2023-02-28 Mid-term Exam

5072DASC6Y Data Science 23/24 (2.1) · 7 exercises · 25.0 points

General Data Science Questions

6.0 points · 6 questions

1 If data frame A and B both have no missing data, which of the following operations will definitely **NOT** produce missing data?

1.0 point · Multiple choice · 4 alternatives

- (●) Inner join — 1.0
- ( ) Left join — 0.0
- ( ) Right join — 0.0
- ( ) Outer join — 0.0

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 pairs is used to calculate precision?

1.0 point · Multiple choice · 4 alternatives

- (●) True Positive and False Positive — 1.0
- ( ) True Negative and False Positive — 0.0
- ( ) True Positive and False Negative — 0.0
- ( ) True Negative and False Negative — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

3 Suppose we fit a binary classification model to predict if a bad smell event will happen in the city for the next 8 hours. The positive label means that there will be smell events, and the negative label means no events. Which of the following statements is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (selected) **If the model predicts that there would be an event, but it turns out there is nothing happening, this is called a False Positive.** 1.0
- 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 Positive. 0.0
- If the precision of the model is very high, it means that the model catches almost all the smell events and does not miss them. 0.0
- If the recall 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. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

4 What is the main purpose of computing feature importance?

1.0 point · Multiple choice · 4 alternatives

- (selected) **To determine which features are most important in making predictions** 1.0
- To increase the number of features used in the model 0.0
- To improve the accuracy of the model on the training set 0.0
- To prevent overfitting of the model 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

5 Which of the following practices is **recommended** in the data science pipeline?

1.0 point · Multiple choice · 4 alternatives

- (●) **Visualizing and exploring data in addition to using descriptive statistics** 1.0
- ( ) Sticking to a linear data science pipeline that starts from problem framing and data preparation to model building 0.0
- ( ) Assuming that someone else has already framed the data science problem 0.0
- ( ) Using the same modeling technique to deal with all different types of data 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

6 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.** 1.0
- ( ) Entropy reaches the minimum when the coin is fair, meaning two sides have equal probability. 0.0
- ( ) 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. 0.0
- ( ) Entropy is always one in this case because the coin has only two sides. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Decision Tree and Random Forest Questions

4.0 points · 4 questions

7 Which of the following is a common metric that is used to evaluate the quality of a node split in a Decision Tree model?

1.0 point · Multiple choice · 4 alternatives

- (●) Entropy — 1.0
- ( ) R-squared — 0.0
- ( ) Mean squared error — 0.0
- ( ) F1 Score — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

8 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). — 1.0
- ( ) Random Forest is more likely to overfit the data than the Decision Tree model. — 0.0
- ( ) Random Forest contains multiple Decision Tree models, and the best tree is used for performing the task. — 0.0
- ( ) Random Forest contains multiple Decision Tree models that are trained identically using the same set of features. — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

9 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

- (selected) Reduced variance in the error — 1.0
- Reduced bias in the error — 0.0
- There is not really an advantage — 0.0
- Reduced noise in the error — 0.0

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 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
rainy	cold	windy	daytime	no
rainy	warm	calm	daytime	no

1.0 point · Multiple choice · 4 alternatives

- (selected) Weather 1.0
- Feeling 0.0
- Wind 0.0
- Time 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Neural Networks and Deep Learning Questions

4.0 points · 4 questions

11 Which of the following about the Gradient Descent algorithm is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (selected) The learning rate determines how large or small the step will be when updating model parameters. — 1.0
- Gradient Descent is used to minimize the error/cost in regression problems only. — 0.0
- Gradient Descent is guaranteed to find the global minimum. — 0.0
- The mini-batch size determines the number of iterations required to find the optimal parameters in the neural network. — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

12 What is the purpose of regularization when performing Gradient Descent?

1.0 point · Multiple choice · 4 alternatives

- (selected) To prevent the algorithm from overfitting the training data — 1.0
- To speed up the algorithm and save computer memory — 0.0
- To reduce the number of features in the dataset — 0.0
- To schedule the learning rate appropriately — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

13 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** 1.0
- ( ) To scale the input features to a consistent range 0.0
- ( ) To reduce the dimensionality of the input data 0.0
- ( ) To regularize the model parameters 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

14 Which of the following statements about the backpropagation algorithm when training neural networks is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (●) **Backpropagation iteratively updates the weights in previous layers in the neural network.** 1.0
- ( ) Backpropagation is only used to train the final layer of a neural network. 0.0
- ( ) Backpropagation is used to transform the weighted sum of the input non-linearly. 0.0
- ( ) Backpropagation is used to add new layers to a neural network. 0.0

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

6.0 points · 6 questions

15 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** 1.0
- ( ) To regularize the model parameters 0.0
- ( ) To reduce the number of features 0.0
- ( ) To increase the accuracy of the model on the training set 0.0

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 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.** 1.0
- ( ) To find the optimal values of the input variables (features) that minimize the output variable (responses). 0.0
- ( ) To reduce the variance in the model and prevent overfitting. 0.0
- ( ) To ensure that the residuals (the differences between predicted and true values) are normally distributed. 0.0

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 statements about overfitting is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (selected) **Overfitting usually happens when using a very complex model.** 1.0
- Most of the time, we can deal with overfitting by removing outliers. 0.0
- We can combat overfitting by increasing the size of the test set. 0.0
- Overfitting means fitting the test data extremely well. 0.0

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 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

- (selected) **Pearson correlation coefficient between the true response Y and the predictions Z.** 1.0
- Pearson correlation coefficient between the true response Y and the feature X. 0.0
- Pearson correlation coefficient between the predictions Z and the feature X. 0.0
- The coefficient of feature X. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

19 In the setting of fitting a model Y=F(X), Which of the following statements about the coefficient of determination (R-squared) is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (●) **R-squared increases as we add more features X.** 1.0
- ( ) R-squared is a good evaluation metric for classification. 0.0
- ( ) R-squared intuitively means the unexplained variation for the response variable Y. 0.0
- ( ) A bad R-squared means that there is no pattern in the data. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

20 Which of the following about data modeling is **TRUE**?

1.0 point · Multiple choice · 4 alternatives

- (●) **Classification is used to predict categorical labels, and regression is used to predict continuous values.** 1.0
- ( ) R-squared is a common evaluation metric for classification models. 0.0
- ( ) We can just use the training data to evaluate if the model will work well. 0.0
- ( ) When using Random Forest, we can usually put raw data into the model without feature engineering. 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Calculation Related Questions

2.0 points · 2 questions

21 Suppose we want to classify if an image contains a banana or mango by using a Decision Tree model. We have 1 green banana image, 1 yellow banana image, 3 green mango images, and 3 yellow mango images. 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"?

The formula of entropy H is given below:
$H = p_1 * log_2(1/p_1) + p_2 * log_2(1/p_2)$

1.0 point · Multiple choice · 4 alternatives

- (●) 0 — 1.0
- ( ) 0.33 — 0.0
- ( ) 0.5 — 0.0
- ( ) 1 — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

22 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.

- 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
- 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

What is the precision of the model based on the evaluation result?

1.0 point · Multiple choice · 4 alternatives

- (x) 0.6 — 1.0
- ( ) 0.3 — 0.0
- ( ) 0.4 — 0.0
- ( ) 0.55 — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

Coding Questions

3.0 points · 3 questions

23 Assume you have a pandas DataFrame named D containing time series data with inconsistent frequency. Which of the following code resamples the data to a frequency of one hour?

1.0 point · Multiple choice · 4 alternatives

- (●) D.resample("60Min") 1.0
- ( ) D.resample("60H") 0.0
- ( ) D.resample("1Min", '1H') 0.0
- ( ) D.resample("1Min").asfreq() 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

24 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, which is a sum of 3, 4, and 20.

C1	C2
1	NaN
2	3
2	4
10	20

1.0 point · Multiple choice · 4 alternatives

- (●) D.dropna().sum()["C2"] — 1.0
- ( ) D.drop("C2", axis=1).sum() — 0.0
- ( ) D["C1"].dropna().sum() — 0.0
- ( ) D.groupby("C2").sum() — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly

25 Suppose we have a pandas data frame D with two columns (Smell, and Zipcode). The data frame contains the complaints of bad smell that citizens submitted in Pittsburgh. The "Smell" column contains ratings of how bad the smell is at the corresponding timestamp. The "Zipcode" column means the zip code where the citizen submitted the report. We want to know the average number of smell ratings for ONLY the zip code 15213. Which of the following code produces the desired output? For example, if D looks like the table below, the code should output 4, which is the average of 3 and 5.

Smell	Zipcode
1	15208
2	15202
3	15213
4	15222
5	15213

1.0 point · Multiple choice · 4 alternatives

- (selected) D[D["Zipcode"]=="15213"].mean()["Smell"] — 1.0
- D["Smell"].groupby("Zipcode").mean()["15213"] — 0.0
- D.groupby("Zipcode").mean()["Smell"].iloc("15213") — 0.0
- D.mean()["Smell"].groupby("Zipcode")["15213"] — 0.0

Feedback

Feedback when the question is answered correctly

Feedback when the question is answered partially correctly

Feedback when the question is answered incorrectly