Instructions
The test will start at 6:40pm. Hand in your solution at or before 7:40pm. 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)

Unless specified otherwise questions are each worth 1 mark.

1. The following is a list of prices from a local car dealership:

<table>
<thead>
<tr>
<th>No.</th>
<th>Model</th>
<th>Luxury?</th>
<th>Year</th>
<th>MPG</th>
<th>Horsepower</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acura MDX</td>
<td>Yes</td>
<td>2017</td>
<td>20</td>
<td>290</td>
<td>$50,000</td>
</tr>
<tr>
<td>2</td>
<td>Honda Accord</td>
<td>No</td>
<td>2017</td>
<td>25</td>
<td>190</td>
<td>$25,000</td>
</tr>
<tr>
<td>3</td>
<td>Honda Civic</td>
<td>No</td>
<td>2012</td>
<td>23</td>
<td>160</td>
<td>$10,000</td>
</tr>
<tr>
<td>4</td>
<td>Honda Civic</td>
<td>No</td>
<td>2016</td>
<td>24</td>
<td>170</td>
<td>$18,000</td>
</tr>
<tr>
<td>5</td>
<td>Nissan Altima</td>
<td>No</td>
<td>2016</td>
<td>30</td>
<td>180</td>
<td>$25,000</td>
</tr>
<tr>
<td>6</td>
<td>Acura MDX</td>
<td>Yes</td>
<td>2015</td>
<td>18</td>
<td>280</td>
<td>$38,000</td>
</tr>
<tr>
<td>7</td>
<td>Lexus RX350</td>
<td>Yes</td>
<td>2015</td>
<td>21</td>
<td>270</td>
<td>$40,000</td>
</tr>
<tr>
<td>8</td>
<td>Toyota Prius</td>
<td>No</td>
<td>2014</td>
<td>45</td>
<td>120</td>
<td>$28,000</td>
</tr>
<tr>
<td>9</td>
<td>Toyota Prius</td>
<td>No</td>
<td>2013</td>
<td>40</td>
<td>120</td>
<td>$24,000</td>
</tr>
</tbody>
</table>

Suppose you train a regressor of the following form to predict a vehicle’s price:

\[ \text{price} \approx \theta_0 + \theta_1[\text{Year}] + \theta_2[\text{MPG}] + \theta_3[\text{Is luxury?}] \]

What would be the feature representation of the first two vehicles?

1: 

2: 

2. List two additional features that might be useful for predicting the price of a car, and how you would encode them:

1: 

2: 

3. Suppose that you train two predictors on similar data to predict the price and obtain:

\[ \text{Price}^{(\text{Predictor 1})} = 40000 - 100 \times [\text{MPG}] \]
\[ \text{Price}^{(\text{Predictor 2})} = 30000 + 10000 \times [\text{Is luxury?}] + 100 \times [\text{MPG}] \]

The coefficient for MPG is negative for the first predictor, but positive for the second. Can you provide a brief explanation / interpretation of why this could be the case?

A: 

4. In class we stated that the best possible constant predictor (i.e., \( y_i \approx \alpha \)) was to set \( \alpha \) to be the mean value of \( y \) (i.e., \( \alpha = \frac{1}{N} \sum_i y_i \)). Show that this is the case when minimizing the MSE (hint: compute the derivative of the MSE and find the critical point by solving \( \alpha \)) (2 marks):

A: 

5. (Hard) What would be the best value of \( \alpha \) if our goal was instead to minimize the Mean Absolute Error (\( \frac{1}{N} \sum_i |y_i - \alpha| \))? Show your work:

A: 

Section 2: Classification and Diagnostics (8 marks)

Suppose you train two (linear) SVM classifiers, A and B, which produce the following separation boundaries:

![Separation Boundaries Diagram]

6. What is the performance of the two classifiers in terms of the following (you may leave your expressions unsimplified) (3 marks):

<table>
<thead>
<tr>
<th>Metric</th>
<th>Class A</th>
<th>Class B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision@5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. Suppose you were using your classifier to rank e-mails from ‘important’ (positive label) to ‘not important.’ Which of the two classifiers would you prefer and why?

A:

8. Imagine that the goal of a classifier is to predict whether a person is ≥ 20 years old. Two features that might be predictive include (a) height, and (b) vocabulary size. Would a Naïve Bayes classifier be suitable to train a predictor based on these two features? Explain why or why not.

A:

9. What if we added a third feature: (c) weight to our classifier from the previous question. Would this change whether a Naïve Bayes classifier was appropriate? Explain why or why not.

A:
10. (Critical thinking) A trivial classifier that we did not cover in class is a nearest neighbor classifier. This classifier has no parameters, and simply classifies points in the test set based on their similarity to points in the training set. That is, given a point $X_i$ that we wish to classify, we consider all $X_j$ in the training set, and select the label of the nearest one:

$$y_i = y_{\text{argmin}_j \|X_i - X_j\|_2^2}$$

Describe two settings (e.g. applications, properties of datasets, computational resources available, etc.) in which the nearest neighbor classifier would be (1) preferable to logistic regression, and (2) less preferable than logistic regression (2 marks)

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

When asked to draw examples, provide 2-d sets of points and/or clusters like the following:

11. Consider running the clique percolation algorithm with $K = 3$ on the following graph (see pseudocode on final page of exam):

what are the communities found by the algorithm? (you can draw your solution directly on the graph)

12. Using the boxes below, draw examples of sets of 2-d point sets for which
   
   (a) PCA would be more appropriate than hierarchical clustering
   (b) Hierarchical clustering would be more appropriate than PCA
   (c) Neither hierarchical clustering nor PCA would be appropriate

   (3 marks)

13. For the examples above, describe a real pair of features that might be described by the points you drew.
   ((b) is provided as an example) (2 marks):

   dimension 1: 
   dimension 2: 
   (a) 

   dimension 1: 
   dimension 2: 
   (b) 

   dimension 1: 
   dimension 2: 
   (c)
Section 4: Recommender Systems (7 marks)

On a popular music streaming website, a few users have listened to the following music:

<table>
<thead>
<tr>
<th>Album</th>
<th>Listened?</th>
<th>Liked?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nathan</td>
<td>Thomas</td>
</tr>
<tr>
<td></td>
<td>Nathan</td>
<td>Thomas</td>
</tr>
<tr>
<td>Lana Del Ray</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Born to Die</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>Ultraviolence</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Honeymoon</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Lust for Life</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

14. Suppose you want to determine which users are similar to each other in terms of their listening behavior. What would be an appropriate metric for determining users’ similarity, and which two users would be most similar under this metric (list multiple in case of a tie)? (2 marks)

A:

15. Suppose you want to determine which users are similar to each other in terms of their preferences. What would be an appropriate metric for determining users’ similarity, and which two users would be most similar under this metric (list multiple in case of a tie)? Describe how you handle the ‘?’ entries. (2 marks)

A:

16. (Critical Thinking) Suppose you wanted to design a recommender system to suggest points of interest in a city based on users’ past activities/behavior/etc. Describe what data you would collect from users, how you would model the problem, and any issues that make this problem different from 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: \[ \frac{1}{2} (\text{False Positive Rate} + \text{False Negative Rate}) \]

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

Naïve Bayes:
\[ p(\text{label}|\text{features}) \approx \frac{p(\text{label}) \prod_i p(\text{feature}_i|\text{label})}{p(\text{features})} \]

Algorithm 1 Clique percolation with parameter k
Initially, all k-cliques in the graph are communities

while there are two communities that have a (k − 1)-clique in common do
merge both communities into a single community

Algorithm 2 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

Write any additional answers/corrections/comments here: