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:	**60 Minutes**	**Total Marks:**	**22**
Paper Date:	**08-April-17**	**Weight**	**15 % approx**
Section:	**ALL**	**Page(s):**	**5**
Exam:	**Mid-II**		

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	5	Total
Marks	/ 11	/ 4	/ 2	/ 2	/3	/ 22

Q1. [2+3+2+2+2 Marks] Short Questions:

A) [2 points] Why do we use F measure instead of simply taking average of precision and recall? What problem will arise if we use average of precision and recall for evaluation, illustrate with an example.

Solution:

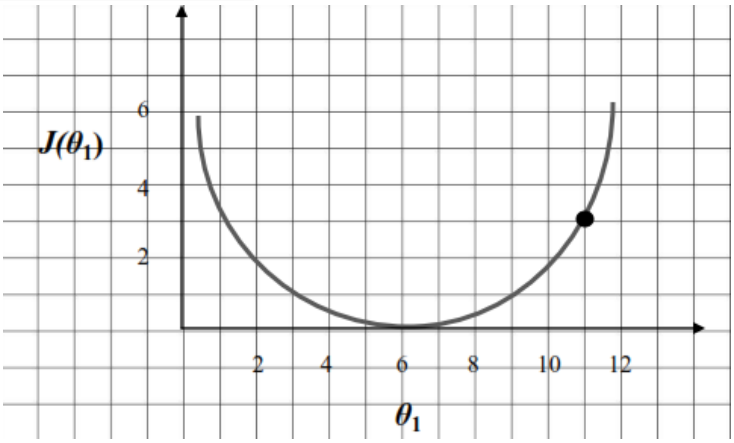
B) [3 points] Suppose you train a logistic regression classifier in order to predict if the aircraft engine is faulty or not. Our model predicts 1 if $h(x) > 0.5$. Given the test data ($m_{test} = 150$), we already know that 10% of the aircrafts have actually fault. On testing, our hypothesis predicted that 22% of the aircrafts have fault. Only 6% of the predicted ones are those, which actually have fault.

a. Create a table with actual number of true positive, true negative, false positive and false negative examples. Moreover, Calculate the F score.

Solution:

b. What should we do if we don't want to miss-classify too many cases of the faulty engines (engines which are actually faulty)?

C) 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 $\theta_1 = 11$, where the cost $J(\theta_1) = 3$ (as shown the dot on the curve). What will be the updated value of θ_1 after single iteration of the Gradient Descent Algorithm (assume $\alpha = 2$)?



D) For a specific choice of model, as the number of training points goes to infinity, describe the changes to the bias and variance exhibited by a model trained on the data.

E) What down at least one advantage (other than simplicity) and one disadvantage of Mean Imputation?

Name: _____ Reg #: _____ Section: _____

Q2. [2+2 marks]

a) Suppose a company X has m programmers in pool and you as a project manager want to assign teams of programmers to N projects. Assuming N=3 and m=4, you assign teams of 3, 1, 2 programmers respectively from a pool of m programmers. Moreover, one programmer can be assigned to multiple projects.

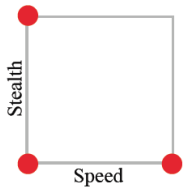
i) What is the total number of Factors and what will be a Design point for this problem?

ii) What will be the total number of combinations?

b) If we apply "One-at-a -Time Variation" Design to "Capture-the-flag" example studied in the class, we get the following experiments.

i) Can you guess which factor is contributing to success? Explain your selected choice in one or two sentences.

Speed	Stealth	Success?
Low	Low	No
High	Low	No
Low	High	No



ii) In which scenarios, "One-at-a-time Variation" Design will not work?

Name: _____ Reg #: _____ Section: _____

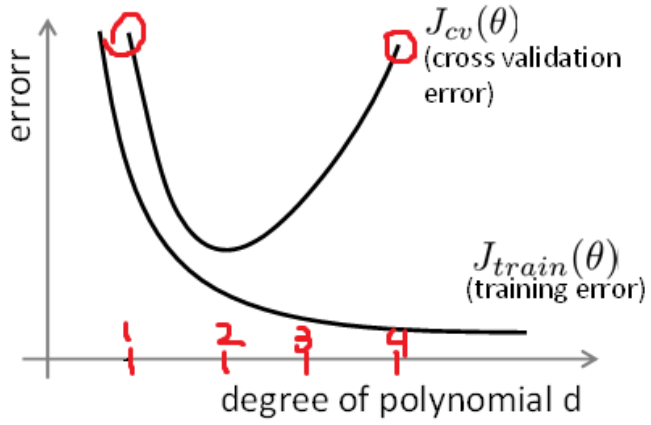
Q3. [2 marks] **Diagnosing bias vs. variance: Answer the following questions:**

(1). If $J_{cv}(\theta)$ and $J_{train}(\theta)$ are high such that ($J_{cv}(\theta) \approx J_{train}(\theta)$). Is it a bias problem or variance problem?

(2). If $J_{train}(\theta)$ is low and $J_{cv}(\theta)$ >> $J_{train}(\theta)$. Is it a bias problem or variance problem?

(3). For what value of d (degree of polynomial), the problem is underfit?

(4). For what value of d (degree of polynomial), the problem is overfit?



Q4. [2 marks]: Convert the following XML into JSON

```
<Semester8>
 <student>
  <name> ABC </name>
  <reg_no> 3112 </reg_no>
  <section> A </section>
 </student>

 <student>
  <name> xyz </name>
  <reg_no> 3244 </reg_no>
  <section> B </section>
 </student>
</Semester8>
```

Name: _____ Reg #: _____ Section: _____

Q5. [3 marks] **Clearly circle the correct options.**

i) 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).

ii) 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.

iii) 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.

(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.