[Mock1Mid] Mid-term Exam

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

[Mock1Mid] General Data Science Questions #32308818

6 pts · Last updated 22 Jan, 18:52 · Part of flow 2 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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

Left join	0 pts
Inner join	1 pt
Outer join	0 pts
Right join	0 pts

1 PT · MULTIPLE CHOICE

Which of the following pairs is used to calculate precision?

True Negative and False Negative	0 pts
True Positive and False Negative	0 pts
True Negative and False Positive	0 pts
True Positive and False Positive	1 pt

1 PT · MULTIPLE CHOICE

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**?

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 pts
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 pt
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 pts
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 pts

1 PT · MULTIPLE CHOICE

What is the main purpose of computing feature importance?

To determine which features are most important in making predictions	1 pt
To increase the number of features used in the model	0 pts
To improve the accuracy of the model on the training set	0 pts
To prevent overfitting of the model	0 pts

1 PT · MULTIPLE CHOICE

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

Using the same modeling technique to deal with all different types of data	0 pts
Assuming that someone else has already framed the data science problem	0 pts
Sticking to a linear data science pipeline that starts from problem framing and data preparation to model building	0 pts
Visualizing and exploring data in addition to using descriptive statistics	1 pt

1 PT · MULTIPLE CHOICE

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

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 pts
Entropy is always one in this case because the coin has only two sides.	0 pts
Entropy reaches the minimum when the coin is fair, meaning two sides have equal probability.	0 pts
Entropy intuitively means the averaged surprise when we flip the coin.	1 pt

[Mock1Mid] Decision Tree and Random Forest Questions #32308824

4 pts · Last updated 22 Jan, 18:52 · Part of flow 3 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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

Entropy	1 pt
R-squared	0 pts
Mean squared error	0 pts
F1 Score	0 pts

1 PT · MULTIPLE CHOICE

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

Random Forest contains multiple Decision Tree models that are trained identically using the same set of features.	0 pts
Random Forest uses randomly selected features and bootstrapped samples (i.e., sample with replacement).	1 pt
Random Forest is more likely to overfit the data than the Decision Tree model.	0 pts
Random Forest contains multiple Decision Tree models, and the best tree is used for performing the task.	0 pts

1 PT · MULTIPLE CHOICE

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.

Reduced variance in the error	1 pt
There is not really an advantage	0 pts
Reduced bias in the error	0 pts
Reduced noise in the error	0 pts

1 PT · MULTIPLE CHOICE

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

Wind	0 pts
Time	0 pts
Feeling	0 pts
Weather	1 pt

[Mock1Mid] Neural Networks and Deep Learning Questions #32308828

4 pts · Last updated 22 Jan, 18:52 · Part of flow 4 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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

The mini-batch size determines the number of iterations required to find the optimal parameters in the neural network.	0 pts
The learning rate determines how large or small the step will be when updating model parameters.	1 pt
Gradient Descent is used to minimize the error/cost in regression problems only.	0 pts
Gradient Descent is guaranteed to find the global minimum.	0 pts

1 PT · MULTIPLE CHOICE

What is the purpose of regularization when performing Gradient Descent?

To prevent the algorithm from overfitting the training data	1 pt
To schedule the learning rate appropriately	0 pts
To reduce the number of features in the dataset	0 pts
To speed up the algorithm and save computer memory	0 pts

1 PT · MULTIPLE CHOICE

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

To scale the input features to a consistent range	0 pts
To reduce the dimensionality of the input data	0 pts
To regularize the model parameters	0 pts
To introduce non-linearity into the model	1 pt

1 PT · MULTIPLE CHOICE

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

Backpropagation iteratively updates the weights in previous layers in the neural network.	1 pt
Backpropagation is only used to train the final layer of a neural network.	0 pts
Backpropagation is used to transform the weighted sum of the input non-linearly.	0 pts
Backpropagation is used to add new layers to a neural network.	0 pts

[Mock1Mid] Regression and Classification Questions #32308832

6 pts · Last updated 22 Jan, 18:52 · Part of flow 5 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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

To regularize the model parameters	0 pts
To reduce the number of features	0 pts
To increase the accuracy of the model on the training set	0 pts
To measure the error between the prediction and the ground truth	1 pt

1 PT · MULTIPLE CHOICE

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

To find the line that best fits the data by minimizing the distance between the predicted and true values.	1 pt
To reduce the variance in the model and prevent overfitting.	0 pts
To ensure that the residuals (the differences between predicted and true values) are normally distributed.	0 pts
To find the optimal values of the input variables (features) that minimize the output variable (responses).	0 pts

1 PT · MULTIPLE CHOICE

Which of the following statements about overfitting is **TRUE**?

Most of the time, we can deal with overfitting by removing outliers.	0 pts
Overfitting means fitting the test data extremely well.	0 pts
Overfitting usually happens when using a very complex model.	1 pt
We can combat overfitting by increasing the size of the test set.	0 pts

1 PT · MULTIPLE CHOICE

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:

The coefficient of feature X. 0 pts

Pearson correlation coefficient between the predictions Z and the feature X. 0 pts

Pearson correlation coefficient between the true response Y and the feature X. 0 pts

Pearson correlation coefficient between the true response Y and the predictions Z. 1 pt

1 PT · MULTIPLE CHOICE

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

R-squared increases as we add more features X. 1 pt

R-squared is a good evaluation metric for classification. 0 pts

A bad R-squared means that there is no pattern in the data. 0 pts

R-squared intuitively means the unexplained variation for the response variable Y. 0 pts

1 PT · MULTIPLE CHOICE

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

When using Random Forest, we can usually put raw data into the model without feature engineering. 0 pts

Classification is used to predict categorical labels, and regression is used to predict continuous values. 1 pt

We can just use the training data to evaluate if the model will work well. 0 pts

R-squared is a common evaluation metric for classification models. 0 pts

[Mock1Mid] Calculation Related Questions #32308839

2 pts · Last updated 22 Jan, 18:52 · Part of flow 6 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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 pts
0.33	0 pts
0	1 pt
0.5	0 pts

1 PT · MULTIPLE CHOICE

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?

0.4	0 pts
0.55	0 pts
0.3	0 pts
0.6	1 pt

[Mock1Mid] Coding Questions #32308842

3 pts · Last updated 22 Jan, 19:03 · Part of flow 7 · Saved in Data Science [5072DASC6Y]

1 PT · MULTIPLE CHOICE

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?

D.resample("60Min")	1 pt
D.resample("1Min").asfreq()	0 pts
D.resample("1Min", '1H')	0 pts
D.resample("60H")	0 pts

1 PT · MULTIPLE CHOICE

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

D.dropna().sum()["C2"]	1 pt
D.drop("C2", axis=1).sum()	0 pts
D["C1"].dropna().sum()	0 pts
D.groupby("C2").sum()	0 pts

1 PT · MULTIPLE CHOICE

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. The “Zipcode” column contains only strings (not integers). 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

D.groupby("Zipcode").mean()["Smell"].iloc("15213")	0 pts
D["Smell"].groupby("Zipcode").mean()["15213"]	0 pts
D.mean()["Smell"].groupby("Zipcode")["15213"]	0 pts
D[D["Zipcode"]=="15213"].mean()["Smell"]	1 pt