6.S079 Quiz 1 Sample Questions
Topics

● SQL
  ○ All concepts on PS 1, e.g., joins, aggregates, window functions, recursion
● Database Tuning
  ○ When to create & use indexes
● Pandas
  ○ Understanding python code & semantics of operations
● Regular Expressions
  ○ Sed, awk, grep
● Text processing
  ○ Stemming
  ○ Similarity metrics, e.g., cosine and jaccard similarity
● Missing Value Substitution
● Entity Resolution
● Outliers / Box Plots
● Feature engineering
● Clustering Algorithms
● Classification Algorithms
  ○ Decision Tree (ID3)
  ○ Differences of common algorithms
  ○ Decision boundaries
● EM algorithm
● Evaluation
  ○ Accuracy / Precision / Recall / F1
  ○ Cross-validation
  ○ Training vs Test vs Validation error
  ○ Bias vs Variance
SQL 1

Consider the following English language and corresponding SQL queries over a database recording books, authors, bookstores, stores, and sales

The tables in this database are:

Authors (a_id, name)
Books (b_id, title, b_a_id) % b_a_id references a_id
Stores (st_id, name, address)
BookStores (bs_st_id, bs_b_id) % bs_st_id references st_id,
  % bs_b_id references b_id
Sales(s_b_id, s_st_id, date, price) %s_b_id references b_id
  %s_st_id references st_id
Complete the query

1. Find the stores that sell no books by the author 'John Smith'

```sql
WITH JSBookStores as ( 
    SELECT bs_st_id 
    FROM BookStores JOIN Books ON bs_b_id = b_id 
    JOIN Authors on b_a_id = a_id 
    WHERE name = 'John Smith' 
)

SELECT bs_st_id FROM BookStores WHERE _______________

SELECT bs_s_id FROM BookStores WHERE bs_s_id NOT IN (SELECT bs_s_id FROM JSBookStores)
```
Find the number of books written by each author

```sql
SELECT name, count(*) FROM
Authors
GROUP BY name
```

```sql
SELECT name, count(*) FROM
Authors LEFT JOIN Books ON a_id = b_a_id
GROUP BY name
```
Regular Expressions

Suppose you need to extract the month from a sequence of ID codes in a file 'input.csv', where each line is of the form:

N…CCC MMM DD

Where N… is one or more numerical digits, CCC is 3 characters, MMM is a 3 character month abbreviation, and DD is a 2 digit date.

Write a sed expression to output the MMM from a number of rows of this form, using the sed command:

s/pattern/replacement/g

Where pattern is a regular expression and replacement is the output of the pattern.

You can assume that you are using "extended sed" such that you don't need to escape the characters '{', '}', '+', '*', '(', and ')' when using them for counting and grouping in the regular expression.
Regular Expressions Solution

s/^[0-9]+[a-zA-Z]{3}([a-zA-z]{3})[0-9]{2}/\1/g

Explanation:

^ - newline
[0-9]+ - leading digits
[a-zA-Z]{3} - 3 characters
( … ) - the backreferences we want to capture
[a-zA-Z]{3} - 3 character month
[0-9]{2} - trailing two digits

\1 is the value of captured backreference (3 character month), which we want to output
Text Similarity

Given the following two passages, compute their Jaccard and Cosine similarity

The quick brown fox runs
The beige slow fox runs hard

<table>
<thead>
<tr>
<th></th>
<th>The</th>
<th>Quick</th>
<th>Brown</th>
<th>Fox</th>
<th>Runs</th>
<th>Beige</th>
<th>Slow</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Quick</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Brown</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Fox</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Runs</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Beige</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Slow</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hard</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Cos(\(\Theta\)) = \(\mathbf{V1} \cdot \mathbf{V2} / \|\mathbf{V1}\| \times \|\mathbf{V2}\|\)

\[
3 / (\sqrt{5} \times \sqrt{6}) = .55
\]

\[
sim(s1, s2) = \frac{s1 \cap s2}{s2 \cup s2}
\]

\[
3 / 8 = .375
\]
We have six data points in the euclidean plane, and we want to use k-means algorithm to cluster them.

Suppose we set $K=2$. If the initial position of the centroids are at (-4, -4) and (-3, -3), after running the algorithm for one iteration, what is the position of each centroid? Which centroid will each point be assigned to?
An online retailer Valmart wants to know if any of the product listings will run out of stock on Black Friday. As a machine learning engineer at the firm, your job is to build a classifier to predict whether a given product listing will go **Sold Out**. **Show the first split** for an ID3 decision tree on the following examples. (10 points)

<table>
<thead>
<tr>
<th>Category</th>
<th>On Sale</th>
<th>Price Range</th>
<th>Sold Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home &amp; Elec</td>
<td>Yes</td>
<td>$100 ~ $200</td>
<td>Yes</td>
</tr>
<tr>
<td>Home &amp; Elec</td>
<td>No</td>
<td>$100 ~ $200</td>
<td>Yes</td>
</tr>
<tr>
<td>Books &amp; Music</td>
<td>Yes</td>
<td>$5 ~ $25</td>
<td>Yes</td>
</tr>
<tr>
<td>Books &amp; Music</td>
<td>No</td>
<td>$5 ~ $25</td>
<td>No</td>
</tr>
<tr>
<td>Books &amp; Music</td>
<td>Yes</td>
<td>$5 ~ $25</td>
<td>No</td>
</tr>
<tr>
<td>Books &amp; Music</td>
<td>No</td>
<td>$30 ~ $90</td>
<td>Yes</td>
</tr>
<tr>
<td>Clothing &amp; Shoes</td>
<td>Yes</td>
<td>$30 ~ $90</td>
<td>No</td>
</tr>
<tr>
<td>Clothing &amp; Shoes</td>
<td>No</td>
<td>$30 ~ $90</td>
<td>No</td>
</tr>
</tbody>
</table>
Evaluation

Suppose you have two classifiers (C1 and C2). For each classifier, we provide you with two confusion matrices: one for the training set, the other one from the test set (you can assume it is based on the average of 10-fold cross-validation).

**For Classifier 1:** Is the classification accuracy good? Does the classifier suffer from high bias or high variance? Explain your answers. It is close to random guessing. It suffers from high bias. Error is high on training and test data.

**For Classifier 2:** Is the classification accuracy good? Does the classifier suffer from high bias or high variance? Explain your answers. High Variance