University of Central Punjab

Faculty of Information Technology

Midterm Examination

Course Title : Programming for Big Data

Course Instructor: Saeed Iqbal Khattak	Semester: SPRING 2021
Course Code: CSDS4423	Mid Term: 20%
Duration: 90 Minutes	Time: 09:00 AM to 10:30 AM
Date: May 18, 2021	Program: BSCS
Room# :	Review Date: TBA

Instructions

1. **Make sure your mobile phone is switched off and place it in/on the bag/table.** Students are **allowed** to use INTERNET, smartwatches, books, computer, laptops, notes, and cheat sheets.
2. You are strongly discouraged from plagiarism (copy & paste) and do not copy from fellows, otherwise you will be graded "F".
3. Total paper time : **90 Minutes**.
4. Try to finish your exam within prescribed time.
5. Be at the exam venue at least 15 minutes before the scheduled start time.
6. Kindly submit a **zip folder (Jupyter Notebook containing short questions and MS Excel containing a long question)** on CMS or MS Team before the deadline.

PART II – Short Questions

1. What is correlation between variables or features of a dataset? Why we need to find correlation to extract the best predictor?

2. In the following table each row represents one observation, or the data about one employee (either Ann, Rob, Tom, or Ivy). Each column shows one property or feature (name, experience, or salary) for all the employees.

Name	Years of Experience	Annual Salary
Ann	30	120,000
Rob	21	105,000
Tom	19	90,000
Ivy	10	82,000

   If you analyze any two features of a dataset, then you'll find some type of correlation between those two features. Explain what type of correlation is expressed in the following three graphs?

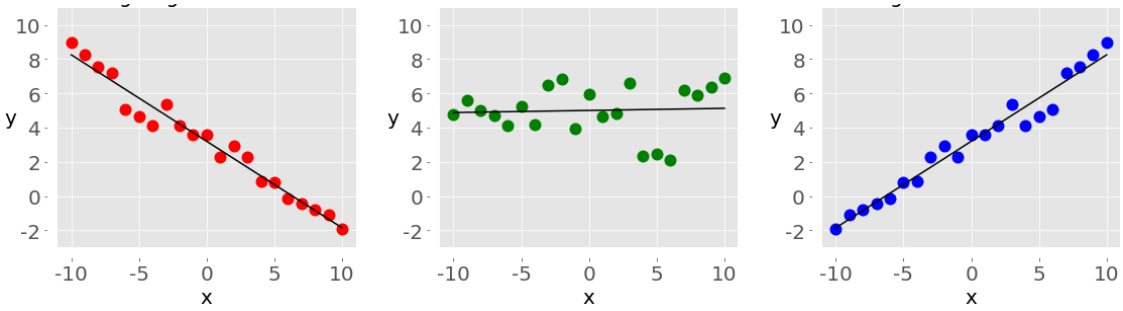

Figure 1: Different forms of Correlations

3. What is the difference between correlation and linear regression?

PART III – Long Question

A botanist has collected data on **three** flowers and forgot which data belonged to which flower. Now to help the botanist you must use your skills to group the data into **two** parts based on the similarity of the features.

There are two features namely **petal length** and **sepal length**. We need to group the data into **2 groups** based on the similarity of the features and we will do that by finding the distance between the features.

What you are going to do is that you will take 2 random data points from the data given and find the distance of all the points from these 2 referenced data points. These reference data points will be used to create groups that is each point is a group.

After calculating the distance, each point will be assigned to the group with which it has the minimum distance. Now after all the data points have been assigned their groups you need to calculate 2 new reference points. You will do that by taking the average of all the points (their features) in a specific group. For example, we have 2 features in a data point so you will calculate 2 averages (column wise) per group which will create a new data point.

You will calculate the average for all two groups and will have 2 new reference data points. You are going to repeat this 3 times or until the new data points calculated have very small difference than those calculated at the previous iteration up to two decimal points.

To calculate the distance, you can use the Manhattan distance formula 1 given below for 2 features:

$$|a-x|+|b-y| \tag{1}$$

Petal Length	Sepal Length
5.1	1.4
4.9	1.4
4.7	1.3
4.6	1.5
5	1.4
6	5.1
5.4	4.5
6	4.5
6.7	4.7
6.3	4.4
7.9	6.4
6.4	5.6
6.3	5.1
6.1	5.6
7.7	6.1