Multiple-choice

2p **1a** Consider a DataFrame df with columns 'Name', 'Department', and 'Salary'. Which of the following code snippets will give you the average salary per department?

a
a) df.groupby('Department').mean()['Salary']
b) df.mean().groupby('Department')['Salary']
c) df['Salary'].mean().groupby('Department')

1p **1b** True or false: The number of clusters has to be specified to perform hierarchical clustering.

b
a) True
b) False

1p **1c** True or false: Hierarchical clustering and K-means are equivalent in terms of computational complexity.

b
a) True
b) False

Multiple-choice

1a 2 / 2

Consider a DataFrame df with columns 'Name', 'Department', and 'Salary'. Which of the following co

a) df.groupby('Department').mean()['Salary']
b) df.mean().groupby('Department')['Salary']
c) df['Salary'].mean().groupby('Department')

1b 1 / 1

True or false: The number of clusters has to be specified to perform hierarchical clustering.

a) True
b) False

1c 1 / 1

True or false: Hierarchical clustering and K-means are equivalent in terms of computational complexi

a) True
b) False

Case Study 1: Introduction to AI

Coach is an AI agent which is in charge of picking the best robot football players from a collection of robots. For each robot player, Coach has an internal value representing the score of how many games it lost and won in the past, as well as variables representing the settings in which the robots won or lost. Based on these variables, Coach runs a regression model to predict which robots are the best in the current settings, and then aims to pick those for the team.

2p **2a** Which one of the four main definitions of AI suits Coach best? Explain your answer.

"Thinking rationally". Due to the process behind the Coach AI is rating the players and picking the best, it shows the problem-solving ability. In this view, the Coach applies ~~the~~ to scenarios that require logic and reasoning.

2p **2b** Imagine that Coach needs to pass the Turing test. It is asked the question 'why did you pick these players?'. Should coach answer truthfully? Explain your answer.

I think so. Coach is designed to implement the algorithms that can combine and involve all the information and data from the robot player. Therefore, its answer can be based on the rating process and with the pretraing, the real answer can be reasonable and truthfully in most case as long as the data is plenty enough for its regression.

2p **2c** Which of the four main definitions of AI is the Turing test trying to test? Explain your answer.

"Acting humanly". Because origin Turing test's procedure performs by human asking questions to test whether the one that answers is or is not human. So after all, it's testing its human behavior directly. All the other definations in this test is the support of the ~~action~~ human action as backup to pass the test.

Case Study 1: Introduction to AI

2a 2 / 2

Which one of the four main definitions of AI suits Coach best? Explain your answer.

2b 0 / 2

Imagine that Coach needs to pass the Turing test. It is asked the question 'why did you pick these players?'. Should coach answer truthfully? Explain your answer.

2c 2 / 2

Which of the four main definitions of AI is the Turing test trying to test? Explain your answer.

0003.pdf 0325055604

3p **2d** In AI we can differentiate goals and techniques. Based on the story about Coach, what is this agent's goal, and what the technique? Explain your answer.

The goal of Coach is picking the best 7 robot football players for the team as what it designed to take in charge of.
The technique of Coach is its implementation of regression model of all the variables involved in to predict the best robot football players.

2d 2 / 3

In AI we can differentiate goals and techniques. Based on the story about Coach, what is this agent's goal, an

Case Study 2: Data

3p **3a** How would you create a new column 'total' in a DataFrame (df) that is the sum of columns 'A' and 'B'?

```
# Assume that the pandas is imported and df already exists.
df = df.assign("total" = lambda d: d.A + d.B)
```

3p **3b** Write a Pandas code snippet to join two DataFrames, df1 and df2, on a common column named 'ID', using an left join.

```
df = pd.merge([df1, df2], column = 'ID', join = left)
```

3p **3c** Write a Pandas code snippet which use the iloc() function to select the first three rows and first two columns from a DataFrame named df.

```
df_sub = df.iloc[:3, :2]
```

Case Study 2: Data

3a 3 / 3

How would you create a new column 'total' in a DataFrame (df) that is the sum of columns 'A' and 'B'?

3b 3 / 3

Write a Pandas code snippet to join two DataFrames, df1 and df2, on a common column named 'ID', using an left join.

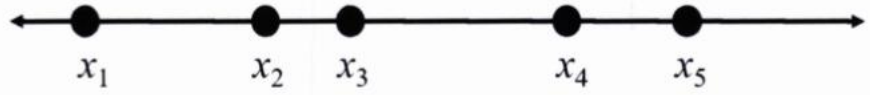
3c 3 / 3

Write a Pandas code snippet which use the iloc() function to select the first three rows and first two columns from a DataFrame named df.

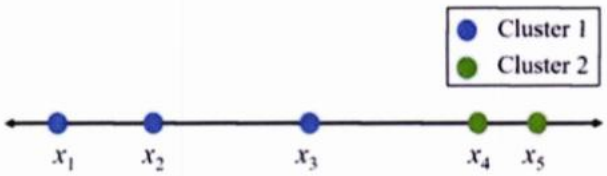
0003.pdf 0325055606

Case Study 3: Clustering

Consider five data points $\{x1, x2, x3, x4, x5\} = \{0, 1.4, 1.6, 3.2, 3.8\}$ as shown in the figure below,



Use the k-means clustering and the Euclidean distance, please answer the following questions:



4a The figure above shows a clustering result, can you give the cluster number k and the final centroid locations of this result?

k = 2 ,

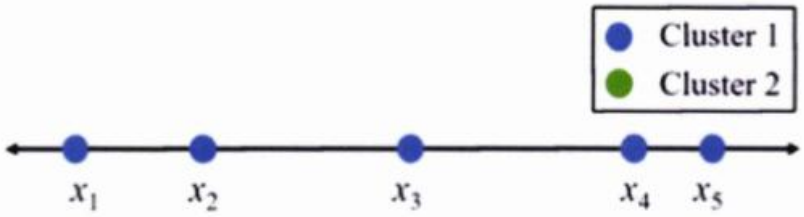
Cluster 1 centroid = 1

Cluster 2 centroid = 3.5

Case Study 3: Clustering

4a 2 / 2

The figure above shows a clustering result, can you give the cluster number k and the final centroid locations of this result?



2p **4b** The figure above shows a clustering result, can you give the cluster number k and the final centroids of this result?

k = 2

Cluster 1 centroid = 2

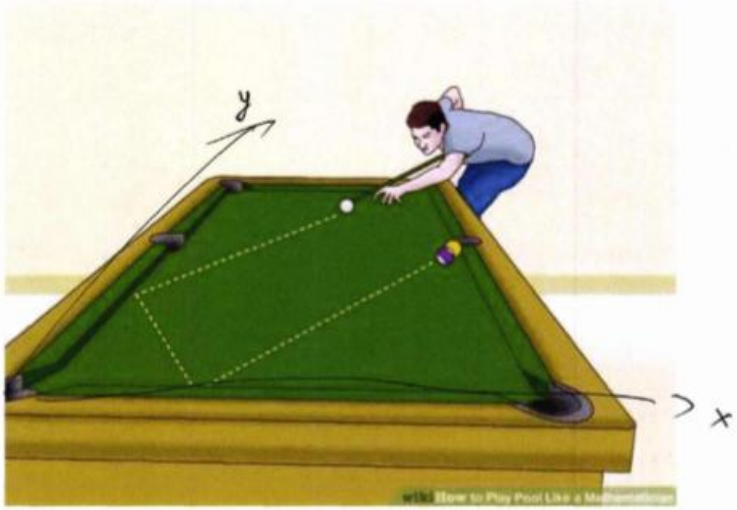
Cluster 2 centroid is unknown but based on the information I have, it should be either < -2.8 or > 7.6

4b 2 / 2

The figure above shows a clustering result, can you give the cluster number k and the final centroids of this result?

Case Study 4: Reinforcement Learning

The picture below shows the task to pocket the yellow ball from the given position in a simulation. The game has always three balls (white, magenta and yellow). The player is only allowed to strike the white ball with the cue stick when all balls are at rest. The strike determines how far and in which direction the white ball rolls. Collisions can push other balls until all balls have come to rest again. The initial ball positions are random, but the player must always play the yellow ball into one of the pockets (i.e. the six holes) in as few strikes as possible. If the magenta or white ball drops into a pocket the player fails and the game ends.



2p **5a** Define a low dimensional continuous state space of the above game.

The state space can be defined as

$S = [w, m, ye, \dot{w}, \dot{m}, \dot{ye}]$

in which $w = [x_1, y_1]$ white ball's location $\dot{w} = [\dot{x}_1, \dot{y}_1]$ velocity

$m = [x_2, y_2]$ magenta ball's location $\dot{m} [\dot{x}_2, \dot{y}_2]$ velocity

$ye = [x_3, y_3]$ yellow ball's location $\dot{ye} [\dot{x}_3, \dot{y}_3]$ velocity

When the ye reach the 6 hole's location, the round will finishes

as long as w and m are not at the 6 hole's location

and the $\dot{w}, \dot{m}, \dot{ye} = 0$, the next strike can be done.

Case Study 4: Reinforcement Learning

5a 1 / 2

Define a low dimensional continuous state space of the above game.

p **5b** Define a low dimensional continuous action space of the above game.

$A = [\dot{s}]$

where $\dot{s} = [\dot{x}_s, \dot{y}_s]$ means the velocity of the stick which gave the white ball the same velocity (it's a tension so the angle is included)

3p **5c** Define a reward function that would yield the optimal behavior described above. Explain why your reward solves the task.

$R = 1, 100 - |\text{closest hole's location} - y_0| - 10 - 1, 100$

where ① if yellow ball reaches the hole, it satisfies the first positive reward and gets 100 point.

② negative reward for the distance between the closet hole's location and the yellow ball's once the velocity of all balls is 0.

③ negative reward of 10 for each ~~strick~~ strike

④ if white or magenta ball's location is at the hole, -100 points.

5b **1** / 1
Define a low dimensional continuous action space of the above game.

5c **3** / 3
Define a reward function that would yield the optimal behavior described above. Explain why your reward solves the task.

Case Study 5: Neural Networks

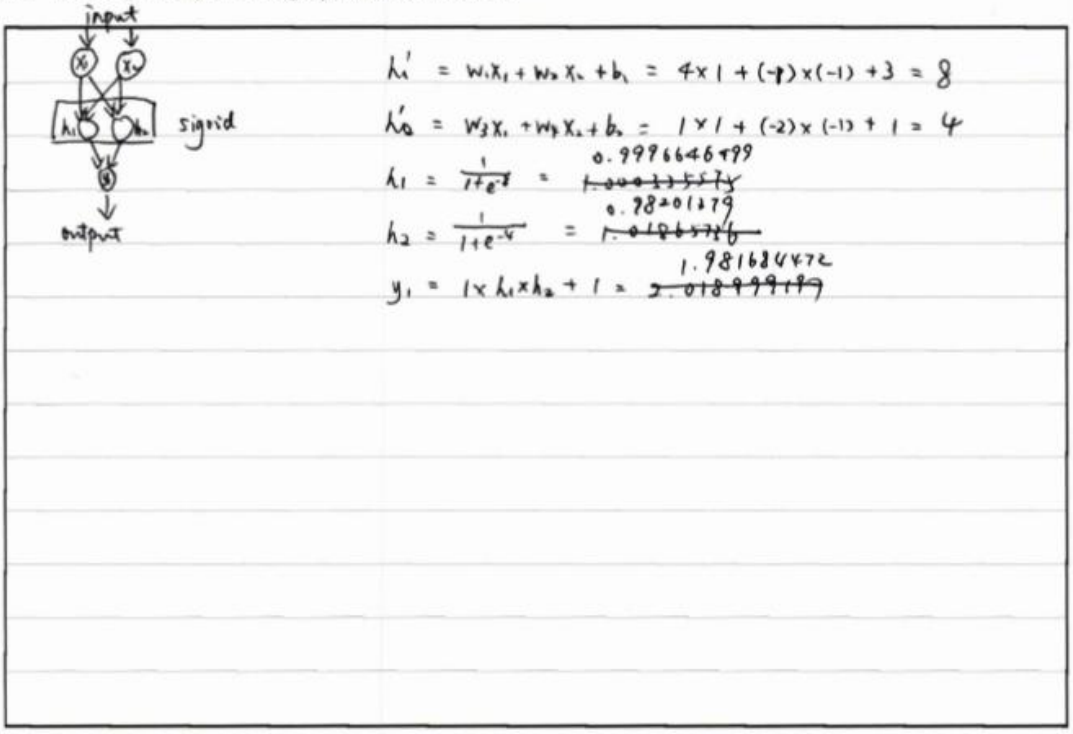
Consider the following function representing a network with 2 input neurons, 1 hidden layer with 2 neurons and sigmoid activation functions σ, and 1 output:

$$\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} \sigma(w_1 x_1 + w_2 x_2 + b_1) \\ \sigma(w_3 x_1 + w_4 x_2 + b_2) \end{pmatrix}$$

$$y_1 = w_5 h_1 h_2 + b_3$$

Assuming $x_1 = 1, x_2 = -1, b_1 = 3, b_2 = 1, b_3 = 1, w_1 = 4, w_2 = -1, w_3 = 1, w_4 = -2, w_5 = 1$, do the following:

6a Build the computation graph of the function.



Case Study 5: Neural Networks

6a 0 / 2

Build the computation graph of the function.

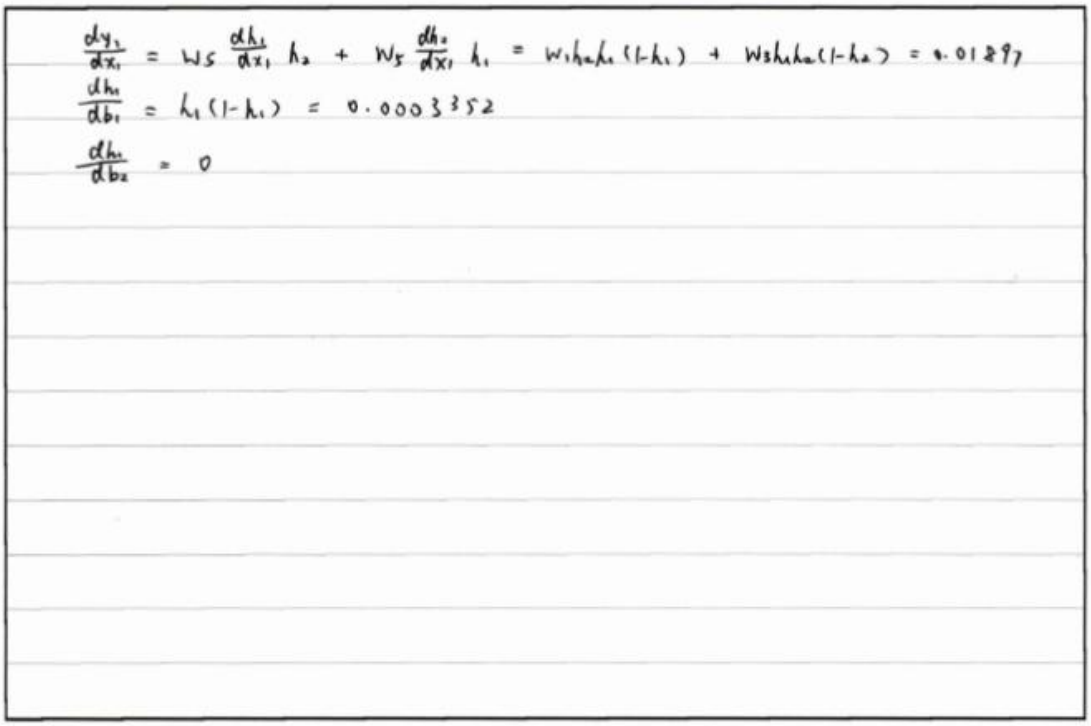
6b

By knowing that

$$\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$$

compute the following derivatives:

$$\frac{dy_1}{dx_1}, \frac{dh_1}{db_1}, \frac{dh_1}{db_2}$$



6b 3 / 3

By knowing that

$$\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$$

compute the following derivatives:

$$\frac{dy_1}{dx_1}, \frac{dh_1}{db_1}, \frac{dh_1}{db_2}$$

Case Study 6: ML for Balance Prediction + Ethics

Banks aim to predict balance for their customers. This has several goals; for example, the bank will try to warn a customer if they think their balance will become negative in the future.

1p **7a** Discuss whether Mean Squared Error (MSE) or Mean Absolute Error (MAE) is more appropriate for predicting customer balances in this context. Provide your reasoning in one or two sentences.

For the case that if the bank is more concerned about the small amount negative that it can not stand for less than 1 euro negative, use MAE cuz MSE will lower the penalty. But as for in the business for the number of going more than 1 euro's prediction, use MSE to enlarge the penalty fuction.

The bank treats the prediction of a negative balance as a classification task. If the balance becomes negative, label is negative, if the balance remains positive, the label is positive.

2p **7b** Provide the definition of Bayes error. Will the Bayes error in this classification task be zero? Explain why or why not.

The Bayes error is the lowest possible error for any classifier of a random outcome and is analogous to the irreducible error. It occurs because of the natural overlap in the distribution of the classes within the feature space. Therefore, in this classification, the Bayes error will not be zero. There can be overlap in the characteristics that the bank uses and also might for a person have have sudden issue that needs lots of money that all the features is the same for another one that still will be positive.

Case Study 6: ML for Balance Prediction + Ethics

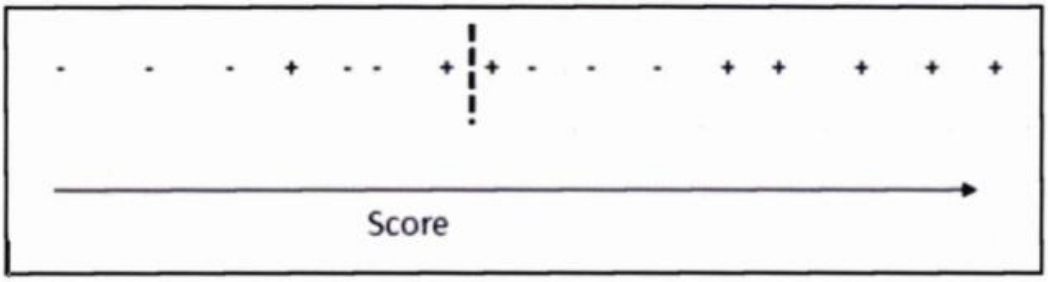
7a 0 / 1

Discuss whether Mean Squared Error (MSE) or Mean Absolute Error (MAE) is more appropriate for predicting customer b in one or two sentences.

7b 2 / 2

Provide the definition of Bayes error. Will the Bayes error in this classification task be zero? Explain why or why not.

The bank has trained a classifier and visualizes some of the objects versus their score. The objects are sorted from left (low score) to right (high score). A "-" indicates negative class, a "+" indicates the positive class.



7c Give the confusion matrix. Clearly indicate what the rows and columns mean.

Actual \ Predicted	positive	negative
positive	6	2
negative	3	5

(1,1) is true positive > truely, accurately predicted

(2,2) is true negative

(1,2) is false negative, which where it should be positive but predicted negative

(2,1) is false positive, where it should be negative but predicted positive

7c 2 / 2

Give the confusion matrix. Clearly indicate what the rows and columns mean.

Three features for this classification task are: age, income, amount of savings.

2p **7d** Explain how these features should be pre-processed for a classifier like logistic regression or SVM and why.

Second
First, they need to be scaled to the same scaling using min-max, or other algorithms; First, they need to be checked with outliers. If so, the outliers need to be dealed properly like deleted or replaced with normally largest number like 99% pentile; third, once find the strongly, significaly relationship between each other like maybe income and savings, they need to be either merged or delete one. (~~Before~~ if there are missing values, delete the row or fill with the mean or the upper)

3p **7e** The bank is trying to optimize the hyperparameters of the SVM for this classification task. The bank has settled on the F1 score as evaluation measure. How should the bank split its data in order to get a fair estimate of the F1 score for the tuned model?

Use gene index or information ratio or other splitting criteria like entropy to choose the feature with the most scores (chaos) to split its data.

7d 1 / 2
Explain how these features should be pre-processed for a classifier like logistic regression or SVM and why.

7e 0 / 3
The bank is trying to optimize the hyperparameters of the SVM for this classification task. The bank has settle bank split its data in order to get a fair estimate of the F1 score for the tuned model?

The task of balance prediction is also used by the bank to decide whether a customer is able to apply for a loan or not.

1p **7f** Give one reason why a decision tree could be an attractive choice for model for the bank for this task.

Because decision tree is also a good approach to model the ~~ex~~ dummy y with ~~and~~ the mutiple X as layers of tree like using ID3 to classifer the target. XGBoost or Randomforest can be used as more advanced decision tree dealing with error and stability.

In addition the bank is considering to use a machine learning system to detect money laundering and terrorism financing among its customers. The system has as input a transaction, consisting of the personal details and bank accounts of sender and receiver, the amount and description of the transaction. In order to ensure that the personal information is complete the Dutch banks will share their client data for the sole purpose of detecting fraud. As output the system then generates a risk score that a transaction is an instance of money laundering or terrorism financing. The banks automatically block transactions pending an investigation (carried out by employees of the bank) above a certain threshold.

2p **7g** What kind of privacy is most at stake here and why?

Privacy II : Privacy as the proper flow of personal information. This is for controlling how personal ~~pf~~ information is shared, who has access to ?t, and the conditions under which it is disseminated. In this case, the Dutch banks share their client personal details and accounts data which is the major process of its system. The contextual flow of information need to be concerned.

7f 0 / 1

Give one reason why a decision tree could be an attractive choice for model for the bank for this task.

7g 2 / 2

What kind of privacy is most at stake here and why?

2p **7h** What kind of bias is likely in this case? Give reasons as to why this is the most likely form(s) of bias.

Technical bias. The decision of bank is automatically made by the system based on data and model for the prediction. In the process, the not properly preparing data or the limitation of hardware might cause the inacurate result. Also, the emergent bias. the model and parameters of it are already designed and once some ~~thing~~ changed happened in the society or the policy, it causes the bias.

2p **7i** What statistical definition of fairness would be the most suitable to enforce in this case? Motivate your choice.

Equality of opportunity. Not overall acceptance, but percentage of correct positive decisions should be the same, which help to reduce the problems in fairness of the banking system. Don't blindly follow the statistics. User feedback mechanism. Incorporate this for user feedback and ~~some~~ contestability. This leads to continuous improvement in the model's performance and fairness

7h 1.5 / 2
What kind of bias is likely in this case? Give reasons as to why this is the most likely form(s) of bias.

7i 1.5 / 2
What statistical definition of fairness would be the most suitable to enforce in this case? Motivate your choi

4p **7j** Identify two value conflicts that are likely to play a role in this case. Briefly mention the conflicts and suggest a trade-off strategy that you consider appropriate for dealing with the conflict.

① Pra Privacy & fuctionality conflict : To enhance the accuracy of the system and the personalo data using could conflict with the users' privacy expectations.
Trade-off strategy : Implement "satisficing" approach to ensure the system meets an "adequate good" level of privacy protection , involving privacy-enhancing tech like local processing or made users' consent before using system .

② Safty & users Convenience : the block of transactions automatically ~~using~~ might conflict once mistakenly done and that user is emergently in need .
Trade-Off strategy : Employ an innovation approach to find new tech or methods that enhance both ~~the~~ users' convenience and the bank's safety . Designing flexible interation protocals for the issue reporting and new algs to improve error detection .

7j 3 / 4

Identify two value conflicts that are likely to play a role in this case. Briefly mention the conflicts and suggest a trade-
dealing with the conflict.

2p **7k** What kind of transparency would the bank employees need? What would the customers of the bank need? Explain your answer.

The bank employees need performance and consistency transparency so that the accuracy, false positive rate, and any limitations of the alg in predicting addition risk and also provides consistent output given similar input, which allow them to better judge and trust the output. The customers need value transparency. Clearly communicate the objects and values the alg is designed to optimize. This will allow them to understand the motivations and trade-off that underlie the alg's decision and help fairness and consideration.

7k 2 / 2

What kind of transparency would the bank employees need? What would the customers of the bank neec