Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Introduction to Data science: Tools and Techniques**	**Course Code:**	**DS500**
Program:	**MS (Computer Science)**	**Semester:**	**Spring 2019**
Duration:	**180 Minutes**	**Total Marks:**	**48**
Paper Date:	**31-May-19**	**Weight**	**47 %**
Sections:	**All**	**Page(s):**	**13**
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	1	2	3	4	4	5	**Total**
Marks	/ 9	/ 12	/5	/8	/7	/7	**48**

Q1. [9 points] Short Questions:

1) [2 points] Given 4 data points in 2-d space, (2, 4), (1, 1), (5, 4) and (2, 2). Perform hierarchical agglomerative clustering using Average inter-cluster similarity and answer the following questions.

a) What is the 3-cluster lifetime for the data given above?

b) What do you think which is more sensitive to outliers: average-link or complete-link? Explain with an example.

2) [3.5 points] Considering the spam classification example, as a data scientist, you build a classifier for spam detection. Assume you have a test collection of 1000 emails. First 5 emails (email# 1 to 5) are labelled as spam (y = 1) and rest of the emails are not-spam (y = 0). On testing, you trained model detects 9 emails as spam, which are email# 1, 5, 10, 120, 300, 420, 500, 700, and 930.

(i) What is the accuracy of your trained model?

(ii) On the contracry a fresh trainee (who does not know machine learning) designed a simple function which always predicts y = 0, irrespective of the email features. It means the function takes a 'feature vector' as an argument and always classify it as not-spam (y = 0). What will be the accuracy of this system?

(iii) Do you think accuracy is a good measure in this case? If yes, please elaborate why? If not, which measure will be better? Compute the suggested measures using the data given above and show how your trained model is better.

3) [1.5 points] **Bias and Variance:** A set of data points is generated by the following process: $Y = w_0 + w_1X_1 + w_2X_2 + w_3X_3 + w_1X_4$, where X is a real-valued random variable. You use two models to fit the data:

Model 1: $Y = aX + b$

Model 2: $Y = w_0 + w_1X_1 + ... + w_9X_9$

i. Model 1, when compared to Model 2 using a fixed number of training examples, has a bias which is:

(a) Lower (b) Higher (c) The Same

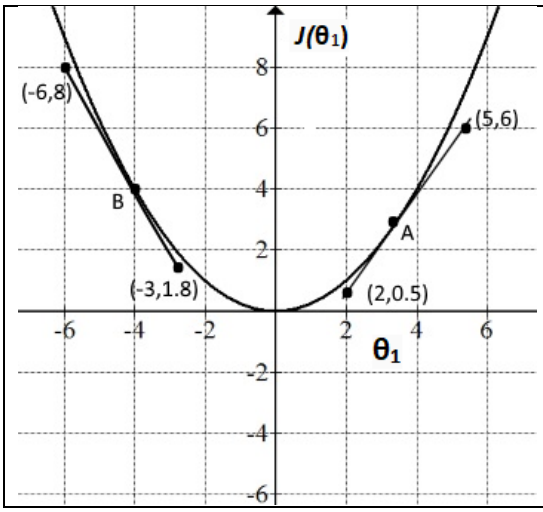
ii. Model 1, when compared to Model 2 using a fixed number of training examples, has a variance which is:

(a) Lower (b) Higher (c) The Same

iii. Given 10 training examples, which model is more likely to overfit the data?

(a) Model 1 (b) Model 2 (c) cannot say (d) none

4) [1 point] Assuming linear model for prediction $h_\Theta(x) = \theta_0 + \theta_1 x$, (where $\theta_0 = 0$), the cost function curve is shown in the Figure below. Suppose we initialize $\theta_1 = -4$, where the cost $J(\theta_1) = 4$ (as shown point B on the curve). What will be the updated value of θ_1 after single iteration of the Gradient Descent Algorithm (assume $\alpha = 1.5$)?



Solution:

5) [1 point] You are hired by a company as a data scientist and your first task is to build a machine learning system for ``email spam'' detection. After looking at related work and frquent words, you figured out the following 5 important features: (i) Buy, (ii) Discount, (iii) now, (iv) Deal, (v) cash. Construct the **feature vector** for the following email example.

From: sales@cheapsales.com
To: xyz@nu.edu.pk
Subject: Buy Now!

```
Deal of the week! Buy now!
Rolex w4tchs - $100
Med1cine (any kind) - $50
Also low cost M0rgages
available.
```

Solution:

Name: _____ Reg #: _____ Section: _____

Q2. [12 points] **Clearly circle the correct option/s and explain your choice with reasoning where required.**

i). A Data Scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result.

The models should be evaluated based on the following criteria:

1) Must have a recall rate of at least 80%

2) Must have a false positive rate of 10% or less

3) Must minimize business costs

After creating each binary classification model, the Data Scientist generates the corresponding confusion matrix. Which confusion matrix represents the model that satisfies the requirements?

A) TN = 91, FP = 9
FN = 22, TP = 78

B) TN = 99, FP = 1
FN = 21, TP = 79

C) TN = 96, FP = 4
FN = 10, TP = 90

D) TN = 98, FP = 2
FN = 18, TP = 82

ii). A regression model in which more than one independent variable is used to predict the dependent variable is called

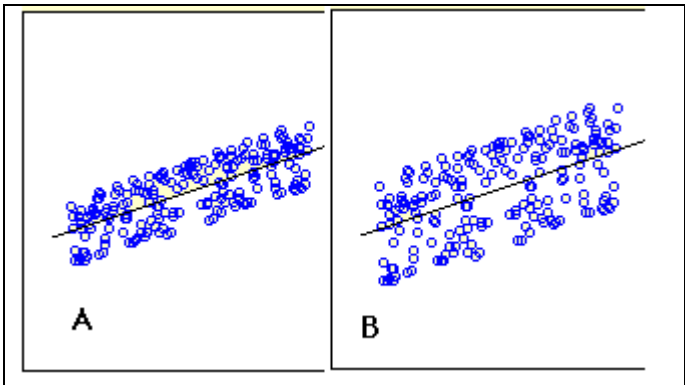
(A) an independent model (B) a simple linear regression model

(C) a multiple regression model (D) none of the choices

iii). Which of the following statement is true about sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note: Scale is same in both graphs for both axis. X axis is independent variable and Y-axis is dependent variable.



A) A has higher sum of residuals than B

B) A has lower sum of residual than B

C) Both have same sum of residuals

D) None of these

iv). What is wrong with the elbow method for selecting number of clusters?

A). It is hard to find the elbow when clusters are well separated

B). It is hard to find the elbow when clusters are poorly separated

C). It usually picks too many clusters D). It usually picks too few clusters

v). Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix}5\\5\end{bmatrix}$, $\mu_2 = \begin{bmatrix}-4\\3\end{bmatrix}$, $\mu 3 = \begin{bmatrix}3\\4\end{bmatrix}$. Furthermore, we have training example $\mathbf{x^{(i)}} = \begin{bmatrix}3\\-4\end{bmatrix}$. After a cluster assignment step, what will $\mathbf{C^{(i)}}$ be?

(a) $\mathbf{C^{(i)}} = 2$ (b) $\mathbf{C^{(i)}} = 1$ (c) $\mathbf{C^{(i)}}$ not assigned (d) $\mathbf{C^{(i)}} = 3$

vi). We can use multi-objective Genetic Algorithm for clustering the different types of news. What do you think which will be the relevant objective functions:

(a) Minimize the both intra-cluster distance and inter-cluster distance
(b) Maximize the both intra-cluster distance and inter-cluster distance
(c) Maximize the intra-cluster distance and minimize the inter-cluster distance
(d) Minimize the intra-cluster distance and maximize the inter-cluster distance
Note: Here intra-cluster means "within the same cluster"

vii). A Data scientist is preparing a data frame for a supervised learning task. He/She notices the target label classes are highly imbalanced and multiple feature columns contain missing values. The proportion of missing values across the entire data frame is less than 5%. What should the data scientist do to minimize bias due to missing values?

A) Replace each missing value by the mean or median across non-missing values in same row.

B) Delete observations that contain missing values because these represent less than 5% of the data.

C) Replace each missing value by the mean or median across non-missing values in the same column.

D) For each feature, approximate the missing values using supervised learning based on other features

viii). In Linear Regression problems, when you fit a line (model) on the training data, your aim is to minimize the error between prediction and actual values. Statisticians have developed different **error metrics** to judge the quality of a model, and two of them are:

(i) Mean Squared Error (**MSE**): $\frac{1}{m}\sum_{i=1}^{m}(h(x^i) - y^i)^2$, (ii) Mean Absolute Error (**MAE**): $\frac{1}{m}\sum_{i=1}^{m}|h(x^i) - y^i|$

a) MAE is less robust (more sensitive) to outliers than MSE.

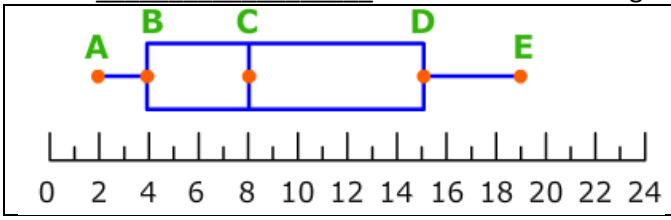
b) MSE is less robust (more sensitive) to outliers than MAE.

c) MSE can have more than one solutions (can fit more than one lines with optimal cost).

d) MAE can have more than one solutions (can fit more than one lines with optimal cost).

ix). What does the point labeled C represent on the box plot?

Name: _____ Reg #: _____ Section: _____



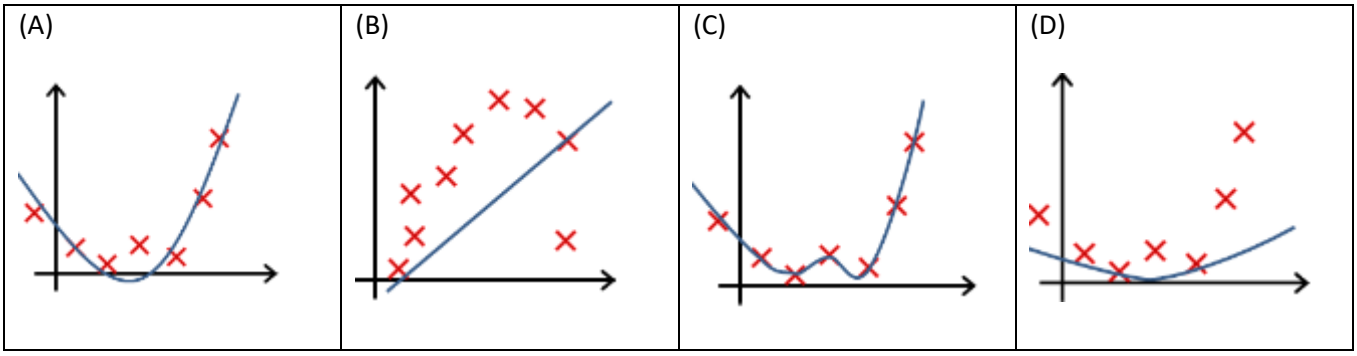
a) Mean b) Mode

c) Median d) Average

x). A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set. Explain your choice with reason. (a) True (b) False

Reason:

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



xii). Wrapper-based and Filter-based methods are two well-known approaches for

(a) Clustering (b) Feature selection

(c) Classification (c) Dimensionality reduction

Q3. [5 points] Regression and Regularization

1) [2.5 points] Consider the following dataset D in the one-dimensional space.
Here each row is one training example. Given the hypothesis below, we optimize using mean squared error cost function.

x	y
0	-1
1	2
1	0

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

a) Find the optimal θ_0, and θ_1 given the aforementioned dataset D and justify your answer.

b) What is the minimum number of training examples that are required to obtain a unique solution using the using mean squared error cost function?

2) [2.5 points] Regression with Regularization:

You are asked to use regularized linear regression to predict the target $Y \in \mathbb{R}$ from the eight-dimensional feature vector $X \in \mathbb{R}^8$. You define the model $Y = w^T X$ and then you recall from class the following three objective functions:

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \tag{5.1}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} w_j^2 \tag{5.2}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} |w_j| \tag{5.3}$$

a) [points: 1.5 pt] The following table contains the weights learned for all three objective functions (not in any particular order):

	Column A	Column B	Column C
w_1	0.60	0.38	0.50
w_2	0.30	0.23	0.20
w_3	-0.10	-0.02	0.00
w_4	0.20	0.15	0.09
w_5	0.30	0.21	0.00
w_6	0.20	0.03	0.00
w_7	0.02	0.04	0.00
w_8	0.26	0.12	0.05

Beside each objective write the appropriate column label (A, B, or C):

Objective 5.1: ----------

Objective 5.2: ----------

Objective 5.3: ----------

b) [Points: 0.5 pts] For large values of λ in objective 5.2 the bias would:

(a) increase (b) decrease (c) remain unaffected

c) [Points: 0.5 pts] For large values of λ in objective 5.3 the variance would:

(a) increase (b) decrease (c) remain unaffected

Name: _____ Reg #: _____ Section: _____

Q4. [8 marks] Support Vector Machine (SVM):

Part-I [2 marks]:

Soft-Margin Linear SVM. Given the following dataset in 1-d space (Figure 1), which consists of 4 positive data points $\{0, 1, 2, 3\}$ and 3 negative data points $\{-3, -2, -1\}$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft-margin linear SVM can be formalized as the following constrained quadratic optimization problem. In this formulation, C is the regularization parameter, which balances the size of margin (i.e., smaller $w^t w$) vs. the violation of the margin (i.e., smaller $\sum_{i=1}^{m} \epsilon_i$).

$$\text{argmin}_{\{w,b\}} \ \frac{1}{2} w^t w + C \sum_{i=1}^{m} \epsilon_i$$

$$\text{Subject to : } y_i(w^t x_i + b) \geq 1 - \epsilon_i$$

$$\epsilon_i \geq 0 \ \forall i$$

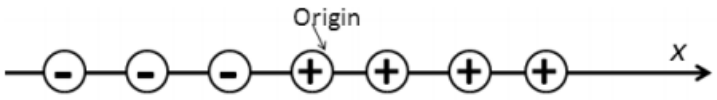


Figure 1: Dataset

a) if C = 0, how many support vectors do we have? (Support your answer with the reasoning.)

b) if C → ∞, how many support vectors do we have? (Support your answer with the reasoning.)

Part-II [2.5 marks]: Suppose you want to train a binary classifier using SVM, the data set (with 2 features) is given in the following table. Answer the following 2 parts.

x_1	x_2	Y (class label: y=1 or y=0)
1.5	-2	1
-1.5	2	0

(i) Calculate the P (Projection length) of the two training examples given in the table. Suppose the learned parameter values are $\Theta_0 = 0$, $\Theta_1 = 2$, and $\Theta_2 = 0$. Moreover draw the vectors, projection of examples on the parameter vector, and the corresponding decision boundary.

(ii) Calculate the P (Projection length) of the two training examples given in the table. Suppose the learned parameter values are $\Theta_0 = 0$, $\Theta_1 = 2$, and $\Theta_2 = 2$. Moreover draw the vectors, projection of examples on the parameter vector, and the corresponding decision boundary.

(iii) Which decision boundary will be calculated by SVM and why?

Part-III [3.5 marks]: Suppose you want to train a classifier using SVM, the data set (with 1 feature) is given in the following table. In the table, we can see that there are 4 training examples (m=4). 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. Given the two landmarks L1 = 2, and L2=8.

x_1	Y (class label: y=1 or y=0)
2	0
8	1
4	1
10	0

Answer the following questions.

a) Using Gaussian Kernel with **σ = 2**, compute the new features for all the examples.

b) After training using new computed features, we get the parameter values as $\Theta_0 = 1.5$, $\Theta_1 = 1$, and $\Theta_2 = 2$. Given a test example p having **x_1 = 9**, show that if your trained model will classify it as 0 or 1.

Assume you predict y=1 when $\Theta^T f >= 0$.

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 = 2**

Solution:

Q5. [7 marks] **Principal Component Analysis (PCA):**

Given 3 data points in 2-d space, (1, 1), (2, 2) and (3, 3),

(a) (2 pt) what is the first principle component? (Find unit vector)

(b) (1 pt) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data? Also write the projected data.

(c) (1 pt) For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error?

Name: _____ Reg #: _____ Section: _____

(d) (2 pts) If we project the given 3 data examples on a vector U = $\begin{bmatrix}1\\0\end{bmatrix}$, compute the projected data and also calculate the variance of the projected data. Do you think there will be any information loss?

e) [1 point] In order to find the Eigen vectors of a matrix A, why do we solve $|A - \lambda I| = 0$? Why determinant of this matrix is set to 0.

Q6: [7 points] Exploratory Data Analysis: Data Transformation

Suppose the relationship between X (independent variable) and Y (dependent variable) is represented by a power function **$Y = 2X^3$**. The data for X is given in the table below. Suppose for the given data we fit a line $h(x) = \Theta_0 + \Theta_1 X$ such that $\Theta_0 = 0$, and $\Theta_1 = 20$.

Example #	1	2	3	4
X	**1**	**2**	**3**	**4**

A) [3 points] Draw the standardized residual plot and discuss if there is any room for improvement in the model? Moreover compute coefficient of determination $R^2 = 1 - (SE_{line}/SE_{\bar{Y}})$, (Here SE means squared error, so you have to take square for both terms).

B) [2 points] If there is any room for improvement then transform the data in such a way that we can fit a Linear Line using Linear Regression. Plot the transformed data on a graph and fit a linear line.

C) [1 point] Find out the approximated intercept term and slope of the line (on transformed data).

D) [1 point] Compute the required measure and tell if the transformation is successful (better than the previous model)?

Good luck ☺