CSE 40647/60647: Data Science **Fall 2021**

Homework Written Assignment 1: Data Processing

Handed Out: August 24, 2021 *Due: September 12, 2021 11:55pm*

Save and submit your solution file as *NETID-hw1-written.pdf*.

1 Concepts: Single-choice questions (10 points)

1.1 [2 points] Given a dataset of FIFA player profiles (e.g., age, country, position, height, weight), suppose a hypothesis is "generally goalkeeper is taller than other positions," which statistical description should be calculated?
a) Average height of players per country
b) Maximum height of players per country
c) Average height of players per position
d) Maximum height of players per position

1.2 [2 points] Suppose we have a dataset of 1,000 ND employees' personal wealth. Suppose Jeff Bezos (CEO of Amazon) joins ND's faculty. So, we have this new data point, which statistical description will **NOT** change significantly?
a) Mean value of personal wealth
b) Median value of personal wealth
c) Variance of personal wealth
d) Standard deviation of personal wealth

1.3 [2 points] Given a large dataset, how can we have the first impression of it? Take a look at the top 20 rows, the bottom 20 rows, and a 20 random rows. Here what are these 60 rows?
a) A subset of data objects and all of their attribute values
b) A subset of attributes and all the associated data objects

1.4 [2 points] The more features (dimensions) we have about a particular data point, the better performance a machine learning model will deliver.
a) True
b) False

1.5 [2 points] Given social network data as a "user-to-user" friendship graph, the data object is a user. We have two object-feature data matrices. The features in the first matrix (data I) are node attributes, i.e., user's profiling information such as age, gender, and education. The features in the second matrix (data II) are the set of users in the network and the feature values are binary (1, if the "object user" has a relation

with the "feature user"; 0, otherwise). Which of the following is correct?
a) Data II has more dimensions and higher density than data I
b) Data II has more dimensions and lower density than data I
c) Data II has fewer dimensions and higher density than data I
d) Data II has fewer dimensions and lower density than data I

2 Correlation Analysis (10 points)

[10 points] Suppose two stocks X_1 and X_2 have the following value of price in 5 days:

$$(X_1, X_2) : (\$2, \$5), (\$3, \$8), (\$5, \$10), (\$4, \$11), (\$6, \$14)$$

Are their prices rising/falling together or in different trends? Calculate the covariance without using any computing tools.

3 Regression Analysis (10 points)

[10 points] Suppose we have n data samples. The i-th sample ($i = 1, \dots, n$) has k numerical features ($\{x_{i,j}\}_{j=1}^{k}$) and 1 numerical label (y_i). A standard linear regression model M makes a coefficient of determination R^2 from the n samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Now we add one more feature, that is the average of the original k features:

$$x_{i,k+1} = \frac{\sum_{j=1}^{k} x_{i,j}}{k},$$

and build a new linear regression model M'. Will the new coefficient of determination R'^2 be bigger than, or smaller than, or equal to R^2? Please mathematically prove your answer.