Computational Statistics

ⓘ This is a preview of the draft version of the quiz

Started: Oct 20 at 5:28pm

Quiz Instructions

Aids:

- Non-programmable calculator

Hint:

- Use period as decimal separator, not comma

Question 1

1 pts

You are given a two-dimensional Numpy array A:

```
array([[1, 2, 3],
       [4, 5, 6]])
```

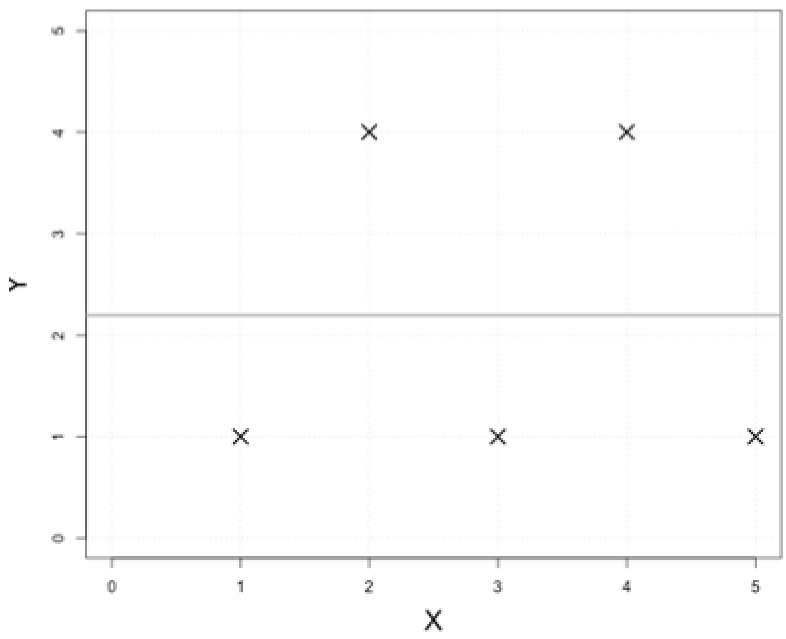
You want variable b to contain the sums of each column, i.e.:

```
array([5, 7, 9])
```

What needs to be put in for "[...]" in the following code to achieve that?

```
b = np.sum(A, axis=[...])
```

Data frame *df1* contains 5 points: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 4)$, $(x_3, y_3) = (3, 1)$, $(x_4, y_4) = (4, 4)$, and $(x_5, y_5) = (5, 1)$ as shown here:



Question 2

2 pts

What is the MSE (mean squared error) of the model shown as the gray line at $\hat{y} = 2.2$ for the entire data set? Round your answer to two decimal places.

Question 3

3 pts

Consider *df1* again. You randomly sampled points 1, 3, and 5 into the validation set. What is the validation MSE of the linear regression model predicting Y from X? Provide your calculations for partial credit.

Edit View Insert Format Tools Table

12pt Paragraph

p

0 words </>

You collect a set of data points (n = 500 observations) containing a single feature X and a quantitative target Y. You first split the data into 80% training data *train* and 20% validation data *validation*. Using the training data and Python, you then fit two regression models: a **quadratic regression model** using the single, non-transformed feature X as well as the additional, transformed feature X^2, and a **cubic regression model** using the original feature X as well as additional, transformed features X^2 and X^3.

Question 4

1.5 pts

Suppose that the **true relationship between X and Y is quadratic**, i.e., $Y = b_2 X^2 + b_1 X + b_0 + \epsilon$ for some b_0, b_1, b_2 and with some Gaussian noise ϵ with $\mathbf{E}[\epsilon] = \mathbf{0}$. Consider the **training MSE** for the quadratic regression and the training MSE for the cubic regression. Would we expect:

- There is not enough information to tell which regression model has lower error.
- The errors of both regression models to be about the same.
- The error of the cubic regression to be lower.
- The error of the quadratic regression to be lower.

Question 5

1.5 pts

Suppose that the true relationship between X and Y is still quadratic but consider now the **validation MSE** for the quadratic regression and the cubic regression. Would we expect:

- ◯ The error of the quadratic regression to be lower.
- ◯ The errors of both regression models to be about the same.
- ◯ There is not enough information to tell which regression model should have lower error.
- ◯ The error of the cubic regression to be lower.

Question 6

1.5 pts

Suppose now that the **true relationship between X and Y is not quadratic, but we don't know how far it is from quadratic**. Consider the **training MSE** for the quadratic regression and the cubic regression. Would we expect:

- ◯ There is not enough information to tell which regression model should have lower error.
- ◯ The error of the quadratic regression to be lower.
- ◯ The error of the cubic regression to be lower.
- ◯ The errors of both regression models should be about the same.

Question 7

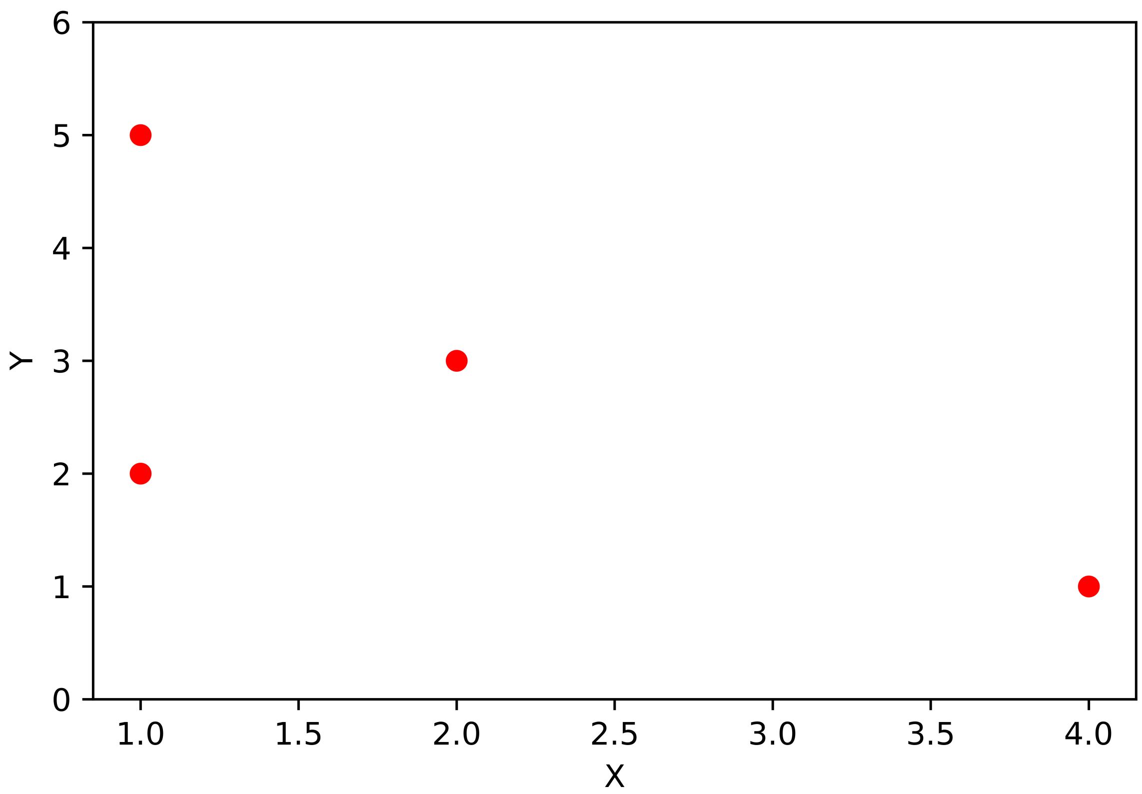
1.5 pts

Suppose the true relationship between X and Y is still not quadratic and we still don't know how far it is from quadratic. Consider the **validation MSE** for the quadratic regression and the cubic regression. Would we expect:

- ◯ The error of the cubic regression to be lower.

- ○ There is not enough information to tell which regression model should have lower error.
- ○ The error of the quadratic regression to be lower.
- ○ The errors of both regression models should be about the same.

When growing decision trees, the greedy recursive binary splitting algorithm needs to find the optimal split with respect to a given loss function. Consider the following example of a regression task with a single feature X and a target Y:



Question 8

3 pts

What is the SSE (sum of squared errors) on this training data for the split at X=1.5?

Question 9

3 pts

What is the SSE on this training data for the split at X=3.0?

Question 10

2 pts

What is the SSE when splitting at both X=1.5 and X=3.0 on this training data?

Multiple Choice Questions:

Question 11

1 pts

No other hyperplane has a lower training error than the maximal margin classifier when the training data is linearly separable.

- ◯ True
- ◯ False

Question 12

1 pts

If we do not limit the number of splits or leaf nodes, classification trees will eventually obtain a training loss of 0.

- ○ True
- ○ False

Question 13

1 pts

When growing a random forest, for each tree, one first samples a subset of features, and then restricts each split in that tree to only using one of those tree-specific subsampled features (with replacement).

- ○ True
- ○ False

Question 14

1 pts

If one multiplies a proper scoring rule with any real-valued number, the resulting scoring rule is still proper.

- ○ True
- ○ False

Question 15

1 pts

Random forests determine feature importance by looking at how often that feature appears at the root node.

- ○ True

◯ False

You have a dataset cancer_data.csv that contains data on patients who have gotten a sample of breast mass taken with each row representing one patient. There are 4 features and 1 target variable in the columns. The features describe characteristics of the cell nuclei present in the image of those samples and the target denotes whether a patient's sample is malignant (0) or benign (1). The objective is to train a random forest classifier using the features to predict the probability that the sample is benign. To evaluate your model, you are using the Brier (quadratic) score and the validation set method.

Large parts of the code are already given. Your task is to fill in the missing pieces. The brier function should return the average Brier score. The parts with missing pieces are marked in yellow (the length of the yellow parts are not necessarily the same as the length of the missing code):

```
1  import pandas as pd
2  import numpy as np
3  from sklearn.model_selection import train_test_split
4  from sklearn.ensemble import RandomForestClassifier
5
6  def brier(                ):
7      return 1-((y-model.           (X)[          ])**2)
8
9  df = pd.read_csv("cancer_data.csv", index_col=0)
10
11            = df.    [:,4]
12            = df.    [:,:4]
13
14 X_train, X_validate, y_train, y_validate = train_test_split(features, targets, test_size=0.1, random_state=3)
15
16
17
18
19 print("Average Brier score of Random Forest on validation set:", brier(rf,                     ))
```

Question 16

2 pts

Please enter (only!) the **missing code for line 6**:

Question 17

1 pts

Please enter (only!) the missing code for the **first yellow part of line 7**:

Question 18

1 pts

Please enter (only!) the missing code for the **second yellow part of line 7** (you can assume that the model was trained on a dataset containing both classes):

Question 19

1 pts

Please enter (only!) the missing code for the **third yellow part of line 7**:

Question 20

1 pts

Please enter (only!) the missing code for the **first yellow part of line 11**:

Question 21

0.5 pts

Please enter (only!) the missing code for the **second yellow part of line 11**:

Question 22

1 pts

Please enter (only!) the missing code for the **first yellow part of line 12**:

Question 23

0.5 pts

Please enter (only!) the missing code for the **second yellow part of line 12**:

Question 24

2 pts

Please enter the missing code for **line 16** (creating the random forest object; you can leave all parameters at the default values):

Question 25

2 pts

Please enter the missing code for **line 17** (training the model):

Question 26

1 pts

Please enter (only!) the missing code for **line 19**:

You are a probabilistic forecaster competing with two other forecasters for a single prize that will be given to the forecaster with highest quadratic scoring rule score. (Ties will be broken uniformly at random, so that all forecasters with highest quadratic score receive the prize with equal probability.)

The quadratic scoring rule is: $R_q\left(y,\, x\right) \;=\; 1 - \left(y - x\right)^2$

The forecasting competition has only a single question and you believe that the event will happen with probability p = 0.7. You know that the two other forecasters also believe that the probability of the event occurring is 0.7. Moreover, you know that they will report truthfully.

Question 27

1 pts

What is your subjective **probability of being selected** if you report 0.6?

- ○ 0
- ○ 0.7
- ○ 1/2
- ○ 1
- ○ None of these
- ○ 0.3

Question 28

2 pts

Which of these **forecasts maximizes your subjective probability of winning the prize?**

- ○ 1.0
- ○ 0.7
- ○ None of these
- ○ Any of these

Question 29

2 pts

Which of these forecasts maximizes your **subjective probability of winning the prize if the other two forecasters are not truthful but report 1.0 (i.e., 100%) instead?**

- ○ Any of these
- ○ 0.5
- ○ None of these
- ○ 0.0

Question 30

2 pts

What is the **highest expected probability of winning the prize that you can obtain assuming the other two forecasters report 1.0 (i.e., 100%)?**

Question 31

4 pts

In the lecture, we have seen that we can implement truthful forecasting competitions by giving the prize to forecaster i with probability

$$f_i = \frac{1}{n} + \frac{1}{n}\left(R_q(y_i, x) - \frac{1}{n-1}\sum_{j \neq i} R_q(y_j, x)\right).$$

For $x = 0$, compute the probability that forecaster 1 receives the prize when she reported $y_1 = 0.5$ and the other two forecasters reported $y_2 = y_3 = 0.0$.

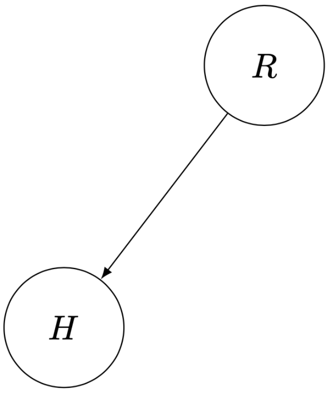
Edit View Insert Format Tools Table

12pt Paragraph

p

0 words

Consider the following Bayesian network representing the beliefs of a weather forecaster seeking to predict the probability of rain (R). The forecaster has been offered access to a (noisy) hygrometer (H) with the conditional probability table shown in the graphic.



$\Pr(R=1)=0.75$	$\Pr(R=0)=0.25$

R	$H=1$	$H=0$
1	0.9	0.1
0	0.6	0.4

Question 32

2 pts

What is the forecaster's belief that the hygrometer signal will be 1? That is, what is Pr(H=1)? (use **at least three decimals** in your calculations wherever possible)

Question 33

2 pts

What is the forecaster's belief of rain given that the hygrometer signal is 0? That is, what is Pr(R=1|H=0)? (use **at least three decimals** in your calculations wherever possible)

Question 34

4 pts

What is the expected quadratic score of the truthfully-reporting forecaster if she does not learn the outcome of H (i.e., if she does not have access to the hygrometer)? (The numerical result alone is sufficient but you can provide your calculation for partial credit; use **at least three decimals** in your calculations wherever possible.)

Edit View Insert Format Tools Table

12pt Paragraph

p

0 words

</>

Question 35

3 pts

The forecaster still seeks to predict the probability of R=1 and is still reporting truthfully. Her expected quadratic score following H=1 is 103/121 = 0.851, her expected quadratic score following H=0 is 37/49 = 0.755.

What is the forecaster's expected improvement in quadratic score from learning the outcome of H? (The numerical result alone is sufficient but you can provide your calculation for partial credit; use **at least three decimals** in your calculations wherever possible.)

Edit View Insert Format Tools Table

12pt Paragraph

p

0 words

Saving...

Submit Quiz