Carnegie Mellon University in Qatar

AI for Medicine

15-182/282 - Spring 2021

Assignment 3

Name: ______________________________

Andrew ID: ______________________________

Due on: April 14, 2021 by midnight

Instructions:

- This assignment has maximum scores of 100 points for 15-282 and 42 for 15-182.
- You should submit your solutions through Gradescope.

Question	Points	Score
Inference Using Logistic Regression	10	
Learning Using Logistic Regression	40	
Recommending Doctors Using AI	50	
Total:	100	

Problem 1: Inference Using Logistic Regression (10 Points)

Suppose you have collected data for past students in the "All for Medicine" (Not "AI for Medicine"!) class with features x_1 = “Hours Studied” and x_2 = “Cumulative GPA”, and label y = “Received an A” or y = “Did Not Receive an A”. Assume also that you have learnt a logistic regression model out of this data and ended up with a parameter vector $\theta = [-6,\ 0.05,\ 1]$.

Answer the following questions and make sure to show all your work:

5pts (a) Estimate the probability that a student who studies for 40 hours and has a cumulative GPA of 3.5 gets an "A" in the class.

5pts (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an "A" in the class?

Assignment continues on the next page(s)

Problem 2: Learning Using Logistic Regression (40 Points)

The following training dataset assumes two classes, "1" and "0", and obeys the rule that the examples of class "1" all have vectors whose components sum to 10 or more, while the sum is less than 10 for the examples of class "0".

([3, 4, 5], 1) ([2, 7, 2], 1) ([5, 5, 5], 1)

([1, 2, 3], 0) ([3, 3, 2], 0) ([2, 4, 1], 0)

Answer the following questions and make sure to show and submit all your work:

4pts (a) Propose a parameter vector θ such that the hypothesis function defined by $\frac{1}{1+e^{-\theta^T x}}$ renders a good classifier for the "1" and "0" examples.

22pts (b) Starting with your answer to part (a), use gradient descent to find the "optimal" θ in less than 15 rounds, assuming a learning rate of 0.5. In this problem, we consider "optimum" to be the case when *every* example in the training dataset is predicted correctly by the hypothesis function. When this happens, you can assume that gradient descent has converged and stop training. (***Hint***: your choice in part (a) impacts the number of rounds gradient descent will take to converge; hence, you may want to revisit part (a) if it does not converge in less than 15 rounds).

6pts (c) Use the parameter vector learnt in part (b) to infer the classes of several new (say, 5 to 10) different feature vectors of your choice, some whose components sum to less than 10 and some whose components sum to 10 or more. How many of these vectors were classified correctly and why in your opinion some has been misclassified, if any?

8pts (d) **This part is only for 15-282**: Write Python Code to train a logistic regression model using the above training dataset. How many rounds did your code take to converge to the "optimal" parameter vector (as defined in part (b))?

Assignment continues on the next page(s)

Problem 3: Recommending Doctors Using AI (50 Points)

This problem is ONLY for 15-282. In it, you will build a recommendation system that recommends doctors to patients.

Table 1 captures the ratings (between 0 and 5, inclusive) of several patients for several doctors. For example, as shown in Table 1, Patient 1 has given a rating of 4.5 to Doctor 1, while Patient 4 has given a rating of 0 to the same doctor.

	Patient 1	**Patient 2**	**Patient 3**	**Patient 4**
Doctor 1	4.5	4.9	0	0
Doctor 2	3.8	4.6	0	0
Doctor 3	?	4.3	0	0
Doctor 4	0	0	4	5
Doctor 5	0	0	5	?

Table 1: Ratings of Patients for Doctors (ratings are between 0 and 5; ? entails no rating)

Answer the following questions, assuming a learning rate $\alpha = 0.5$, and make sure to show and submit all your work:

10pts

(a) Suppose you are given the following table (i.e., Table 2) that captures certain values between 0 and 1 for two features x_1 and x_2 about each doctor in our dataset shown in Table 1.

	x_1	x_2
Doctor 1	1	0.1
Doctor 2	1	0
Doctor 3	0.8	0.2
Doctor 4	0.3	1
Doctor 5	0.1	1

Table 2

Predict the missing values in Table 1 (i.e., the ratings of Patient 1 for Doctor 3 and Patient 4 for Doctor 5) using a content-based recommendation algorithm. Write Python code to implement your algorithm. You can stop your algorithm after 10 rounds and make the predictions accordingly.

20pts

(b) Suppose you are now given the parameter vectors for Patients 1, 2, 3, and 4 as follows:

$\theta_1 = [0, 5, 0]$
$\theta_2 = [0, 5, 0]$

$\theta_3 = [0, 0, 5]$

$\theta_4 = [0, 0, 5]$

On the flip side, assume you are NOT given the values of the two features x_1 and x_2 about each doctor (see Table 3).

Predict the missing values in Table 3 using a collaborative filtering algorithm. Write Python code to implement your algorithm. You can stop your algorithm after 15 rounds and make the predictions accordingly.

	x_1	x_2
Doctor 1	?	?
Doctor 2	?	?
Doctor 3	?	?
Doctor 4	?	?
Doctor 5	?	?

Table 3

For parts (c), (d), (e), (f), and (g) assume Patient 1 has given Doctor 3 a rating of 4.1 and Patient 4 has given Doctor 5 a rating of 5. Clearly, this completes Table 1.

4pts (c) Write Python code to produce and show the TF.IDF matrix that corresponds to Table 1.

4pts (d) Write Python code to produce and show the normalized TF.IDF matrix that corresponds to Table 1.

4pts (e) Who is the closest patient to Patient 4? Write Python code to answer this question (***Hint***: think about applying cosine similarity to rank all the patients against Patient 4).

4pts (f) Who is the closest doctor to Doctor 4? Write Python code to answer this question (***Hint***: think about applying cosine similarity to rank all the doctors against Doctor 4).

4pts (g) If the closest patient to Patient 4 has seen Doctor 4 before, but Doctor 4 is currently not available (their schedule is closed for the coming two months due to COVID-19!). Which doctor would you recommend for Patient 4 and why? Write Python code to answer this question.