German University in Cairo
Department of Digital Media Engineering and Technology
CSEN 1083: Data Mining

CSEN 1083: Data Mining
Spring 2019
Quiz #1

Name: ....

App #: ....

Group #: ....

Instructions: Read carefully before proceeding.

1) The duration of this quiz is **20 minutes**.
2) Write your name, application number, and group number in the provided space above.
3) No books, notes or other aids are permitted for this quiz.
4) When you are told that time is up, please stop working on the test.
5) Calculators are allowed to the quiz.

Good Luck!

Question 1: **(4 marks)**

For each of the following applications, **choose** what type of data mining task it represents (Classification, Regression, Clustering, Association Rule Discovery, or Anomaly Detection). Each task could be chosen more than once. **Justify** your answer.

1) In a call center, analyzing the relationship between wait times of callers and number of complaints received

2) Identifying strange patterns in network traffic that could signal a hack

3) When uploading an image on Facebook, it suggests tagging people whose faces appear in the image

4) Identifying websites that talk about the same topic (regardless of the topic type)

Answer:

1) Regression. We are trying to find the relationship between two variables.
2) Anomaly Detection. Strange patterns corresponding to hacking would represent anomalies as they don't happen that frequently
3) Classification. Assigning images to specific people (discrete classes)
4) Clustering. We don't have labels. Only assigning websites (points) that are similar to the same group.

Question 2: **(4 Marks)**

A box contains 3 red balls and 6 blue balls. A second box contains 5 red balls and an unknown number of blue balls. A single ball is drawn from each box. The probability that both balls are of the same color is 19/36. **Calculate** the number of blue balls in the second box.

Answer:

Pr(Both balls of same color) = Pr(Ball1 = red, Ball2 = red) + Pr(Ball1 = blue, Ball2 = blue)

Given the independence between the two picked balls

Pr(Both balls of same color) = Pr(Ball1 = red) Pr(Ball2 = red) + Pr(Ball1 = blue) Pr(Ball2 = blue)

Let x be the number of blue balls in the second box.

Pr(Both balls of same color) $= (3/9)*(5/(5+x)) + (6/9)*(x/(5+x)) = 19/36$

$x = 7$

Question 3: **(2 Marks)**

For each of the following datasets, **choose** the best dataset type to represent the data (Record Data, Graph-based Data, Ordered Data). Each dataset type could be chosen more than once. **Justify** your answer.

1) Heart signals for patients

2) Data about bank clients including their addresses, age and savings

3) Images of red blood cells for cancer detection

Answer:

1) Ordered Data. Signals represent a time-series which is one form of ordered data.
2) Record Data. Dataset with three attributes: address, age and savings.
3) Ordered Data. Images represent spatial data.