


National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Data Science	Course Code:	CS4048
	Degree Program:	BS(Computer Science)	Semester:	Spring 2022
	Exam Duration:	60 Minutes	Total Marks:	68
	Paper Date:	9-May-2022	Weight	10%
	Section:	ALL	Page(s):	4
	Exam Type:	Midterm-2		

Student : Name: _____ **Roll No.** _____ **Section:** _____

Instruction/Notes:

1. Provide precise answers in the given space. You don't have to write complete sentences. A phrase consisting of only a few words that conveys the point is expected. For example you can write "Yes, because ..." or provide a value/expression according to the question.
2. Longer answers will be discarded. (E.g. Using small handwriting or putting multiple lines in place of single line space)

Q1. [14 marks]

Following is the cost function for collaborative filtering recommender system algorithm:

$$J(x, \theta) = \frac{1}{2} \sum_{(i,j):r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (x_k^{(i)})^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^n (\theta_k^{(j)})^2$$

- n_u = number of users
 - n_m = number of movies
 - $r(i, j) = 1$ if user j has rated movie i
 - $y(i, j) =$ rating given by user j to movie i (defined only if $r(i,j)=1$)
1. Identify the parameters that the collaborative filtering algorithm learns during the training.
It learns both x and θ parameters. x being movie features and θ being weights to determine ratings.
 2. What is being represented by x in the collaborative filtering algorithm?
 x being movie features.
 3. Once the collaborative filtering algorithm learning completes, how do we get ratings for movies not rated by user j ?
Rating for user $j = (\theta_j^T x)$
 4. PCA is a linear dimensionality reduction algorithm. What is the relationship of features extracted using PCA with the original features of data?
The extracted features are a linear combination of original features.

- One way to reduce overfitting is to remove unnecessary/redundant features and use only a subset of features relevant to the problem. Can we use PCA to extract a small number of features in this scenario to address overfitting? Why or why not?
No. Because the redundant features will also be factored in the extracted features.
- In what situation should you apply anomaly detection instead of classification?
When there are very few positive examples. OR The positive examples are of many different types
- When is the advantage of performing anomaly detection using “Multivariate Gaussian Distribution” instead of the original model?
It captures the correlations between different features.

Q2. [30 marks]

	Hypothesis	Cost Function
Linear Regression	$h_{\theta}(x) = \theta^T x$	$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$
Logistic Regression	$h_{\theta}(x) = g(\theta^T x)$ $z = \theta^T x$ $g(z) = \frac{1}{1 + e^{-z}}$	$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$ $\text{Cost}(h_{\theta}(x), y) = -\log(h_{\theta}(x)) \quad \text{if } y = 1$ $\text{Cost}(h_{\theta}(x), y) = -\log(1 - h_{\theta}(x)) \quad \text{if } y = 0$
Support Vector Machine	$h_{\theta}(x) = \begin{cases} 1 & \text{if } \Theta^T x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$J(\theta) = C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \frac{1}{2} \sum_{j=1}^n \Theta_j^2$ $\text{cost}_0(z) = \max(0, k(1 + z))$ $\text{cost}_1(z) = \max(0, k(1 - z))$

- What is the range of linear regression hypothesis?
 $-\text{Infinity} < h(x) < \text{Infinity}$
- What is the range of logistic regression hypothesis?
 $0 \leq h(x) \leq 1$
- What is the range of SVM hypothesis?
 $\{0, 1\}$
- What complication would arise if logistic regression reuses the cost function of linear regression?
The cost function will not be convex introducing the risk of local minima.
- What is the value of logistic regression cost function for $y = 1$ and $h(x)$ approaches 0?
J approaches Infinity.
- What is the value of logistic regression cost function for $y = 0$ and $h(x)$ approaches 1?

J approaches Infinity.

7. What is the value of logistic regression cost function for $y = 1$ and $h(x)$ approaches 1?

J approaches zero.

8. What is the value of logistic regression cost function for $y = 0$ and $h(x)$ approaches 0?

J approaches zero.

9. What is the value of SVM cost function for $y=0$, $k=1$ and z approaches Infinity

J approaches Infinity.

10. What is the value of SVM cost function for $y=1$, $k=1$ and z approaches Negative Infinity

J approaches Infinity.

11. For what values of $\theta^T x^{(i)}$, the first term of SVM cost function will be zero for $y = 0$

$$\theta^T x^{(i)} \leq -1$$

12. For what values of $\theta^T x^{(i)}$, the first term of SVM cost function will be zero for $y = 1$

$$\theta^T x^{(i)} \geq 1$$

13. For SVM cost function when the value of C is very large, how the optimization algorithm will minimize the cost function. State your intuition.

It will try to make the first term zero by keeping $\theta^T x^{(i)}$ less than -1 for $y=0$ and more than 1 for $y=1$

14. Rewrite the linear regression cost function with regularization

$$\frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right]$$

15. What exactly is the impact on parameters when minimizing a regularized cost function with a large value of λ ?

The values of parameter θ are pushed to be smaller in order to minimize the function.

Q3. [24 marks]

Consider the regularized cost function for linear regression and logistic regression and answer the following questions

1. If your model has high variance, what can you do to try to address that?

Increase the value of λ

2. If your model has underfitting, what can you do to try to address that?

Decrease the value of λ

3. In what situation should you consider adding more features to your data?

In case of high bias or underfitting

4. In what situation should you consider collecting more data points?

In case of high variance or overfitting

5. In what situation should you consider dropping some of the features from your data?

In case of high variance or overfitting

Consider logistic regression hypothesis $h(\mathbf{x})$ for binary classification. In the following three cases when you set different thresholds for prediction, what is the impact on Precision and Recall?

6. You predict class 1 if $h(\mathbf{x}) \geq 0.5$

Neutral for balanced classes OR proportionate to the number of examples of each class.

7. You predict class 1 if $h(\mathbf{x}) \geq 0.8$

High Precision OR low Recall

8. You predict class 1 if $h(\mathbf{x}) \geq 0.2$

High Recall OR low Precision

Recall that m is the number of data points and n is the number of features in a dataset. Answer the following questions:

9. If m is small and n is large, should you consider adding more features? Why or why not?

No, because complex features will increase variance/overfitting with less examples.

10. If n is small and m is large, should you consider adding more features? Why or why not?

Yes, because complex features will allow the model to learn better and decrease bias/underfitting.

11. SVM with kernel allows us to remain in feature space but get the effect of working in instance space. How does the similarity function conceptually facilitate classification?

Conceptually the similarity function generates new features according to the location of data points.

12. If m is large, should you consider using SVM with kernel? Why or why not?

No, because it will be computationally expensive as the $X^T X$ matrix becomes very large.