Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Data Science**	**Course Code:**	**CS481**
Program:	**BS (Computer Science)**	**Semester:**	**Spring 2017**
Duration:	**180 Minutes**	**Total Marks:**	**47**
Paper Date:	**28-May-18**	**Weight**	**45 %**
Section:	**All**	**Page(s):**	**12**
Exam:	**Final**		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	Objective	1	2	3	4	**Total**
Marks	/ 10	/ 6	/ 5	/ 15	/9	/ **45**

Section 1

(Objective part) [points 10]

Clearly circle the correct options.

Q1. Let f be some function so that $f(\Theta_0, \Theta_1)$ outputs a number. For this problem, f is some arbitrary/unknown smooth function (not necessarily the cost function of linear regression, so f may have local optima). Suppose we use gradient descent to try to minimize $f(\Theta_0, \Theta_1)$ as a function of Θ_0 and Θ_1. Which of the following statements are true? (select all that apply.)

(A) If the first few iterations of gradient descent cause $f(\Theta_0, \Theta_1)$ to increase rather than decrease, then the most likely cause is that we have set the learning rate α to too large a value.

(B) If the learning rate α is too small, then gradient descent may take a very long time to converge.

(C) Even if the learning rate α is very large; every iteration of gradient descent will decrease the value of $f(\Theta_0, \Theta_1)$.

(D) No matter how Θ_0 and Θ_1 are initialized, so long as α is sufficiently small, we can safely expect gradient descent to converge to the same solution.

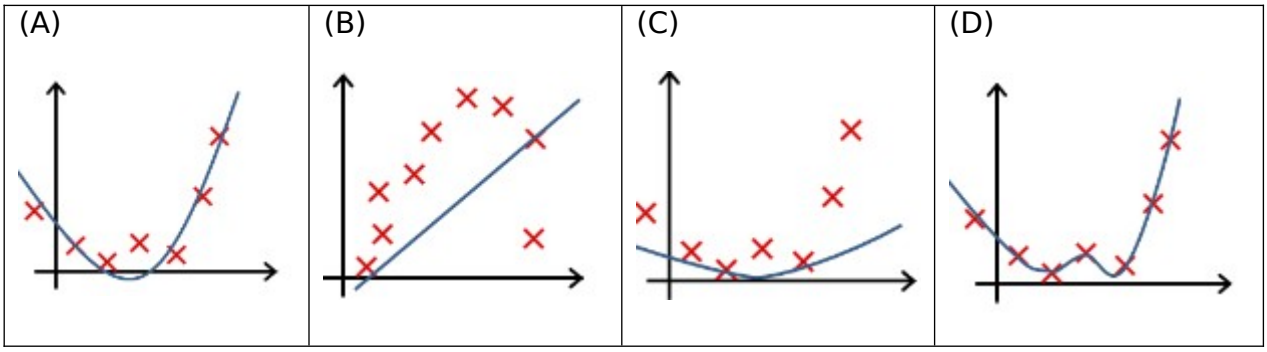
Name: _____ Reg #: _____ Section: _____

Q2. Suppose that you have trained a logistic regression classifier, and it outputs a new example **x** a prediction $\mathbf{h_\theta(x) = 0.7}$. This means (select all that apply):

(A) our estimate for P(y = 0| x; ɵ) is 0.3 (B) our estimate for $P(y = 0| x; \Theta)$ is 0.7

(C) our estimate for $P(y = 1| x; \Theta)$ is 0.3 (D) our estimate for $P(y = 1| x; \Theta)$ is 0.7

Q3. In which of the following figure do you think the hypothesis is over-fitting the training set?



Q4. Will the removal of non-support vector points affect the decision boundary for SVM?

a) True b) False

Q5. You are training a classification model with logistic regression, which of the following statement are true. Select all that apply.

(A) Introducing regularization to the model always results in equal or better performance on examples not in the training set.
(B) Adding many new features to the model helps prevent overfitting on the training set.
(C) Adding many new features to the model makes it more likely to overfit the training set.
(D) Adding a new feature to the model always results in equal or better performance on examples not in the training set.

Q6. Which statements are true about Data Wrangling?

(A) We should use imputation when we have a lot of data.
(B) Due to partial deletion, we can compromise the representativeness of our sample.
(C) Imputation is the process of approximating the missing values.
(D) Pairwise deletion is more useful when we have only few records (data).

Q7. Which of the following statement about regularization are true. Select all that apply.

(A) Using too large value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.
(B) Using too large value of λ can cause your hypothesis to underfit the data.

Name: _____ Reg #: _____ Section: _____

(C) Using too small value of λ can cause your hypothesis to overfit the data.
(D) Using very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.

Q8. Suppose you have a dataset with $m = 1000000$ examples and $n = 15$ features for each example. You want to use multivariate linear regression to fit the parameters θ to our data. Should you prefer gradient descent or the normal equation?

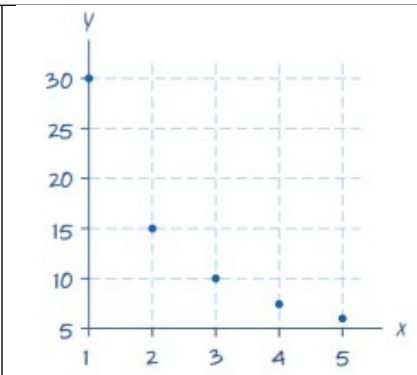
A. Gradient descent, since $(X^T X)^{-1}$ will be very slow to compute in the normal equation.
B. Gradient descent, since it will always converge to the optimal θ.
C. The normal equation, since gradient descent might be unable to find the optimal θ.
D. The normal equation, since it provides an efficient way to directly find the solution.

Q9. In the context of regression analysis, which of the following statements are true?

I. When the sum of the residuals is greater than zero, the data set is nonlinear.
II. A random pattern of residuals supports a linear model.
III. A random pattern of residuals supports a non-linear model.

(A) I only (B) II only (C) III only (D) I and II (E) I and III

Q10. The relationship between two variables y and x as shown in the scatterplot is non-linear.

Which Transformation is most likely to linearize the relationship? **Ans**: ----------	

Section 2 (Subjective part) (marks 35)

Q1: [5 + 1 marks] Principal Component Analysis (PCA):

Part-I: [5 marks] Suppose you want to train a classifier, the data set (with 2 features) is given in the following table. In the table, we can see that there are 3 training examples (m=3). Assume the given data is also already mean normalized, and you don't need to scale it.

X_1	X_2	Y (class label: y=1 or y=0)
1	2	1

Name: _____ Reg #: _____ Section: _____

2	1	0
1	5	1

As a data scientist, your task is to apply PCA and reduce the given dataset to one dimensional data. Answer the following questions.

a) What are the Eigen vectors for the data set? Write down all the Eigen vectors in matrix form (U matrix in PCA).
b) What will be the projected data (Z) in reduced form? Write the reduced form for all the training examples.

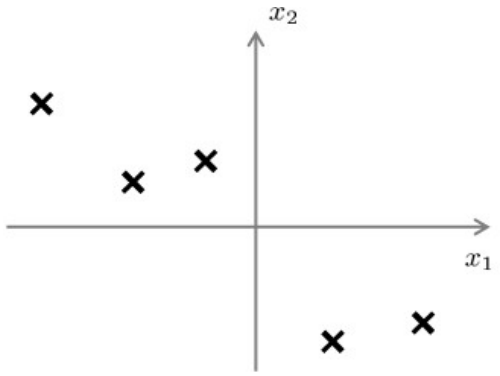
The helpful formulas are given as follows:

A (Covariance Matrix) $¿\frac{1}{m}(X^T.X)$, where X is mxn data matrix.

$Z^{(i)} = (U_{reduce})^T * x^{(i)}$

Name: _____ Reg #: _____ Section: _____

Part-II: [1 mark] Considering the dataset given below in the figure, find all principal components (Eigen Vectors). You don't need any calculation, just draw and show the direction of the vectors?

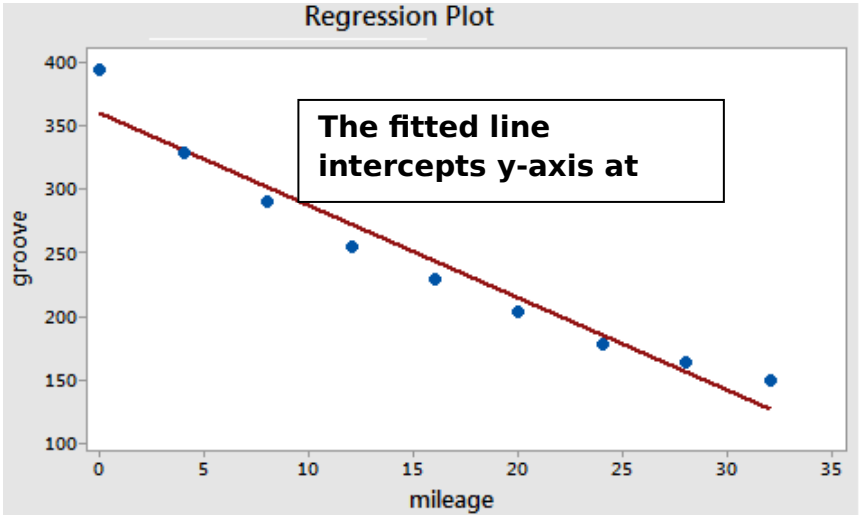


Q2. [5 marks] A scientific laboratory conducted an experiment in order to answer the following research question: "**Is tire tread wear linearly related to mileage?"**
As a result of the experiment, the researchers obtained a data set as shown in Table below, containing the mileage (x, in 1000 miles) driven and the depth of the remaining groove (y, in mils).

mileage	groove
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

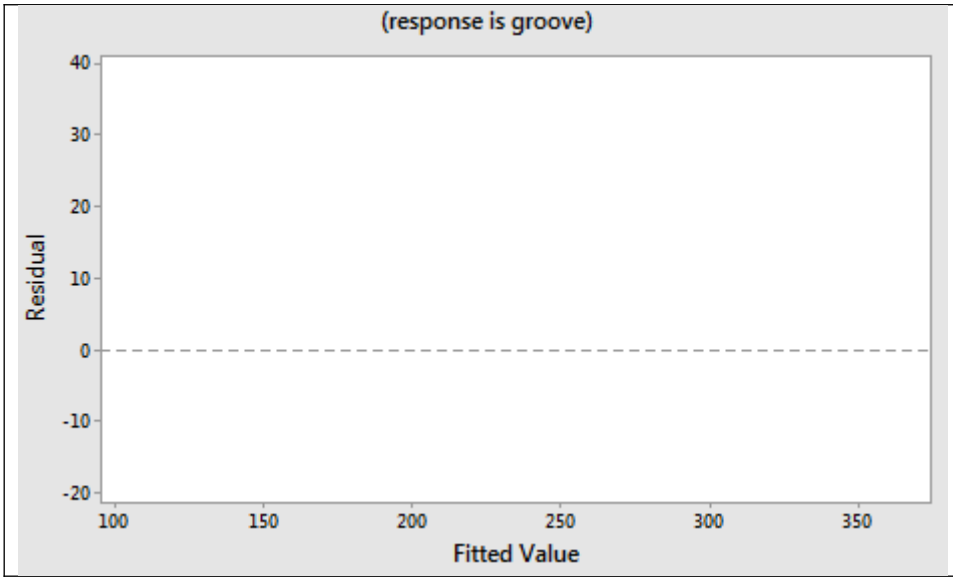
The  fitted line plot of the resulting data suggests that there is a relationship between groove depth and mileage.

Name: _____ Reg #: _____ Section: _____



a) What will be the value of θ_0 and θ_1 for our simple linear hypothesis: **groove = θ_0 + θ_1 mileage**

b) Plot the raw residual of the simple linear regression model of the given data set against the fitted value (predicted values). Plot the raw residuals on the Figure below, and moreover, fill in the table with Fitted (Predicted) values and Residuals.



	Fitted value	**Residual**
1		
2		
3		
4		
5		
6		
7		
8		
9		

c) How does your Residual plot helps to determine if your regression function is linear or not-linear.

Name: _____ Reg #: _____ Section: _____

Q3. [15 Marks] **Short Questions**:

A) [2 points] You are a reviewer for the top Impact Factor journal, and you read papers with the following experimental setups. Would you accept or reject each paper? Provide a one sentence justification.
i). **accept/reject** "The proposed algorithm in the paper is better than the algorithm B as far as training accuracy is concerned."

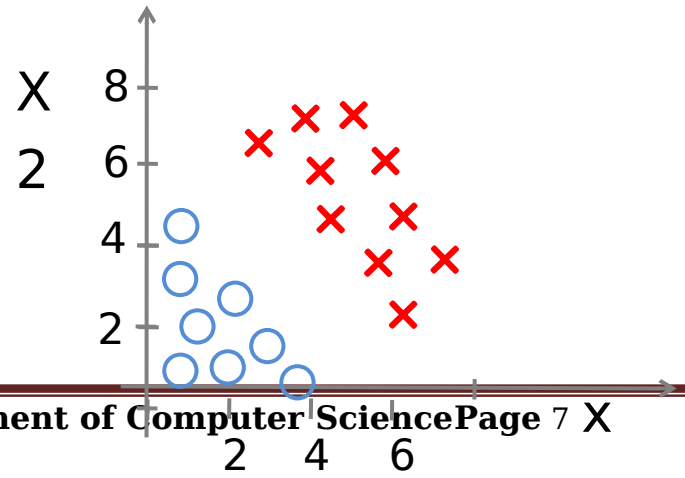
ii). **accept/reject** "The proposed algorithm in the paper is better than the algorithm B as far as test accuracy is concerned. Suppose the λ is selected based on the cross validation data."

B) [2 marks] Logistic Regression – Decision Boundary:

We consider the following model of logistic regression for binary classification with a sigmoid function

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1^2 + \theta_2 x_2)$$

Model:



Suppose the trained parameter values are $\Theta_0 = -8$, $\Theta_1 = 2$ AND $\Theta_2 = 2$

Predict ''y = 1'' if h(x) >= 0.25

Calculate and Draw the decision boundary according to the threshold given above. Show your working here. If you just draw the boundary without working, you will not get any point.

C) [2 marks] EDA - Data Transformation: Suppose the relationship between X (independent variable) and Y (dependent variable) is represented by a power function $\mathbf{Y = 5X^3}$. Your task is to transform the data in such a way that we can fit a Linear Line using Linear Regression. What will be the approximated intercept term and slope of the line.

Intercept: ---------- **Slope: ----------**

D) [2 marks] Does K-mean algorithm always guarantee you the global optimum? Explain with Reasoning. Moreover how can we choose optimal number of clusters (K) in K-mean?

Name: _____ Reg #: _____ Section: _____

E) [1 point] What is the **overfitting problem** and what can be the possible cause for this problem? Write down all possible options for addressing the overfitting problem.

F) [1 point] What is difference between Covariance and Correlation?

G) [1 point] Why the parameter vector θ is perpendicular to the obtained Decision boundary? Prove it.

H) [2 marks] Suppose you train a logistic regression classifier in order to predict if the patient has cancer or not. Given the test data ($m_{test} = 100$), we already know that 10 patients have actually

cancer. On testing, our hypothesis predicted that 16 patients have cancer. Among the predicted ones, only 6 patients are those which actually have cancer (true positive).

Calculate the precision and recall for the case mentioned above.

I) [2 marks] Visualization:

i) In order to show the distribution of the values of a single variable X, which of the following visualization plots are useful.

a) Box Plot b) Histogram c) Scatter plot d) line chart

ii) In order to show the relationship of two variables, which of the following visualization plots are useful.

a) Box Plot b) Histogram c) Scatter plot d) line chart

Q4. [5 + 4 marks] Support Vector Machine (SVM):

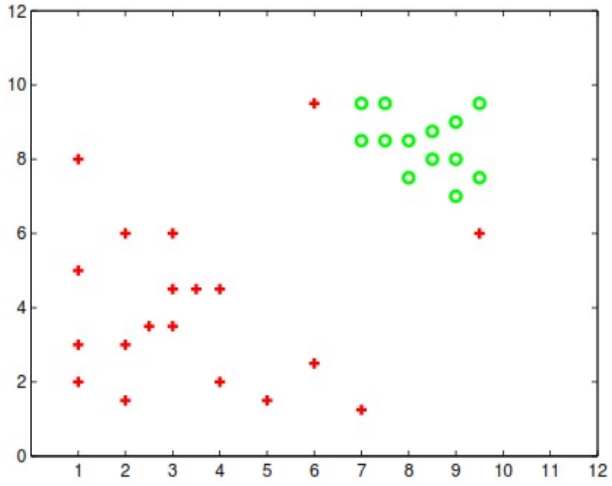
Part-I [5 marks]: The goal of this problem is to correctly classify test data points, given a training data set. For this problem, assume that we are training a SVM with a quadratic kernel– that is, our kernel function is a polynomial kernel of degree 2. You are given the data set presented in the Figure below. The slack penalty C will determine the location of the separating hyperplane. Please answer the following questions qualitatively. Give a one sentence answer for each and draw your solution in the appropriate part of the Figure at the end of the problem.

1. [1 point] Where would the decision boundary be for very large values of C (i.e., $C \rightarrow \infty$)? (remember that we are using an SVM with a quadratic kernel.) Draw on **figure (a) Part 1** below. Justify your answer.

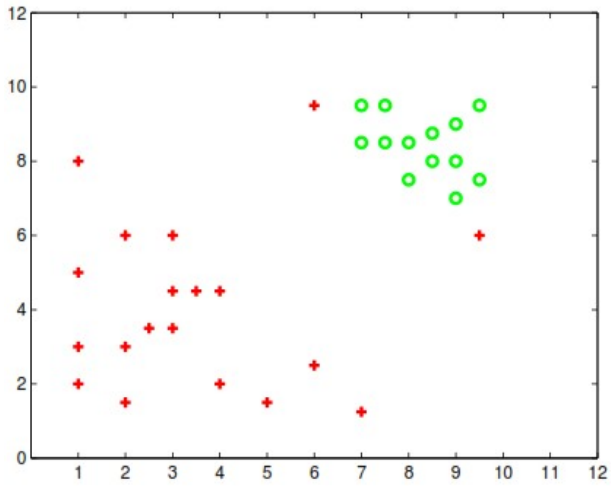
Name: _____ Reg #: _____ Section: _____

2. [1 point] For C ≈ 0, indicate in **figure (b) Part 2** below, where you would expect the decision boundary to be? Justify your answer.
3. [1 point] Which of the two cases above would you expect to work better in the classification task? Why?
   Ans :

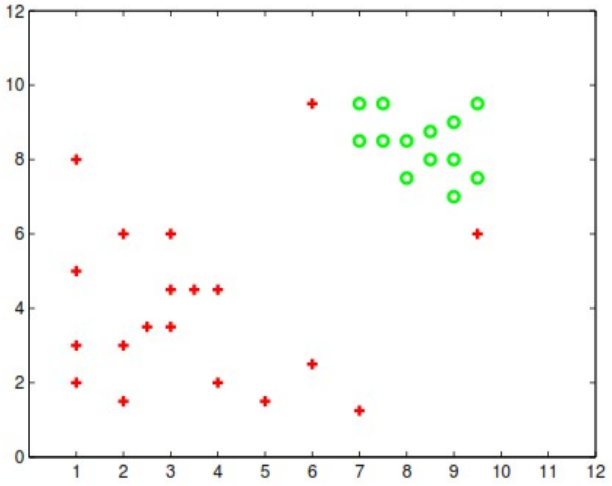
4. [1 point] Draw a new data point (on **figure (c) Part 3**) which will not change the decision boundary learned for very large values of C. Justify your answer.
5. [1 point] Draw a new data point (on **figure (d) Part 4**) which will significantly change the decision boundary learned for very large values of C. Justify your answer.



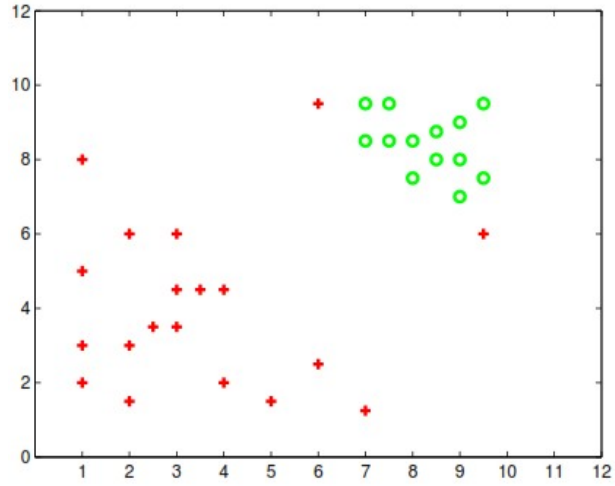
Part 1



(b) Part 2



(c) Part 3



(d) Part 4

Part-II [4 marks]: Suppose you want to train a classifier using SVM, the data set (with 2 features) is given in the following table. In the table, we can see that there are 3 training

Name: _____ Reg #: _____ Section: _____

examples (m=3). You don't need to normalize or scale the data. The data is not linearly separable so as a data scientist you want to classify it using SVM with Gaussian Kernel.

x_1	x_2	Y (class label: y=1 or y=0)
1	2	1
5	2	0
10	2	1

Answer the following questions.

a) Using Gaussian Kernel with **σ = 1**, compute the new features for all the examples.

b) After training using new computed features, we get the parameter values as $\Theta_0 = 0.5$, $\Theta_1 = 2$, $\Theta_2 = 0$, and $\Theta_3 = 1$. Given a test example p having **x_1 = 3** and **x_2 = 4**, show that if your trained model will classify it as 0 or 1.

Assume you predict y=1 when $\Theta_0 + \Theta_1 f_1 + \Theta_2 f_2 + \Theta_3 f_3 >= 0$.

The helpful formula is given as follows:

$f_1 = \text{similarity}(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$, where **σ = 1**

Name: _____ Reg #: _____ Section: _____