2024-02-27 Mid-term Exam

5072DASC6Y Data Science 23/24 (2.1) · 6 exercises · 25.0 points

General Data Science Questions

2.0 points · 2 questions

1 Given the following tables A and B, what will be the result after performing an outer join by User ID?
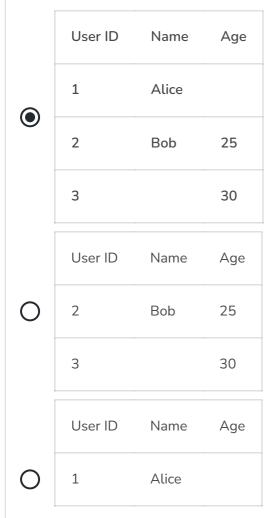
Table A

User ID	Name
1	Alice
2	Bob

Table B:

User ID	Age
2	25
3	30

1.0 point · Multiple choice · 4 alternatives

◉
User ID	Name	Age
1	Alice	
2	Bob	25
3		30

○
User ID	Name	Age
2	Bob	25
3		30

○
User ID	Name	Age
1	Alice	

User ID	Name	Age
2	Bob	25

○

User ID	Name	Age
2	Bob	25

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

2 Which of the following would be the best way to deal with missing data when you do not have a lot of data available when building a regression model, and you know that your data is MNAR (Missing Not At Random)?

1.0 point · Multiple choice · 4 alternatives

◉ **Find more data about the causes for the MNAR missingness**

○ Remove all observations with missing data

○ Use imputation techniques to model and fill in the missing values

○ Replace the missing values with value -1

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Regression and Classification Questions

9.0 points · 9 questions

3 Which of the following statements about overfitting is **FALSE**?

1.0 point · Multiple choice · 4 alternatives

- ⦿ We can combat overfitting by increasing the size of the test set.

- ○ Overfitting usually happens when using a very complex model.

- ○ Using cross-validation can help us tackle overfitting.

- ○ Overfitting means fitting the training data extremely well but performing poorly on the validation or test data.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

4 Which of the following pairs is used to calculate recall?

1.0 point · Multiple choice · 4 alternatives

- ⦿ True Positive and False Negative

- ○ True Positive and False Positive

- ○ True Negative and False Positive

- ○ True Negative and False Negative

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

5 Suppose we fit a binary classification model to predict if a bad smell event will happen in the city for the next 6 hours. The positive label means that there will be smell events, and the negative label means no events. Which of the following statements is **FALSE**?

1.0 point · Multiple choice · 4 alternatives

- ⦿ **If the model predicts that there would be NO event, but it turns out that the model made a wrong prediction, this is called a True Negative.**

- ○ If the model predicts that there would be an event, but it turns out there is nothing happening, this is called a False Positive.

- ○ If the recall of the model is very high, it means that the model catches almost all the smell events and does not miss them.

- ○ If the precision of the model is very low, it means that the model is not very reliable when it predicts that there would be smell events in the future.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

6 What is the purpose of using a loss function in classification or regression?

1.0 point · Multiple choice · 4 alternatives

- ⦿ **To measure the error between the prediction and the ground truth**

- ○ To regularize the model parameters

- ○ To reduce the number of features

- ○ To increase the accuracy of the model on the training set

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

7 You already have a linear classifier $f(x_1, x_2)$ that can predict a binary label y (yes/no) using two features, x_1 and x_2. Which of the following is the best way to use the linear classifier to determine the label of a new incoming data point with the values of these two features? For example, these two features could be the number of special characters and digits in an email message, and the labels could be whether the email is spam or not.

1.0 point · Multiple choice · 4 alternatives

◉ Calculate $f(x_1, x_2)$ and check if it is larger or smaller than zero.

○ Calculate $f(x_1, 0)$ and $f(0, x_2)$ and check which value is larger.

○ Calculate the average of $f(x_1, 0)$ and $f(0, x_2)$, and then check if the average is larger or smaller than zero.

○ Calculate $f(x_1, x_2)$ and check if it is larger or smaller than the average of x_1 and x_2.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

8 What is the goal of minimizing the sum of squared errors in linear regression?

1.0 point · Multiple choice · 4 alternatives

◉ To find the line that best fits the data by minimizing the distance between the predicted and true values.

○ To find the optimal values of the input variables (features) that minimize the output variable (responses).

○ To reduce the variance in the model and prevent overfitting.

○ To ensure that the residuals (the differences between predicted and true values) are normally distributed.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

9 Which of the following is the order that best describes the process of training a linear classifier for detecting whether an email is spam or not?
- A: Initialize a linear decision boundary randomly (e.g., a line, plane, or hyperplane)
- B: Take a separate set of emails with labels
- C: Update the decision boundary iteratively using an optimization algorithm with an error metric
- D: Transform the spam and non-spam messages into feature vectors
- E: Normalize the feature vectors to ensure they are on a similar scale
- F: Evaluate the classifier's performance using precision, recall, and F1 score
- G: Randomly shuffle the labels to test the classifier's robustness against noise

1.0 point · Multiple choice · 4 alternatives

- ● DEACBF

- ○ ADGCBF

- ○ DAEGC

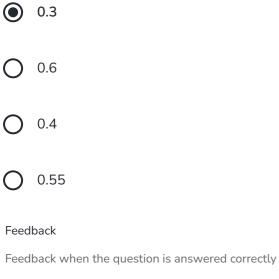- ○ DEACF

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

10 Suppose we fit a simple linear regression model F using the training data with one feature X and one true response Y. We then use the model to output prediction Z=F(X). The coefficient of determination (R-squared) is equal to the square of:

1.0 point · Multiple choice · 4 alternatives

◉ **Pearson correlation coefficient between the true response Y and the predictions Z.**

○ Pearson correlation coefficient between the true response Y and the feature X.

○ Pearson correlation coefficient between the predictions Z and the feature X.

○ The coefficient of feature X.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

11 Suppose that we fit a binary classification model in identifying spam and ham (i.e., non-spam). Spam is the positive label, and ham is the negative label. The following shows the evaluation result of the model. What is the recall of the model based on the evaluation result?

- 70 samples are predicted as ham, but it turns out that they are spam in reality
- 80 samples are predicted as ham, and they are indeed ham in reality
- 30 samples are predicted as spam, and they are indeed spam in reality
- 20 samples are predicted as spam, but it turns out that they are not spam in reality

1.0 point · Multiple choice · 4 alternatives

◉ 0.3

○ 0.6

○ 0.4

○ 0.55

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Structured Data Processing Question

7.0 points · 7 questions

---

12 What is the main purpose of computing feature importance?

1.0 point · Multiple choice · 4 alternatives

- ◉ To determine which features are most important in making predictions

- ◯ To increase the number of features used in the model

- ◯ To improve the accuracy of the model on the training set

- ◯ To prevent overfitting of the model

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

13 Which of the following statements about Random Forest is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- ◉ Random Forest uses randomly selected features and bootstrapped samples (i.e., sample with replacement).

- ◯ Random Forest is more likely to overfit the data than the Decision Tree model.

- ◯ Random Forest contains multiple Decision Tree models, and the best tree is used for performing the task.

- ◯ Random Forest contains multiple Decision Tree models that are trained identically using the same set of features.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

14 Suppose we flip a coin (with two sides) many times and we compute the entropy. Which of the following statements is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

◉ Entropy intuitively means the averaged surprise when we flip the coin.

○ Entropy reaches the minimum when the coin is fair, meaning two sides have equal probability.

○ If we change the probability of one side of the coin (to make it appear more frequently or less frequently), entropy is not sensitive to this change in probabilities.

○ Entropy is always one in this case because the coin has only two sides.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

15 Suppose we want to classify if an image contains a banana or mango by using a Decision Tree model. We have 3 green banana images, 3 yellow banana images, 1 green mango image, and 1 yellow mango image. The Decision Tree uses entropy as the node-splitting strategy. Recall that Information Gain is the difference between the parent node's entropy and the leaf nodes' averaged entropy. What is the Information Gain after we ask the question, "is the fruit color yellow or not"?

1.0 point · Multiple choice · 4 alternatives

◉ 0

○ 0.3333

○ 0.5

○ 1

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

16 What is the main advantage of using an ensemble of decision trees, such as a Random Forest, over a single decision tree in a classification or regression problem? Recall that errors of the model that we trained can be decomposed into bias, variance, and noise.

1.0 point · Multiple choice · 4 alternatives

- (**●**) **Reduced variance in the error**

- (○) Reduced bias in the error

- (○) There is not really an advantage

- (○) Reduced noise in the error

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

17 Which of the following best describes the reason why using misclassification error as the node-splitting strategy when training a decision tree model is not a good idea?

1.0 point · Multiple choice · 4 alternatives

- (**●**) **Misclassification error is less sensitive to changes after splitting a node than other measures such as entropy, making it less effective when deciding which feature to use.**

- (○) Misclassification error directly increases the computational complexity of training decision trees, making the process significantly slower compared to other methods such as entropy.

- (○) Using misclassification error as the node splitting strategy can lead to overfitting the training data, as it always creates more complex trees with a higher number of splits.

- (○) Misclassification error is more prone to being affected by imbalanced datasets, leading to biased trees that favor the majority class.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

18 Suppose we want to train a Decision Tree based on the following dataset to predict whether Alex will go out or not. We use the misclassification error as the strategy when splitting a node. Which feature will the Decision Tree pick to split the first node?

Weather	Feeling	Wind	Time	Going out?
sunny	cold	calm	daytime	yes
rainy	warm	calm	nighttime	no
rainy	warm	windy	daytime	yes
sunny	warm	windy	daytime	no
rainy	warm	calm	daytime	no

1.0 point · Multiple choice · 4 alternatives

🔘 Feeling

⭕ Weather

⭕ Wind

⭕ Time

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Coding Question

5.0 points · 5 questions

19 Which of the following best explains the purpose of the following Python Pandas code?

```
pandas.merge_ordered(A, B, on=A.index.name, how="outer", fill_method=None)
```

1.0 point · Multiple choice · 4 alternatives

◉ **The code merges two pandas DataFrames A and B in order by the index of A, using an outer join, without filling missing values.**

○ The code sorts A based on its index and then appends B to it, without filling missing values.

○ The code performs an inner join on A and B based on their indexes, using the specified fill_method to handle None values.

○ The code concatenates A and B vertically, aligning rows based on the index name of A, using the specified fill_method to handle None values.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

20   Suppose we have a pandas data frame D with 100 rows and two columns (C1 and C2). Column C1 has no missing data, and column C2 has 25% missing data. We want to sum up all valid items in column C2 and get a single integer (not a pandas.Series or pandas.DataFrame). Which of the following code produces the desired output? For example, if D looks like the table below, the code should output only one single integer 27.

C1	C2
3	NaN
15	3
2	4
7	20

1.0 point · Multiple choice · 4 alternatives

🔘 **D.dropna().sum()["C2"]**

⭕ D.drop("C2", axis=1).sum().iloc[0]

⭕ D.dropna().iloc[2].sum()

⭕ D.sum()["C1"]

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

21 Suppose we have a pandas data frame D with the index column and also one value column. The value column has the name "smell". The index column contains pandas datetime objects. The "smell" column contains ratings of how bad the smell is at the corresponding timestamp. What is the following code doing?

```
D.resample("120Min", label="right").mean()
```

1.0 point · Multiple choice · 4 alternatives

◉ **Compute the average smell values from the previous two hours.**

○ Compute the total smell values from the previous two hours.

○ Compute the average smell values from the future two hours.

○ Compute the total smell values from the future two hours.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

---

22 You want to extract the year from a text column in a table using Python Pandas, and you want to use regular expressions (i.e., the pandas.Series.str.extract function). Which of the following regular expressions can achieve the task?

1.0 point · Multiple choice · 4 alternatives

◉ [0-9]{4}

○ [a-zA-Z]{4}

○ {4}[1-10]

○ {4}[0123456789]

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

23 Given a pandas series object S with integers or floats. The index of each row means the time steps. What is the following code doing?

```
S.rolling(3, min_periods=1, closed="right").sum()
```

1.0 point · Multiple choice · 4 alternatives

◉ **For each row in series S, compute the sum of the current value and the previous two values from the two preceding time steps. Then, store the sum in the current row.**

○ Compute the cumulative sum of series S, where the value in each row is the sum of all values up to and including that value.

○ Replace each value in series S with the sum of the current value and the maximum value in a window of size 3. The window is moved one position ahead at a time.

○ For each row in series S, compute the difference between each value and the average of the previous two values from the two preceding time steps.

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Deep Learning Questions

2.0 points · 2 questions

24 What is the main purpose of using an activation function in a neural network?

1.0 point · Multiple choice · 4 alternatives

⦿ To introduce non-linearity into the model

◯ To scale the input features to a consistent range

◯ To reduce the dimensionality of the input data

◯ To regularize the model parameters

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

25 Given an error function $f(x^3)$, what is the value of x after making three gradient descent updates $(x_0 \to x_1 \to x_2 \to x_3)$ using the initial point $x_0 = 10$ and learning rate 0.1? The derivative of x^3 is $3x^2$.

1.0 point · Multiple choice · 4 alternatives

⦿ -6020

◯ 81640

◯ 236379867649310

◯ -38598736831290

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly