CSE 258, Fall 2017: Midterm

Name: JIGYASA GROVER	Student ID: A53234810

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

This study resource was shared via CourseHero.com

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/

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:

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

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

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

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

1: [1 2017 20 1]

2: [1 2017 25 0]

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

1: The type of fuel it runs on, for example: 0 → petrol, 1 → diesel, 2 → CNG etc.

2: Capacity / number of seats the car has, ie: 2 → 2 seater car, 4 → 4 seater car etc

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: In predictor 1, the luxury factor is not taken into account & thus it indicates that as the value of MPG increases, the price seems to drop.
But if the car's luxury factor is accounted for, we see a new trend in the pricing of the car.

4. In class we stated that the best possible constant predictor (i.e., $y_i \simeq \alpha$) was to set α 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 α) (**2 marks**):

A: $MSE = \frac{1}{N}\sum_i (y_i - x_i\theta)^2$

$\frac{\partial MSE}{\partial \theta} = \frac{1}{N} 2\sum_i (y_i - x_i\theta)(x_i)$ → for this to be zero $\frac{1}{N}\sum_i x_i\theta = \frac{1}{N}\sum_i y_i$ gives α the mean value of y (Trivial predictor)

$\frac{\partial^2 MSE}{\partial \theta}$ ⇒ to see if minima or not

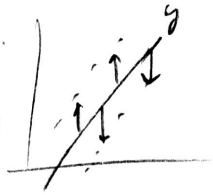
also for trivial predictor we get FVU = 1 ie MSE to be eq to variance ie $x_i\theta = \bar{y}_i$ QED.

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

A: $MAE = \frac{1}{N}\sum_i |y_i - x_i\theta|$

In my opinion, the mean value ie $\alpha = \frac{1}{N}\sum_i y_i$ would again give the desired results

taking a very simple case taking the mean value of y shall minimize the mean absolute error though it penalizes in proportion



This study source was downloaded by 100007680 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

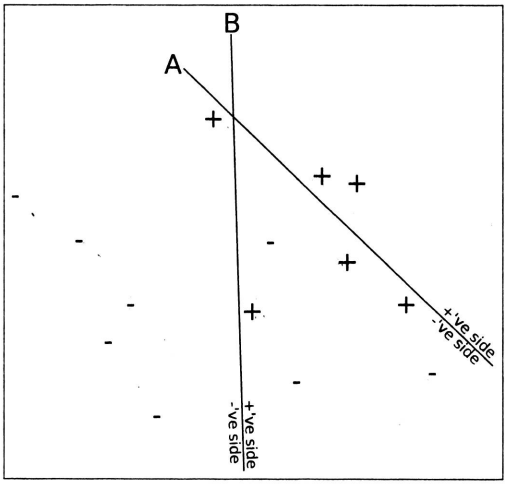
https://www.coursehero.com/file/34472944/MIDTERMpdf/

Section 2: Classification and Diagnostics (8 marks)

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

	A	B
TP :	2	5
TN :	8	5
FP :	0	3
FN :	4	1

Total : 14



Retrieved / Relevant

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

	A:	B:
Accuracy:	A: $10/14$	B: $10/14$
BER:	A: $1 - 1/2\left[2/6 + 8/8\right]$	B: $1 - 1/2\left[5/6 + 5/8\right]$
Precision:	A: $2/10$	B: $5/10$
Recall:	A: $2/6$	B: $5/6$
F-score:	A: $\frac{2 * 2/10 * 2/6}{2/10 + 2/6}$	B: $\frac{2 * 5/10 * 5/6}{5/10 + 5/6}$
Precision@5:	A: $2/5$	B: $5/5$

TPR = 2/6

TPR = 5/6.

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: In my opinion, given a choice I would use ~~predictor~~ classifier B to label my EMails correctly as it gives a better true positive rate in comparison to classifier A.
Also comparing the precision & recall performance metric of B with A.

8. Imagine that the goal of a classifier is to predict whether a person is $\geq$ 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: In my opinion, Naive Bayes classification can be used as a very rudimentary classifier for the above mentioned scenario as the two features "height" & "vocabulary size" are supposedly independent (cond') of each other.

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: In my opinion, we might not be able to use Naive Bayes as our best choice of classifier as "usually" the height & the weight of a human being are corelated. Using these two along with the vocabulary size might not lead to the solution as our assumption of features being conditionally indep would be false in that case.

This study source was downloaded by 100076866883 from CourseHero.com on 10-12-2021 19:45:43 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/

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:

① Preferable to logistic regression

eg for example in a scenario where a user has to be recommended outlets for food delivery KNN will be suitable to logistic regression in that case.

KNN in this case also save computation cost if data/sample set aint that huge.

② less preferrable than logistic regression

Expensive to train on as it requires to check the distance from all the data points For large datasets it shall be huge cost.

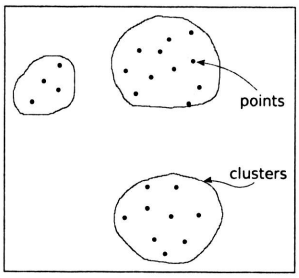
This study resource was shared via CourseHero.com

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

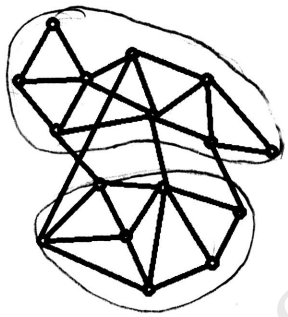
https://www.coursehero.com/file/34472944/MIDTERMpdf/

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)

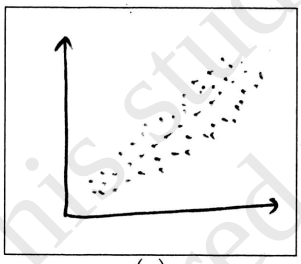
12. Using the boxes below, draw examples of sets of 2-d point sets for which

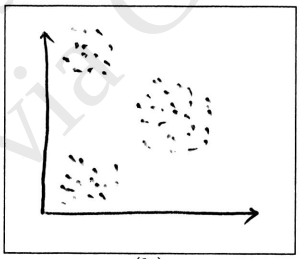
    (a) PCA would be more appropriate than hierarchical clustering

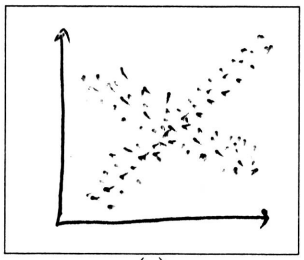
    (b) Hierarchical clustering would be more appropriate than PCA

    (c) Neither hierarchical clustering nor PCA would be appropriate

    (3 marks)


(a)


(b)


(c)

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)**:

(a)	(b)	(c)
dimension 1: Smell of the beer	dimension 1: Latitude	dimension 1: price of running shoes
dimension 2: taste of the beer	dimension 2: Longitude	dimension 2: weight of running shoes

Assuming that costlier the shoes are, better design & hence lighter the weight

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/

Section 4: Recommender Systems (7 marks)

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

Album	Listened?					Liked?				
	Nathan	Thomas	Dhruv	Kevin	Prateek	Nathan	Thomas	Dhruv	Kevin	Prateek
Lana Del Ray	1	0	1	1	0	1	?	1	-1	?
Born to Die	1	0	0	1	0	-1	?	?	1	?
Ultraviolence	0	1	1	1	0	?	1	-1	-1	?
Honeymoon	0	1	1	0	0	?	1	1	?	?
Lust for Life	1	1	0	1	1	-1	-1	?	1	-1

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: Using Jaccard's measure of similiarity.
NT = 1/5 TD = 2/4 DK = 2/5
ND = 1/5 TK = 2/5 DP = 0/4
NK = 3/4 TP = 1/3 KP = 1/4
NP = 1/3
⇒ Nathan & Kevin (most similar)

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: In this scenario, I would use Cosine Similarity and treat ? as 0 ie not listened to the album at all.
NT = 1/9 NK = -3/12 TD = 0 TP = 1/3 DP = 0
ND = 1/9 NP = 1/3 TK = 2/6 DK = 0 KP = 1/4
Seems like Thomas & Prateek are similar.

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: To suggest points of interests.
Let each point of interest be a real valued ID number.
We'll collect the following data of the users :-
- user's location
- user's gender
- user's age
- user's prev visited PoI

} feature set

We can now solve this as a linear regression problem

- Another method could be of applying Binary Classifier (ie recommend or not recommend) on each of the PoI

In cases where a user hasn't visited any place, "cold start" problem perhaps. Also requires lots of data.

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/

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:

$$p(\text{label}|\text{features}) \simeq \frac{p(\text{label}) \prod_i p(\text{feature}_i|\text{label})}{p(\text{features})}$$

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:

This study resource was shared via CourseHero.com

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/

This study resource was shared via CourseHero.com

This study source was downloaded by 100000768665883 from CourseHero.com on 10-12-2021 19:45:15 GMT -05:00

https://www.coursehero.com/file/34472944/MIDTERMpdf/