CSE 258, Winter 2017: Midterm

Name:	Student ID:

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

- Show that when predicting a constant, predicting the average is the best predictor in terms of minimizing the MSE
- What if we're minimizing the MAE? What is the best predictor then?

Section 1: Regression and Ranking (7 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

The following is a list of Vin Diesel's recent films:

No.	Title	Year	IMDB score	MPAA rating	length in minutes	classifier score
1	XXX: The Return of Xander Cage	2017	5.6	PG-13	110	-0.5
2	Billy Lynn's Long Halftime Walk	2016	6.6	R	113	-2.1
3	The Last Witch Hunter	2015	6.3	PG-13	106	-3.2
4	Furious 7	2015	7.4	PG-13	137	4.8
5	Guardians of the Galaxy	2014	8.1	PG-13	121	2.2
6	Riddick	2013	6.4	R	119	-1.2
7	Fast & Furious 6	2013	7.2	PG-13	130	-0.8
8	Fast Five	2011	7.3	PG-13	131	1.2
9	Fast & Furious	2009	6.6	PG-13	107	0.1
10	The Fast and the Furious: Tokyo Drift	2006	6.0	PG-13	104	-0.3

1. Suppose you train a regressor of the following form to predict the IMDB score:

$$\text{IMDB score} \simeq \theta_0 + \theta_1[\text{`Fast' in title}] + \theta_2[\text{`R' rated}] + \theta_3[\text{length in minutes}]$$

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

1:

2:

2. After training the above regressor you obtain $\theta = (1.5, 0.05, -0.25, 0.05)$. What would you predict would be the IMDB score of Vin Diesel's *next* film, *The Fate of The Furious* (released 2017, PG-13, 140 minutes long). You can write down an expression rather than the exact value:

A:

Next, you train a Support Vector Machine to predict the binary outcome 'IMDB score ≥ 7.0' using the same features. Suppose the classifier produces the scores $(X_i \cdot \theta)$ shown in the right column of the table.

3. What is the accuracy, and the Balanced Error Rate of this classifier?

A:

4. What is the classifier's precision and recall?

A:

5. Perhaps you would like to add the 'year' variable to your classifier. Assuming a simple model with no regularizer, show that using the feature [year] is equivalent to using the feature [year − 2006].

A:

6. (Hard) Briefly explain why these two representations would *not* be equivalent when training a model with a regularizer (e.g. $\|\theta\|_2^2$) (**2 marks**).

A:

Section 2: Classification and Diagnostics (6 marks)

Each of the following questions is worth **2 marks**.

Suppose you are trying to train a classifier to detect whether bicycle frames will catastrophically fail during five years of use (i.e., $y_i = 1$ if the i^{th} bicycle fails). You build a dataset containing features about 10,000 bicycle frames manufactured between 1978 and 2012 (weight, material, manufacture year, etc.) and whether or not they failed. You partition the dataset into a training and validation set using a 50%/50% split.

After trying several different classifiers on your data and measuring their error (percentage of incorrect classifications), you obtain some unexpected results. Briefly explain a possible reason for the results and suggest a possible solution.

7. You obtain low (1%) error on your training set but even *lower* error (0.5%) on your validation set.

Diagnosis:

Solution:

8. You build an accurate classifier with only 1% error on your training set, and 1.5% error on your validation set. However, when you deploy the system, it fails to identify any instances of catastrophic failure.

Diagnosis:

Solution:

9. Suppose that to fix the above issue, you want to adjust a logistic regression-based classifier so that it gives 100 times as much weight to false negatives (bicycles that fail but were predicted not to) as to false positives. How would you adjust the objective function to achieve this? Recall that the original objective for logistic regression is

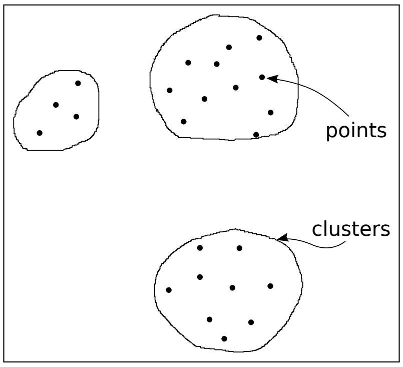
$$\sum_{y_i=1} \log \sigma(X_i \cdot \theta) + \sum_{y_i=0} \log(1 - \sigma(X_i \cdot \theta))$$

A:

Section 3: Clustering / Communities (8 marks)

The following questions are concerned with the *K-means* algorithm (see pseudocode at the end of the exam). Each question is worth **2 marks**.

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



10. Explain why the algorithm provided in the pseudocode will eventually converge (i.e., terminate).

A:

11. The K-Means algorithm will in general converge to a local optimum rather than a global one. Draw a simple 2-d example, containing a set of points and clusters, such that the solution is *not* optimal but for which the algorithm would not make further progress.

A:

12. Suggest simple modifications to the k-means algorithm that might increase its chances of finding a good solution.

A:

13. Compared to PCA, K-means will work well for different types of clusters. Give three examples of 2-d clustered data where (a) K-means will perform *better* than PCA (in terms of reconstruction error); (b) K-means will perform *worse* than PCA; and (c) K-means and PCA will both perform poorly.

A:

Section 4: Recommender Systems (5 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

On a popular movie streaming website, a few users have watched the following recent movies:

Movie	Watched?				Rated?			
	Caroline	Mengting	Ruining	Zachary	Caroline	Mengting	Ruining	Zachary
XXX: Return of Xander Cage	1	1	0	1	5	1		2
La La Land	1	1	1	1	5	2	2	2
John Wick 2	1	1	1	0	4	2	1	
Rogue One	0	0	1	1			5	1
Resident Evil	1	0	0	1	4			1

14. Using 'watched' data: Which two users are *most similar* in terms of their Jaccard similarity (write down all pairs in case of a tie)?

A:

15. Which two *items* are most similar in terms of their Jaccard similarity?

A:

16. Which two users are most similar in terms of their *ratings*, based on their Pearson correlation (defined below)?

A:

$$\text{Sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I_u \cap I_v} (R_{v,i} - \bar{R}_v)^2}}$$

(R = rating matrix, I_u = items rated by user u, $\bar{R}_u$ = average rating by user u)

17. (Hard) Show that a latent factor model of the form

$$\text{rating}(u,i) = \alpha + \beta_u + \beta_i$$

is linear in its parameters ($\theta = (\alpha; \beta_u; \beta_i)$) but that a model of the form

$$\text{rating}(u,i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

is not linear in its parameters ($\theta = (\alpha; \beta_u; \beta_i; \gamma_u; \gamma_i)$) (recall the definition of linearity: $r_{\theta_1+\theta_2}(u,i) = r_{\theta_1}(u,i) + r_{\theta_2}(u,i)$) (**2 marks**).

A:

Algorithm 1 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
```

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:

$$\frac{1}{2}(\text{False Positive Rate} + \text{False Negative Rate})$$

Jaccard similarity:

$$\text{Sim}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Write any additional answers/corrections/comments here: