Tsa Tsa

f74372017@mailst.cjcu.edu.tw

BASIC INFO

Test: Data Science and ML test

Solved: 10/13

Similarity: low

Score: 1200/1500

Finished On: 10 Apr 2022 (CST)

Duration: 58m/1h

Label: None

Candidate's solution	Time Spent	Score	Similarity
customerStatistics	31min	300/300	low
mlConfusionMatrix	7min	0/100	-
mlClusteringCenters	1min	100/100	-
mlDevelopmentProcess	1min	100/100	-
mlSupervisedAndUnsupervised...	1min	100/100	-
mlROCCurve	1min	100/100	-
mlEnsembleAlgorithms	17sec	100/100	-
mlTreesInRF	1min	100/100	-
mlRFvsGB	1min	100/100	-
mlLinearRegressionOptimisation	28sec	100/100	-
mlLinearRegressionDegree	1min	0/100	-

mlLinearRegressionOffsets	28sec	100/100	-
mlLinearRegressionStatements	27sec	0/100	-

Task details: **customerStatistics**

Description:

You are given a `data.csv` file in the `/root/customers/` directory containing information about your customers.

It has the following columns:

```
ID,NAME,CITY,COUNTRY,CPERSON,EMPLCNT,CONTRCNT,CONTRCOST
```

where

```
ID: Unique id of the customer
NAME: Official customer company name
CITY: Location city name
COUNTRY: Location country name
CPERSON: Email of the customer company contact person
EMPLCNT: Customer company employees number
CONTRCNT: Number of contracts signed with the customer
CONTRCOST: Total amount of money paid by customer (float in format dollars.cents)
```

Read and analyze the `data.csv` file, and output the answers to these questions:

- How many total customers are in this data set?
- How many customers are in each city?
- How many customers are in each country?
- Which country has the largest number of customers' contracts signed in it? How many contracts does it have?
- How many unique cities have at least one customer in them?

The answers for second and third questions (the number of customers in each city and in each country) must be sorted by city and country name respectively, in ascending order. If there are several cities that are tied for having the most customers' contracts, print the lexicographically larger one. Please keep in mind that all string comparisons should be considered case-sensitive.

The answers should be formatted as:

```
Total customers:
<number>
Customers by city:
<CITY>: <number>
<CITY>: <number>
...
Customers by country:
<COUNTRY>: <number>
<COUNTRY>: <number>
...
```

```
Country with the largest number of customers' contracts:
<country> (<number> contracts)
Unique cities with at least one customer:
<number>
```

Example

For the following `data.csv`

```
ID,NAME,CITY,COUNTRY,CPERSON,EMPLCNT,CONTRCNT,CONTRCOST
00000001,Breadpot,Sydney,Australia,Sam.Keng@info.com,250,48,1024.00
00000002,Hoviz,Manchester,UK,harry.ham@hoviz.com,150,7,900.00
00000003,Hoviz,London,UK,hamlet.host@hoviz.com,1500,12800,10510.50
00000004,Grenns,London,UK,grenns@grenns.com,200,12800,128.30
00000005,Magnolia,Chicago,USA,man@info.com,1024,25600,512000.00
00000006,Dozen,San Francisco,USA,dozen@dozen.com,1000,5,1000.20
00000007,Sun,San Francisco,USA,sunny@sun.com,2000,2,10000.01
```

the output for this should be:

```
Total customers:
7
Customers by city:
Chicago: 1
London: 2
Manchester: 1
San Francisco: 2
Sydney: 1
Customers by country:
Australia: 1
UK: 3
USA: 3
Country with the largest number of customers' contracts:
USA (25607 contracts)
Unique cities with at least one customer:
5
```

Note that both *USA* and *UK* have the same number of contracts - `25607`, but *USA* is lexicographically larger, so it is the answer.

Solution (main.py3):

```
1  import pandas as pd
2  import csv
3
4  # Write your code here
5  #
6  df_train = pd.read_csv('/root/customers/data.csv')
7
8  print(f'Total customers:')
9  print(len(df_train))
10 print('Customers by city:')
11 # groupbycity = df_train.groupby(["CITY"]).size()
12 groupbyCity = df_train['CITY'].value_counts().sort_index()
13 for i in range(len(groupbyCity)):
```

```
14
print(f'{groupbyCity.index[i]}: {groupbyCity[i]}')
15  groupbyCountry = df_train['COUNTRY'].value_counts().sort_index()
16  print('Customers by country:')
17  for i in range(len(groupbyCountry)):
18      print(f'{groupbyCountry.index[i]}: {groupbyCountry[i]}')
19  print('Country with the largest number of customers\' contracts:')
20  new_df = df_train.groupby('COUNTRY').sum().sort_values(by=['CONTRCNT','CONTRCOST'],
    ascending = False)
21  print(f'{new_df.index.values[0]} ({int(new_df.iloc[0][2])} contracts)')
22  print('Unique cities with at least one customer:')
23  print(len(df_train['CITY'].value_counts()))
```

Task details: **mlConfusionMatrix**

Description:

You have a model for recognising images with faces. When presented with an image containing a face, it recognizes the face with `80%` accuracy. It also misidentifies some round objects as faces in `10%` of cases when presented with an image that doesn't include a face. What will the confusion matrix look like if you have a dataset with `60` images containing faces, and `40` images without faces?

		Image contains face	
		Positive	Negative
Predicted value	Positive	A	B
	Negative	C	D

- ○ A = 48, B = 12, C = 4, D = 36.

  (Incorrect)

- ◉ A = 54, B = 6, C = 8, D = 32.

  (Incorrect)

- ○ A = 48, B = 4, C = 12, D = 36.

  (Correct)

- ○ A = 54, B = 8, C = 6, D = 32.

  (Incorrect)

- ○ A = 80, B = 20, C = 10, D = 90.

(Incorrect)

Task details: **mlClusteringCenters**

Description:

You are working on a model for production.
You need to run a clustering algorithm on `n` objects, which minimizes the sum of the distances from each object to the center of each cluster.

How do you get the right number of clusters?

- ○ `n`, all centers of clusters are at the same points as objects, and the sum of the distances is `0` (minimal possible).

  (Incorrect)

- ○ `1`, all the objects should belong to one cluster.

  (Incorrect)

- ○ Depends on the sum of all distances between all possible pairs of objects.

  (Incorrect)

- ○ Depends on the data; for normalized values you must take a smaller number of clusters.

  (Incorrect)

- ◉ Depends on the task, the number which gives the best result on the validation test.

(Correct)

Task details: **mlDevelopmentProcess**

Description:

Put the following steps in the correct order:

1. Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.
2. Validate the model.
3. Explore the data and become familiar with it.
4. Understand the business problem.
5. Start running the model, analyze the result and tweak the approach.

- [x] 4, 3, 1, 5, 2. *(Correct)*
- [ ] 5, 4, 3, 1, 2. *(Incorrect)*
- [ ] 1, 3, 4, 2, 5. *(Incorrect)*
- [ ] 4, 3, 2, 5, 1. *(Incorrect)*

Task details: **mlSupervisedAndUnsupervisedLearning**

Description:

Categorize the following statements as either 1) Supervised learning, or 2) Unsupervised learning:

A) Linear Regression is an example of this group of algorithms.
B) Support vector machines is an example of this group of algorithms
C) Clustering is an example of this group of algorithms.
D) All input data is labeled.
E) All input data is unlabeled.
F) There is no feedback based on the prediction results.

- ○ 1: A, D. 2: B, C, E, F. *(Incorrect)*
- ○ 1: A, B, C, D. 2: E, F. *(Incorrect)*
- ○ 1: C, E, F. 2: A, B, D. *(Incorrect)*
- ◉ 1: A, B, D. 2: C, E, F. *(Correct)*
- ○ 1: A, C, D. 2: B, E, F. *(Incorrect)*

Task details: **mlROCCurve**

Description:

The area under the ROC-curve could vary from `0` to `1`. What conclusions can be drawn if the area under the ROC-curve is equal to `0.5`?

- [x] The algorithm gives random answers.

(Correct)

- [ ] The algorithm gives all zeros to all the objects.

(Incorrect)

- [ ] The algorithm gives all ones to all the objects.

(Incorrect)

- [ ] The algorithm gives the possible minimum for loss function.

(Incorrect)

- [ ] The algorithm confuses the class for prediction and predicts zeros for classes with answer one, and vice versa.

(Incorrect)

Task details: **mlEnsembleAlgorithms**

Description:

Which of the following algorithms is **not** an example of an ensemble learning algorithm?

- ○ Random Forest. *(Incorrect)*
- ○ AdaBoost. *(Incorrect)*
- ○ Extra Trees. *(Incorrect)*
- ○ Gradient Boosting. *(Incorrect)*
- ✓ Decision Trees. *(Correct)*

Task details: **mlTreesInRF**

Description:

You can generate hundreds of individual trees in a Random Forest. Choose the correct statements about an individual tree in Random Forest:

A. An individual tree is built on a subset of the features.
B. An individual tree is built on all the features.
C. An individual tree is built on a subset of observations.
D. An individual tree is built on the full set of observations.

○ A.	*(Incorrect)*
○ B.	*(Incorrect)*
○ C.	*(Incorrect)*
○ D.	*(Incorrect)*
◉ A, C	*(Correct)*
○ B, D.	*(Incorrect)*
○ A, D.	*(Incorrect)*
○ B, C.	*(Incorrect)*

Task details: **mlRFvsGB**

Description:

Choose the correct statement(s) about Random Forest and Gradient Boosting ensemble methods:

A. Both methods can be used for classification task.
B. Both methods can be used for regression task.
C. Random Forest is used for classification whereas Gradient Boosting is used for regression task.
D. Random Forest is used for regression whereas Gradient Boosting is used for Classification task.

○ A.	*(Incorrect)*
○ B.	*(Incorrect)*
○ C.	*(Incorrect)*
○ D.	*(Incorrect)*
◉ A, B.	*(Correct)*
○ A, C.	*(Incorrect)*
○ A, D.	*(Incorrect)*
○ B, C.	*(Incorrect)*

○ B, D. *(Incorrect)*

Task details: **mlLinearRegressionOptimisation**

Description:

Which of the following methods do we use to find the line of best fit for data in Linear Regression?

A) Least Square Error
B) Maximum Likelihood
C) Logarithmic Loss

- [x] A. *(Correct)*
- [ ] B. *(Incorrect)*
- [ ] C. *(Incorrect)*
- [ ] A, B. *(Incorrect)*
- [ ] A, C. *(Incorrect)*
- [ ] A, B, C. *(Incorrect)*

Task details: **mlLinearRegressionDegree**

Description:

Suppose that you have a dataset and you are designing a degree `3` polynomial regression model. You have found that a degree `3` polynomial perfectly fits the data; in other words, the training and testing error is `0`. Choose correct statements:

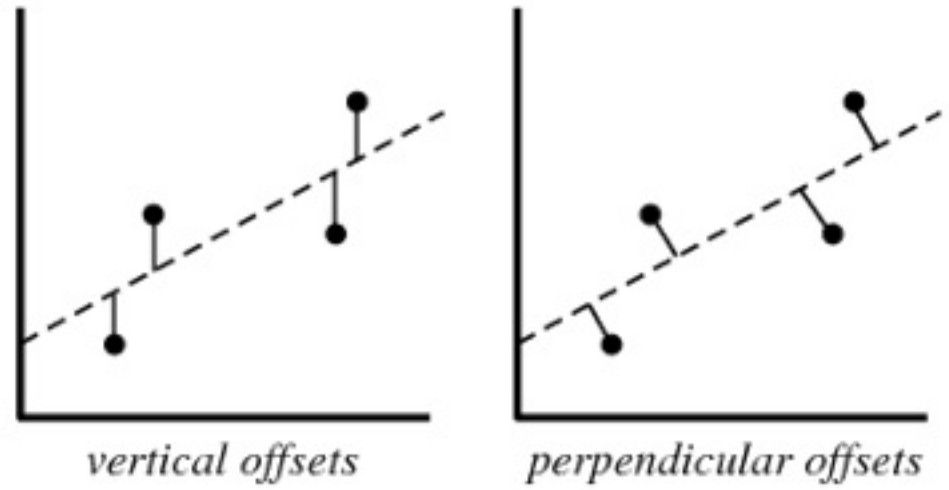
A) There are high chances that a degree 4 polynomial will overfit the data.
B) There are high chances that a degree 4 polynomial will underfit the data.
C) There are high chances that a degree 2 polynomial will overfit the data.
D) There are high chances that a degree 2 polynomial will underfit the data.
E) Can’t say anything about degree 4.
F) Can’t say anything about degree 2.

- ○ B, C. *(Incorrect)*
- ○ D, E. *(Incorrect)*
- ○ A, D. *(Correct)*
- ○ A, C. *(Incorrect)*
- ○ A, F. *(Incorrect)*
- ✓ E, F. *(Incorrect)*

Task details: **mlLinearRegressionOffsets**

Description:

Which of the following offsets do we use in linear regression's least square line fit? Suppose the horizontal axis is the independent variable and the vertical axis is the dependent variable.



- ☑ Vertical offset. *(Correct)*
- ○ Perpendicular offset. *(Incorrect)*
- ○ Both, depending on the situation. *(Incorrect)*
- ○ None of above. *(Incorrect)*

Task details: **mlLinearRegressionStatements**

Description:

Choose the correct statements about Linear Regression:

A) Lasso Regularization can be used for variable selection in Linear Regression.
B) It is possible to design a Linear Regression algorithm using a neural network.
C) Linear Regression is a supervised machine learning algorithm.
D) Overfitting is more likely when you have huge amount of data to train.
E) Linear Regression is sensitive to outliers.

○ A, B, C, E.	*(Correct)*
○ B, C, E.	*(Incorrect)*
✓ A, C, E.	*(Incorrect)*
○ B, C, D, E.	*(Incorrect)*
○ A, B, C, D, E.	*(Incorrect)*