1) Multiple Choice

12 / 20

- Each correctly answered multiple choice question gives 1 point.
- Each incorrect answer results in -1 point.
- However, the minimum scores in each of the four groups are 0 points.

Data processing and feature spaces

3 / 5

Temperature in Celsius is of ☑ interval scale ☐ ratio scale.

The idea of TF-IDF text representation is that globally frequent terms are
☑ more ☐ less informative than less frequent terms.

Every categorical feature with *m* possible values is convertible into *m* binary features.
☑ True ☐ False

Manhattan (L1) distance between two points is always smaller or equal than Euclidean (L2) distance between the same two points.
☐ True ☑ False

When the mean is larger than the median, the underlying distribution is
☑ positively skewed ☐ negatively skewed.

Association Rule Mining

4 / 5

What is the size of the itemset generated from the transaction
{bread, butter, beer, milk, butter, beer}? ☐ 2 ☑ 4 ☐ 6

If confidence of the rule A->BCD is below the confidence threshold then all other rules created from the itemset {ABCD} are necessarily below the threshold as well.
☐ True ☑ False

Which algorithm requires that the items are sorted by frequency?
☐ Apriori ☑ FP Growth

A closed frequent itemset has ☐ no frequent supersets ☐ has no immediate frequent supersets ☑ has no immediate frequent superset of the same support.

Downward closing property:
When an item set X is frequent, all its ☐ subsets ☐ supersets are also frequent.

Classification	0 / 5
Classification is ☐ an unsupervised task ☑ a supervised task.	✓
Decision tree classifiers are of the ☐ lazy ☑ eager learning type.	✓
An attribute, in which all training examples have different values is ☐ problematic ☑ helpful for decision trees based on information gain in terms of generalization error.	✗
The accuracy / error rate evaluation metric is robust with respect to class imbalance. ☐ True ☐ False	–
In a k-fold cross validation, each data point belongs to the test set exactly ☐ once ☑ k times.	✗

Clustering	5 / 5
Clustering is ☑ an unsupervised task ☐ a supervised task.	✓
The k-Means algorithm is ☑ sensitive to outliers ☐ not sensitive to outliers.	✓
DBSCAN is ☐ a partitioning-based clustering ☑ a density-based clustering approach.	✓
Two points A and B are density-connected via another point C only if ☐ A and B are a core points ☑ C is a core point ☐ A, B, and C are core points.	✓
Agnes and Diana both refer to hierarchical clustering approaches. ☑ True ☐ False	✓

2) Association Rule Learning: Apriori Algorithm	19 / 20

The following table shows a list of transactions:

T1	Burger, Wrap
T2	Burger, Coke, Fries
T3	Burger, Coke, Fries
T4	Burger, Fries, Wrap
T5	Burger, Coke, Wrap
T6	Coke, Fries

a) Apply the Apriori algorithm with a minimum support of 0.5. Construct for each step the candidate set C_k and the frequent itemset list L_k starting with $k = 1$ until all frequent itemsets are generated. For each step, also list the itemsets that are pruned based on the apriori property and list the itemsets that are pruned due to the application of the minimum support threshold.

minimum support = 0.5 x 6 = 3

C_1 = Burger(5), Coke(3), Fries(4), Wrap(3) ✓

L_1 = Burger(5), Coke(4), Fries(4), Wrap(3) ✓

no itemsets are pruned in this step ✓

C_2 = {Burger, Coke}(3), {Burger, Fries}(3), {Burger, Wrap}(3)

{Coke, Fries}(3), {Coke, Wrap}(1), {Fries, Wrap}(1) ✓

{Coke, Wrap} and {Fries, Wrap} should be pruned due to minimum support threshold. ✓

L_2 = {Burger, Coke}(3), {Burger, Fries}(3), {Burger, Wrap}(3), {Coke, Fries}(3) ✓

C_3 = {Burger, Coke, Fries}(2), {Burger, Coke, Wrap}(1),

{Burger, Fries, Wrap}(1), (✓)

{Burger, Coke, Wrap} and {Burger, Fries, Wrap} will be pruned based on apriori property, and all the the 3 item sets are pruned based on the minim support threshold. (✓)

b) Generate all possible rules from the frequent itemsets and calculate their confidence.

C: {Burger} → {Coke} = $\frac{S\{\text{Burger, Coke}\}}{S\{\text{Burger}\}} = \frac{3}{5}$ ✓

C: {Coke} → {Burger} = $\frac{S\{\text{Burger, Coke}\}}{S\{\text{Coke}\}} = \frac{3}{4}$ ✓

C: {Burger} → {Fries} = $\frac{3}{5}$ ✓, C: {Fries} → {Burger} = $\frac{3}{4}$ ✓

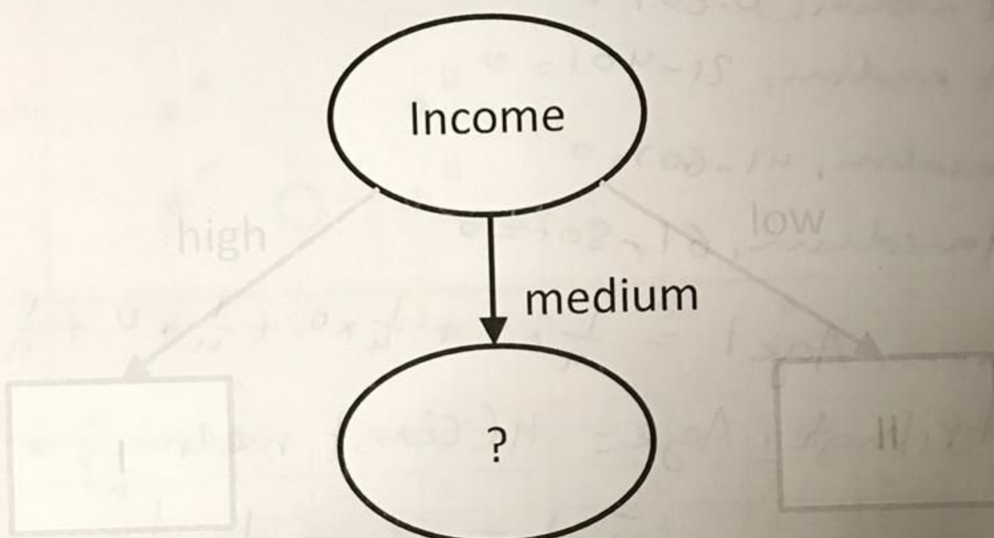
C: {Burger} → {Wrap} = $\frac{3}{5}$ ✓, C: {Wrap} → {Burger} = $\frac{3}{3}$ ✓

C: {Coke} → {Fries} = $\frac{3}{4}$ ✓, C: {Fries} → {Coke} = $\frac{3}{4}$ ✓

3) Classification: Decision Tree 20 / 20

Given the following dataset and partial decision tree:

Age	Car	Income	Class
0-20	no	high	I
0-20	no	medium	II
21-40	yes	medium	I
21-40	no	low	II
41-60	yes	low	II
41-60	no	medium	I
61-80	yes	high	I
61-80	yes	medium	II



Calculate the information gain for the remaining attributes (car, age) to complete the decision tree for the *medium* branch. Decide which attribute should be used for the next split.

$H(\text{class}) = 1$, $H(\text{class} \mid \text{car})$?

$\frac{2}{4}$ no cars , $\frac{2}{4}$ yes cars

↓ $\frac{1}{2}$ class I, $\frac{1}{2}$ class II ; ↘ $\frac{1}{2}$ class I, $\frac{1}{2}$ class II

$H(\text{no car}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = -\frac{2}{4}(-1) - \frac{2}{4}(-1) = 1$

$H(\text{class} \mid \text{yes car}) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = -\frac{2}{4}(-1) - \frac{2}{4}(-1) = 1$

$H(\text{class} \mid \text{Car}) = \frac{1}{2}\times 1 + \frac{1}{2}\times 1 = 1$, $H(\text{class} \mid \text{Car}) = 1$

IG of car attribute $= H(\text{class}) - H(\text{class} \mid \text{Medium, Car}) = 0$

for age :

$\frac{1}{4}$ ~~$\frac{2}{8}$~~ (0-20) ~~$\rightarrow \frac{1}{2}$ class I~~
$\rightarrow \frac{1}{2}$ class II

$\frac{1}{4}$ (21-40) $\rightarrow \frac{1}{2}$ class I
~~$\rightarrow \frac{1}{2}$ class II~~

$\frac{1}{4}$ (41-60) $\rightarrow \frac{1}{2}$ class I
~~$\rightarrow \frac{1}{2}$ class II~~

$\frac{1}{4}$ (61-80) ~~$\rightarrow \frac{1}{2}$ class II~~
$\rightarrow \frac{1}{2}$ class II

~~for medium:~~

H(class| medium, 0-20) = 0
H(class| medium, 21-40) = 0
H(class| medium, 41-60) = 0
H(class| medium, 61-80) = 0

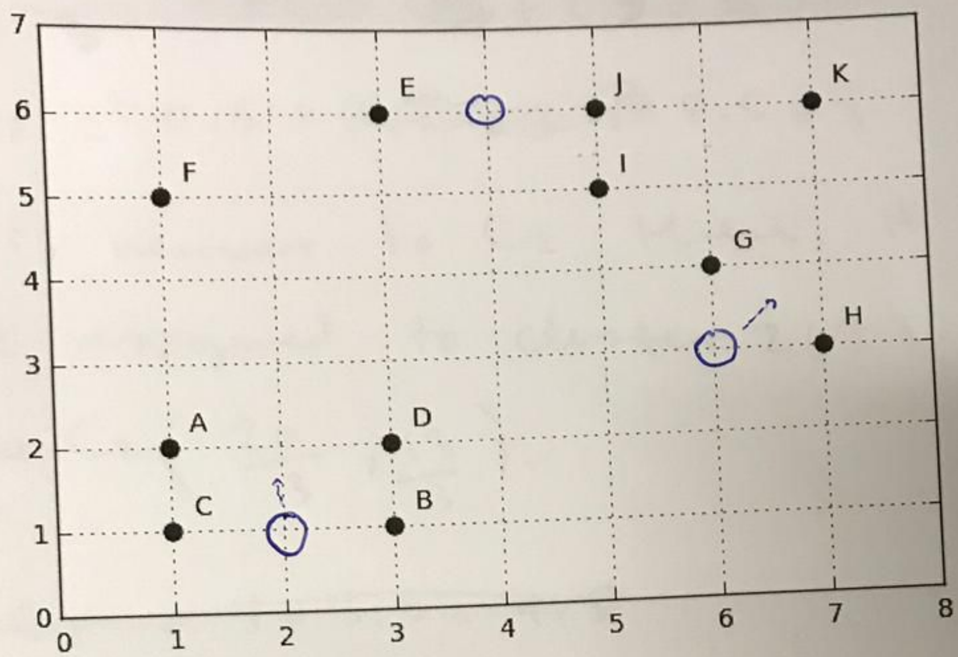
$H(\text{class}|\text{medium, Age}) = \frac{1}{4}\times 0 + \frac{1}{4}\times 0 + \frac{1}{4}\times 0 + \frac{1}{4}\times 0 = 0$

IG of attribute Age = H(class| medium) − H(class| medium, age)
$= 1 - 0 = 1$ ✓

IG of Age larger than IG of Car, then the attribute that should be used on next split is Age. ✓

4) Clustering:k-Means | 5 / 20

The following datapoints are given:



	A	B	C	D	E	F	G	H	I	J	K
x	1	3	1	3	3	1	6	7	5	5	7
y	2	1	1	2	6	5	4	3	5	6	6

Cluster the datapoints with k-Means using Manhattan distance. The initial centroids for k-Means are (2,1), (6,3), and (4,6), i.e. the parameter $k = 3$. Specify for each iteration, to which centroids each point got assigned to and the calculation of the new centroids.

Iteration 1:

C1 = (2,1) : assigned objects = { D, A, C, B } f. $C_1\left(\frac{5}{3}, \frac{4}{3}\right)$ f.

C2 = (6,3) : assigned objects = { G, H } ✓ $C_2\left(\frac{13}{2}, \frac{7}{2}\right)$ ✓

C3 = (4,6) : assigned objects = { E, J, I, F, K } f. $C_3\left(\frac{13}{3}, \frac{17}{2}\right)$ f.

Iteration 2:

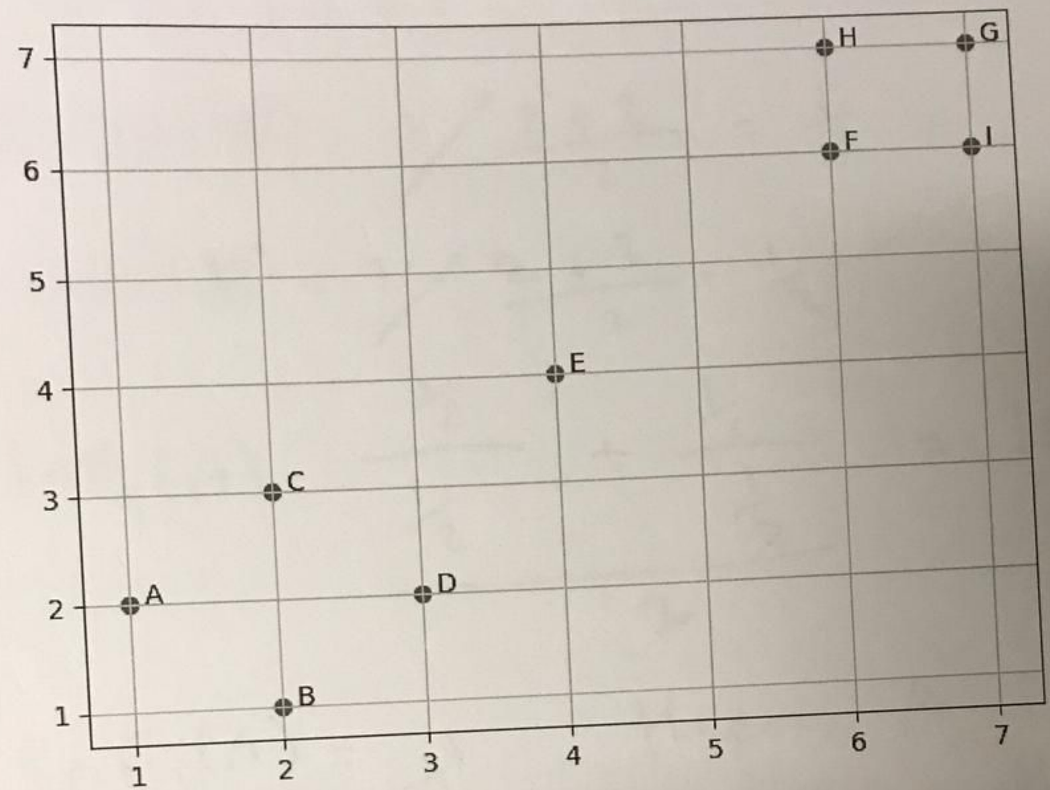
~~assigned to objects~~ D assigned ~~to cluster~~

no assignment

$C_4\left(\frac{13}{2}, \frac{7}{2}\right)$

5) Outlier Detection: Local Outlier Factor — 18 / 20

The following dataset is given. Use Manhattan distance for your calculations.



Use the local outlier factor method (LOF_2) to calculate the scores and decide whether the points A and E are outliers given a threshold of 1.

~~For point E, LOF(E)~~ ~~Path~~

$N_2(E) = \{C, D\}$

$N_2(C) = \{A, D\}$

$N_2(D) = \{C, B\}$

$lrd_2(E) = 1 / \frac{3+3}{2} = \frac{2}{6} = \frac{1}{3}$

$lrd_2(C) = 1 / \frac{2+2}{2} = \frac{1}{2}$

$lrd_2(D) = 1 / \frac{2+2}{2} = \frac{1}{2}$

$$LOF_2(E) = \frac{\frac{\frac{1}{2}}{\frac{1}{3}} + \frac{\frac{1}{2}}{\frac{1}{3}}}{2} = \frac{\frac{3}{2} + \frac{3}{2}}{2} = \frac{3}{2} = \frac{3}{2}$$

$\rightarrow$ ~~E is not outlier~~

$LoF_2(E) = \frac{3}{2} > 1 \rightarrow E$ is outlier

10

5) Outlier Detection: Local Outlier Factor

$N_2(A) = (C, B)$

$lrd_2(C) = \frac{1}{2}$

$N_2(B) = (A, D)$

$lrd_2(A) = 1 / \frac{2+2}{2} = \frac{1}{2}$

$lrd_2(B) = 1 / \frac{2+2}{2} = \frac{1}{2}$

$lof_2(A) = \dfrac{\dfrac{1/2}{1/2} + \dfrac{1/2}{1/2}}{2} = \frac{1+1}{2} = 1$

$lof_2(A) = 1$ then A is not outlier

$lrd_2(C)$?

8 10