be typed as 2/5, would be correct.

A correct answer is: "X_1>0.5"

A correct answer is: "(No conjunct on X_2)"

A correct answer is: "Y=1"

The answer $\frac{2}{5}$, which can be typed as 2/5, would be correct.

A correct answer is: "(No conjunct on X_1)"

A correct answer is: "(No conjunct on X_2)"

A correct answer is: "Y=0"

The answer $\frac{1}{5}$, which can be typed as 1/5, would be correct.

Question 8

Not answered

Marked out of 19.00

Consider two clusterings C and C' of 6 data points defined as follows:

$$C(\boldsymbol{p}_i) = \begin{cases} 1, & i \in \{1,2,3\} \\ 2, & i \in \{4,5,6\}\,. \end{cases}$$

$$C'(\boldsymbol{p}_i) = \begin{cases} 1, & i \in \{1,2\} \\ 2, & i \in \{3,4\} \\ 3, & i \in \{5,6\}\,. \end{cases}$$

(6 points) Complete each entry of the following table with the count of the associated data points.

	$C(\boldsymbol{p}_i) = 1$	$C(\boldsymbol{p}_i) = 2$
$C'(\boldsymbol{p}_i) = 1$		
$C'(\boldsymbol{p}_i) = 2$		
$C'(\boldsymbol{p}_i) = 3$		

(4 points) Using classes-to-clusters evaluation, give the optimal classes-to-clusters assignment and the accuracy of

C' when C is the ground truth.

cluster		class
1	←	(Clear my choice) ⬍
2	←	(Clear my choice) ⬍
3	←	(Clear my choice) ⬍

accuracy=

(1 point) Give the accuracy of C if C' is the ground truth instead.

accuracy=

(4 points) Complete the following table with the counts of the associated ordered pairs $(\boldsymbol{p}_i, \boldsymbol{p}_j)$ of points.

Note that i and j need not be distinct.

	$C(\boldsymbol{p}_i) = C(\boldsymbol{p}_j)$	$C(\boldsymbol{p}_i) \neq C(\boldsymbol{p}_j)$
$C'(\boldsymbol{p}_i) = C'(\boldsymbol{p}_j)$		
$C'(\boldsymbol{p}_i) \neq C'(\boldsymbol{p}_j)$		

(2 points) Give the B-Cubed precision and recall of C if C' is the ground truth:

precision=

recall=

(2 points) Give the B-Cubed precision and recall of C' if C is the ground truth

precision=

recall=

The answer 2, which can be typed as `2`, would be correct.

The answer 0, which can be typed as `0`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer 0, which can be typed as `0`, would be correct.

The answer 2, which can be typed as `2`, would be correct.

A correct answer is: `1`

A correct answer is: `null`

A correct answer is: `2`

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer 10, which can be typed as `10`, would be correct.

The answer 2, which can be typed as `2`, would be correct.

The answer 8, which can be typed as `8`, would be correct.

The answer 16, which can be typed as `16`, would be correct.

The answer $\frac{5}{9}$, which can be typed as `5/9`, would be correct.

The answer $\frac{5}{6}$, which can be typed as `5/6`, would be correct.

The answer $\frac{5}{6}$, which can be typed as `5/6`, would be correct.

The answer $\frac{5}{9}$, which can be typed as `5/9`, would be correct.

Question 9

Not answered

Marked out of 26.00

Consider the following transactional data corresponding to the purchase of items:

$$\begin{aligned} D &:= \{T_1, T_2, T_3, T_4, T_5\} \quad \text{where} \\ T_1 &:= \{1,2,3\} \\ T_2 &:= \{1,2\} \\ T_3 &:= \{1,3\} \\ T_4 &:= \{2,4\}, \\ T_5 &:= \{1,3\} \end{aligned}$$

and the set $\{1,2,3,4\}$ of items are represented by integer labels in their natural ordering.

(14 points) Complete the following table using the Apriori algorithm to identify all frequent itemsets with support count at least $\text{min\_sup} := 2$. Enter one candidate itemset in each row **in the order** they are generated by the algorithm, i.e., the lexicographical order of the item labels. Do **not** include itemsets that are pruned by the prune step.

To input an itemset, use BRACES such as {1,3}. There should be exactly 7 candidates for which the apriori algorithm needs to compute the counts.

Candidate itemset	Count	Frequent?
		(Clear my choice) ⬍
		(Clear my choice) ⬍
		(Clear my choice) ⬍
		(Clear my choice) ⬍
		(Clear my choice) ⬍
		(Clear my choice) ⬍
		(Clear my choice) ⬍

(12 points) Complete the following to list all the association rules and their metrics for rules with **support at least 0.4 and confidence at least 0.6**. List the rules **in descending lexicographical order of (lift, confidence, support)**, e.g., (2,0.6,0.5) should be before (1,1,0.4), which should be before (1,0.6,0.5)

Association Rules			Support	Confidence	Lift
	$\Longrightarrow$				
	$\Longrightarrow$				
	$\Longrightarrow$				

The answer $\{1\}$, which can be typed as {1}, would be correct.

The answer 4, which can be typed as 4, would be correct.

The answer **True** would be correct.

The answer $\{2\}$, which can be typed as {2}, would be correct.

The answer 3, which can be typed as 3, would be correct.

The answer **True** would be correct.

The answer $\{3\}$, which can be typed as {3}, would be correct.

The answer 3, which can be typed as 3, would be correct.

The answer **True** would be correct.

The answer $\{4\}$, which can be typed as {4}, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer $\mathbf{False}$ would be correct.

The answer $\{1, 2\}$, which can be typed as `{1,2}`, would be correct.

The answer 2, which can be typed as `2`, would be correct.

The answer $\mathbf{True}$ would be correct.

The answer $\{1, 3\}$, which can be typed as `{1,3}`, would be correct.

The answer 3, which can be typed as `3`, would be correct.

The answer $\mathbf{True}$ would be correct.

The answer $\{2, 3\}$, which can be typed as `{2,3}`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer $\mathbf{False}$ would be correct.

The answer $\{3\}$, which can be typed as `{3}`, would be correct.

The answer $\{1\}$, which can be typed as `{1}`, would be correct.

The answer $\frac{3}{5}$, which can be typed as `3/5`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer $\frac{5}{4}$, which can be typed as `5/4`, would be correct.

The answer $\{1\}$, which can be typed as `{1}`, would be correct.

The answer $\{3\}$, which can be typed as `{3}`, would be correct.

The answer $\frac{3}{5}$, which can be typed as `3/5`, would be correct.

The answer $\frac{3}{4}$, which can be typed as `3/4`, would be correct.

The answer $\frac{5}{4}$, which can be typed as `5/4`, would be correct.

The answer $\{2\}$, which can be typed as `{2}`, would be correct.

The answer $\{1\}$, which can be typed as `{1}`, would be correct.

The answer $\frac{2}{5}$, which can be typed as `2/5`, would be correct.

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer $\frac{5}{6}$, which can be typed as `5/6`, would be correct.