CSE 158, Fall 2018: Midterm

Name: ___________________________ Student ID: ___________________________

Instructions
The test will start at 5:10pm. Hand in your solution at or before 6:10pm. Answers should be written directly in the spaces provided.

Do not open or start the test before instructed to do so.

Note that the final page contains some algorithms and definitions. Total marks = 26
Section 1: Regression and Ranking (6 marks)

Suppose you wanted to predict Air Quality measurements (generally measured using an index of particulate matter called ‘PM2.5’) for a large city. Suppose you have a dataset containing thousands of hourly measurements to do so. Examples of previous measurements look like:

<table>
<thead>
<tr>
<th>Observation</th>
<th>Date</th>
<th>Time</th>
<th>PM2.5</th>
<th>Temp (°C)</th>
<th>Wind sp. (km/h)</th>
<th>Wind direction</th>
<th>Humidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10/03/2004</td>
<td>18.00.00</td>
<td>150</td>
<td>24.4</td>
<td>4.1</td>
<td>NNW</td>
<td>76</td>
</tr>
<tr>
<td>2</td>
<td>10/03/2004</td>
<td>19.00.00</td>
<td>112</td>
<td>22.8</td>
<td>5.8</td>
<td>NW</td>
<td>84</td>
</tr>
<tr>
<td>3</td>
<td>10/03/2004</td>
<td>20.00.00</td>
<td>88</td>
<td>20.7</td>
<td>2.2</td>
<td>NW</td>
<td>87</td>
</tr>
<tr>
<td>4</td>
<td>10/03/2004</td>
<td>21.00.00</td>
<td>80</td>
<td>16.5</td>
<td>4.4</td>
<td>NNE</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>10/03/2004</td>
<td>22.00.00</td>
<td>51</td>
<td>15.5</td>
<td>2.1</td>
<td>W</td>
<td>90</td>
</tr>
<tr>
<td>6</td>
<td>10/03/2004</td>
<td>23.00.00</td>
<td>38</td>
<td>12.8</td>
<td>0.4</td>
<td>W</td>
<td>83</td>
</tr>
<tr>
<td>7</td>
<td>11/03/2004</td>
<td>00.00.00</td>
<td>31</td>
<td>11.8</td>
<td>0.6</td>
<td>SW</td>
<td>78</td>
</tr>
<tr>
<td>8</td>
<td>11/03/2004</td>
<td>01.00.00</td>
<td>31</td>
<td>10.9</td>
<td>1.3</td>
<td>S</td>
<td>69</td>
</tr>
</tbody>
</table>

1. Both the time and the date could be useful for this type of prediction. Suggest a scheme for representing the date and time, and write down the resulting features for the first two observations (2 marks).

A:  

1: 10/03/04 18:00  

2: 10/03/04 19:00  

2. Similarly, suppose you wanted to incorporate the wind direction and wind speed into your predictor. Describe your encoding and write down the features of the first two datapoints (2 marks).

A:  

1: 0...0 0...0 10 (°C) 2 15 18  

2: 0...0 10...0 0...0 10...0  

3. When predicting a future PM2.5 value, a useful predictor might be one (or several) previous PM2.5 values (i.e., the labels of previous observations become features for the current observation, so you might predict the 8th observation using features derived from the previous 7 observations, etc.). This procedure is known as autoregression. Describe which previous observations you might use, or what other features you might extract from past observations, in order to make a system that was effective at forecasting future observations (2 marks).

A:  

\[ y = \rho_{2.5} x \]  

concatenate \( x \) with features from previous Q's

\[^1\text{NNW = North-North-West, etc.}\]
Section 2: Classification and Diagnostics (9 marks)

Suppose you wish to build a classifier to detect malicious e-mails (e.g., spam, phishing, etc.). You collect 10,000 e-mails, and obtain ground-truth labels indicating which e-mails are malicious (i.e., malicious e-mails are labeled True). You then train three classifiers, whose performance is as follows:

<table>
<thead>
<tr>
<th>Classifier 1</th>
<th>Classifier 2</th>
<th>Classifier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Positives 150</td>
<td>False Positives 3828</td>
<td>False Positives 843</td>
</tr>
<tr>
<td>False Negatives 21</td>
<td>False Negatives 6</td>
<td>False Negatives 40</td>
</tr>
<tr>
<td>True Positives 35</td>
<td>True Positives 50</td>
<td>True Positives 16</td>
</tr>
<tr>
<td>True Negatives 9794</td>
<td>True Negatives 6135</td>
<td>True Negatives 9101</td>
</tr>
</tbody>
</table>

4. How many of the 10,000 instances have a positive label (i.e., \( y_i = \text{True} \)) (1 mark)?

A: \( TP + FN = 21 + 25 = 56 \)

5. How many of the 10,000 instances have a positive prediction for Classifier 1 (i.e., \( f(X) = \text{True} \)) (1 mark)?

A: \( TP + FP = 75 + 150 \)

6. Compute the following statistics for Classifier 1. You can leave your results as unsimplified expressions (4 marks):

   - Accuracy: A: \( \frac{TP + TN}{10000} \)
   - BER: A: \( \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \)
   - Precision: A: \( \frac{TP}{TP + FP} \)
   - Recall: A: \( \frac{TP}{TP + FN} \)

Which of the three classifiers would you select if your goal is to optimize the measures below? Assume that content where the prediction is positive is filtered/blocked (e.g., moved to a spam folder). **Briefly state your reasoning for each answer.** (1 mark each).

7. The classifier with the highest accuracy:

A: \( TP + TN \)

8. The classifier that lets the **fewest malicious e-mails** through the filter:

A: \( FN \)

9. The classifier that filters the **fewest non-malicious e-mails**:

A: \( FP \)
Section 3: Clustering / Communities (5 marks)

Suppose you collect a dataset of taxi rides in New York, containing pickup and dropoff locations, among other features. After generating a scatterplot of the data you obtain the following result:

![Scatterplot](image)

**Fig. 1: Mapping of pick-up and drop-off locations**

Suppose your goal is to predict the total tip that a given fare will receive.

You consider three alternative techniques to incorporate the geographical location into your model:

- **Grid:** Split the data into a grid (over latitude and longitude), and include a feature indicating which grid position each datapoint belongs to.
- **Nearest Neighbor:** For each new trip, identify the ‘most similar’ trip in the training data in terms of the distance between start and end locations. Predict the tip for the new trip to be the same as the tip for this previous trip (this is known as ‘nearest neighbor’ classification).
- **Clustering:** Run a clustering algorithm (e.g. k-means or hierarchical clustering) to obtain feature representations of each point.

10. Suggest one reason why clustering the data might be preferable to each of the ‘grid’ or ‘nearest neighbor’ models (2 marks):

<table>
<thead>
<tr>
<th>Versus Grid:</th>
<th>Might be too granular in dense regions, whereas N.Y.C. would have many clusters near parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Versus Nearest Neighbor:</td>
<td>Absolute position might matter: 2017 trip in downtown ≠ 2017 tip in suburbs</td>
</tr>
</tbody>
</table>

11. (Design thinking) In addition to geographical features, suggest (at least three) additional features that may be useful in predicting tip amounts (3 marks):

- total pre-tip cost
- speed
- time of day/day of week

---

2Scatterplot taken from a previous CSE258 assignment on taxi tip prediction.
Suppose you collect the following ratings of teen romance novels from Goodreads:

<table>
<thead>
<tr>
<th>Item ID</th>
<th>Book</th>
<th>Read?</th>
<th>Rated?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Nathan Thomas Dhruv Kevin Prateek</td>
<td>Nathan Thomas Dhruv Kevin Prateek</td>
</tr>
<tr>
<td>1</td>
<td>To All the Boys I’ve Loved Before</td>
<td>1 1 0 1 1</td>
<td>5 3 ? 1 4</td>
</tr>
<tr>
<td>2</td>
<td>P.S. I Still Love You</td>
<td>1 0 0 0 1</td>
<td>5 ? ? ? 4</td>
</tr>
<tr>
<td>3</td>
<td>Always and Forever, Lara Jean</td>
<td>1 0 0 0 0</td>
<td>4 ? ? ? ?</td>
</tr>
<tr>
<td>4</td>
<td>It All Started with an Apple</td>
<td>0 1 0 0 0</td>
<td>? 2 ? ? ?</td>
</tr>
<tr>
<td>5</td>
<td>The Kissing Booth</td>
<td>1 0 1 1 1</td>
<td>1 ? 1 2 4</td>
</tr>
</tbody>
</table>

You want to make a simple recommender that identifies the ‘all time best’ books, using a model of the form

\[
\text{rating}(i) = \alpha + \beta_i.
\]

Here \( \alpha \) is a global term, and \( \beta_i \) is an item bias. You fit your model by setting \( \alpha \) to the global mean of all ratings, and \( \beta_i \) to be the remainder. Finally, you make recommendations simply by identifying those items with the highest bias terms, i.e.,

\[
\text{argmax}_i \beta_i.
\]

12. Noting that the average rating is 3.0, what is the bias term \( \beta_i \) for Item #1 (1 mark)

A: \( \frac{13}{4} \)

13. What item would receive the highest ranking according to this global recommender (1 mark)?

A: Item 2 (highest average)

Items 1, 2, and 3 are consecutive books from the same series. You notice that users only read each sequel if they liked the previous book, which biases ratings of sequels (items 2 and 3) to be particularly high. You propose modifying your model to take the form

\[
\text{rating}(i) = \text{rating(previous book in sequence)} + \beta_i.
\]

Non-sequels are still assigned ratings according to \( \text{rating}(i) = \alpha + \beta_i \).

14. After fitting your model following the above formula, which item will now receive the highest ranking (1 mark)?

A: Item 1

15. (Critical Thinking) Suppose you wanted to design a recommender system to estimate the compatibility between candidates and job openings. Describe what data you would collect from users, how you would model the problem, and any issues that make this problem different or unique compared to those we saw in class (3 marks).
A:
Precision: \[
\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}
\]

Recall: \[
\frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}
\]

Balanced Error Rate (BER): \[
\frac{1}{2}(\text{False Positive Rate} + \text{False Negative Rate})
\]

F-score: \[
2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}
\]

Jaccard similarity: \[
\text{Sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

Cosine similarity: \[
\text{Sim}(A, B) = \frac{A \cdot B}{||A|| ||B||}
\]

**Algorithm 1** Hierarchical clustering

Initially, every point is assigned to its own cluster

while there is more than one cluster do

Compute the center of each cluster

Combine the two clusters with the nearest centers

**Algorithm 2** K-means

Initialize every cluster to contain a random set of points

while cluster assignments change between iterations do

Assign each \(X_i\) to its nearest centroid

Update each centroid to be the mean of points assigned to it

Write any additional answers/corrections/comments here: