0/21 Questions Answered

Quiz 1

Student Name

Search students by name or email... ▾

Q1 SQL

18 Points

Consider the following questions about a database recording basketball players, teams, and games.

The tables in this database are:

```
Teams (t_id, t_name)
Players (p_id, p_name, p_t_id)  % p_t_id references t_id
Games (g_id, t1_id, t2_id, date, t1_score, t2_score) %t1_id references t_id
                                                     %t2_id references t_id
```

Q1.1

4 Points

Which of the following statements are true about this schema? (Choose all that apply.)

☐ This schema can represent a player who plays for multiple teams

☐ The only way to represent a game without a score is to use a NULL value

☐ This schema cannot properly represent the teams that played in each game

☐ Representing the number (e.g., for the jersey) of each player on each team would required adding a new table.

Save Answer

Q1.2

4 Points

Suppose you wanted to track the players who played in each game. How would you represent this? What changes would you make to the database?

Save Answer

Q1.3
10 Points

Suppose you want to find all teams with a player named 'Mary Wu', without a player named 'Marny Vu'. We've provided a partial query below.

```
with mv_players as
(
    SELECT t_id FROM teams JOIN players
    ON p_t_id = t_id
    WHERE p_name = 'Marny Vu'
),
mw_players as (
    SELECT t_id FROM teams JOIN players
    ON p_t_id = t_id
    WHERE p_name = 'Mary Wu'
)
SELECT _______
```

What SQL should go into the select clause to complete the query?

Save Answer

Q2 Pandas
12 Points

Consider the following pandas code:

```
import pandas as pd
df = pd.read_csv("test.csv")

print(df.columns)
print(df.shape)
print(df.a > df.b)
print(df.b > df.c)
print(df.max())
print(df.min())
df.iloc[0] = 10
print(df.min())
print(df.iloc[1].mean())
```

When you run this code, it prints:

```
Index(['a', 'b', 'c'], dtype='object')

(3, 3)

0    False
1    False
2    False
dtype: bool

0    False
1    False
2    False
dtype: bool

a    7
b    8
c    9
dtype: int64

a    1
b    2
c    3
dtype: int64

a    4
b    5
c    6
dtype: int64

5.0
```

Based on the output above, what is one possible value for the contents of "test.csv"?
Your answer should take the form:
a b c
N N N
N N N
N N N

Where you replace the N's with the numbers in test.csv

Save Answer

Q3 Regular Expressions

10 Points

Given the following Python function f(), write a regular expression that matches the same inputs for which f() returns true (you can assume the use of extended regex without the need for escapes; we will be flexible about grading and escapes):

```
def f(line):
    alpha = 'abcdefghijklmnopqrstuvwxyz'
    num = '0123456789'
    els = line.split(',')
    els = line.split(',')
    if len(els) != 3:
        return False
    c1 = [True if c in num else False for c in els[0]]
    c2 = [True if c in alpha else False for c in els[1]]
    c3 = els[2] == 'red' or els[2] == 'blue'

    #python "all" returns true if all elements in an array are true
    return (len(c1) == 2 and all(c1)) and (len(c2) == 3 and all(c2)) and c3
```

Save Answer

Q4 Cosine Similarity

6 Points

Suppose you want to use cosine similarity to compute the similarity of documents with repeated words, weighting words that occur more frequently higher. Your labmate, C. Quill, suggests that you represent each document as a length N vector, where N is the total number of distinct words in the document corpus, and where the ith element of the jth document vector represents the number of times the ith word occurs in document j. Using her idea, compute the term-frequency-weighted cosine similarity between the following sentences:

```
the slug eats the bug in the rug
the bug runs on the rug
```

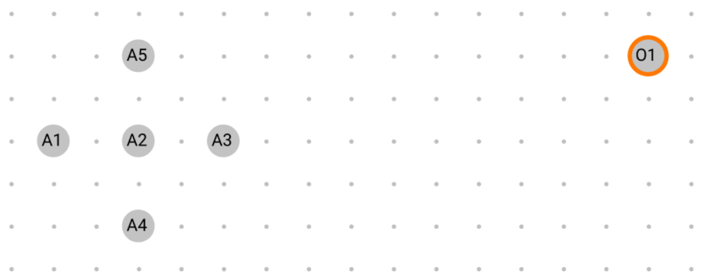
Give your answer and justify your response:

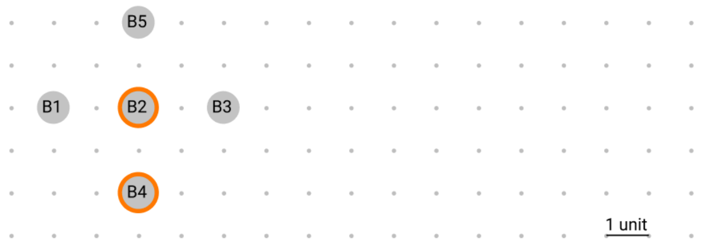
Save Answer

Q5 Clustering

19 Points

You have just received a dataset with 2 features, which appears to come from two distributions.





Q5.1 K-Means Clustering 1-Step

5 Points

Consider the data points above and a Euclidean distance function. Using k-means clustering with three centroids that start at points B2, B4, and O1 (marked in orange), after **one iteration** of the algorithm what are the resulting clusters?

Give your answer in the following format:
C={Point1, Point2....},
C={Point1, Point2,...}
...

Example answer:
C=[A1, A2, A3, A4, B5]
C=[B1, B2, B3, B4]
C={A5, O1}

Note: If you can answer the question intuitively, you do not have to formally calculate the answer. We do NOT require the coordinate of the new centroids.

Save Answer

Q5.2 K-Means Clustering Final Clustering

2 Points

What will be the final clustering (use the same answer format as above).
Note: Again an intuitive approach to derive at the solution is sufficient.

Save Answer

Q5.3 K-Means 2 Centroids

Q5.3 K-Means 2 Centroids

2 Points

There exists two clusters in the dataset above (A1-A5 and B1-B5) and one outlier (O1). Can K-Means clustering with 2 centroids return the right clusters (A1-A5 and B1-B6)? Is it always guaranteed that it groups A1-A5 and B1-B5 will be together? What happens with the outlier data point O1?

Save Answer

Q5.4 DB-Scan

5 Points

Instead of k-means, consider instead using DBScan. How many clusters would DBScan return (with MintPtr=3 and eps=2 units). Use again the same format as for 5.1.

Save Answer

Q5.5 Other clustering algorithms

5 Points

What other algorithm from class would you suggest to cluster A1-A5 into one group and B1-B5 into another?

- □ Agglomerative clustering with average linkage
- □ Agglomerative clustering with single linkage
- □ Gaussian mixture clustering using EM
- □ None of the above

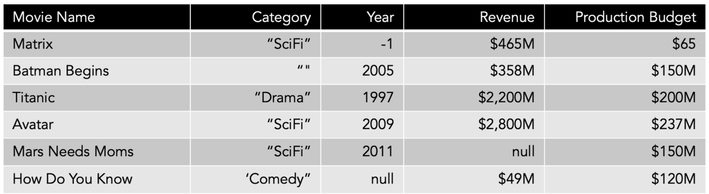
Briefly Explain your answer

Save Answer

Q6 Data Cleani

Q6 Data Cleaning

6 Points



Movie Name	Category	Year	Revenue	Production Budget
Matrix	“SciFi”	-1	$465M	$65
Batman Begins	“ "	2005	$358M	$150M
Titanic	“Drama”	1997	$2,200M	$200M
Avatar	“SciFi”	2009	$2,800M	$237M
Mars Needs Moms	“SciFi”	2011	null	$150M
How Do You Know	‘Comedy”	null	$49M	$120M

Q6.1 Detecting errors

3 Points

Suppose you have a large dataset with thousands of records regarding movies like the one above. What techniques and in what order would you use them to detect potential data problems? For each method describe their pros and cons (bullet points are enough).

Save Answer

Q6.2

3 Points

Considering the dataset above, which contains 4 missing values, what data cleaning technique would most likely lead to the **highest average revenue** (e.g., *SELECT SUM(Revenue)/Count(*) FROM movies*)

Listwise deletion (drop rows with missing values / errors)

Mean substitution

Simple stochastic imputation using random sample

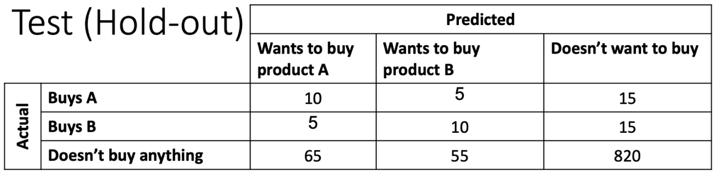
Explain your answer

Save Answer

Q7 Evaluation

21 Points

A startup has built a classifier to predict if one of their marketing contacts will buy product A or B or nothing in the next 6 months. Suppose that the confusion matrix below is the result of testing the classifier on 1000 marketing contacts.



Test (Hold-out)

		Predicted		
		Wants to buy product A	Wants to buy product B	Doesn't want to buy
Actual	Buys A	10	5	15
	Buys B	5	10	15
	Doesn't buy anything	65	55	820

Q7.1

6 Points

What is the accuracy of the classifier based on the confusion matrix above? Would you consider this classifier good? Explain your answer.

Save Answer

Q7.2

8 Points

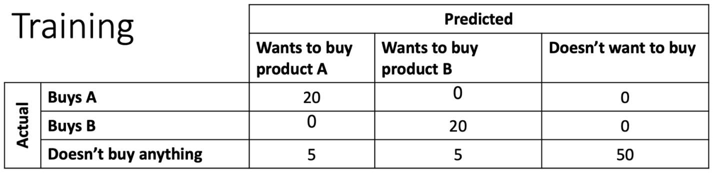
What is the F1 macro score of the classifier above? Is F1 macro score a better metric in this case? Why or why not? (You may use a Calculator app, Python, Excel or Google Spreadsheets to answer the question)

Save Answer

Q7.3

4 Points

Assuming this was the confusion metric on the training dataset.



Training

		Predicted		
		Wants to buy product A	Wants to buy product B	Doesn't want to buy
Actual	Buys A	20	0	0
	Buys B	0	20	0
	Doesn't buy anything	5	5	50

Considering this training error and the test error given above. Does the classifier suffer from high bias or high variance? Explain your answer.

Save Answer

Q7.4

3 Points

Given the above test and training confusion matrices, which of the following techniques would likely improve the F1-score of the classifier on the test data? (Choose all that apply and assume that they are used in isolation and not together)

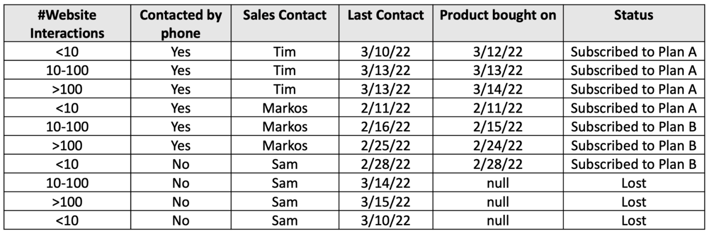
- □ Add more training data
- □ Remove data
- □ Increase regularization
- □ Decrease regularization
- □ Add more features
- □ Remove features

Save Answer

Q8 Decision Trees

8 Points

You work at a new startup that sells two types of subscriptions: Plan A and Plan B. One of your first tasks is to build a classifier to predict if a potential lead (i.e., a potential customer) will eventually convert and buy a subscription based on the data below (the *Status* column is the label):



#Website Interactions	Contacted by phone	Sales Contact	Last Contact	Product bought on	Status
<10	Yes	Tim	3/10/22	3/12/22	Subscribed to Plan A
10-100	Yes	Tim	3/13/22	3/13/22	Subscribed to Plan A
>100	Yes	Tim	3/13/22	3/14/22	Subscribed to Plan A
<10	Yes	Markos	2/11/22	2/11/22	Subscribed to Plan A
10-100	Yes	Markos	2/16/22	2/15/22	Subscribed to Plan B
>100	Yes	Markos	2/25/22	2/24/22	Subscribed to Plan B
<10	No	Sam	2/28/22	2/28/22	Subscribed to Plan B
10-100	No	Sam	3/14/22	null	Lost
>100	No	Sam	3/15/22	null	Lost
<10	No	Sam	3/10/22	null	Lost

Q8.1

2 Points

Looking at the data, would you transform and/or exclude any columns. Explain why?

Save Answer

Q8.2

6 Points

Only **using the first three columns** (*#Website interactions*", *Contacted by phone*, and *Sales Contact*) and the Status column as a label, use ID3 the compute the first split.

What feature will ID3 split on?

Note: You do not have to do the calculation for the splits if you can intuitively tell which split will be picked.

Website Interactions

Contacted by phone

Sales Contact

Save Answer

Q9 Assumptions & Explanations

0 Points

You may use this field to enter any additional explanations or assumptions for multiple choice / short answer questions above.

Save Answer

Q10 Future Topics

0 Points

Which of the following topics would you like to see more of in the 2nd half of the course?

□ Hands-on / practical guide to machine learning

□ Overview of Deep Learning

□ Scaling up Data Science (Parallelism, Multiple Machines, etc)

□ Case Studies - Data Science in Practice

□ Industrial Speakers

□ Flexible Class Time to Discuss Projects with Staff

□ In-Class Coding Exercises / Reversed Lectures

Any other comments / suggestions?

Save Answer

Save All Answers

Submit & View Submission ›