CSE 40647/60647: Data Science **Fall 2019**

Homework 1: Written Assignment

Handed Out: August 29, 2019 *Due: September 16, 2019 11:55pm*

Save your homework submission as *NETID-hw1-written.pdf*.

1 Data Description (20 points)

Suppose the population size is $N = 1,000,000$. We sample $n = 9$ examples x_i $(1 \leq i \leq n)$ from the data. Suppose the mean value of the sample data is $\mu = 10$ and the variance is $v = 18$. Now we sample one more example $x_{n+1} = 20$ from the data. So the sample size is $n + 1 = 10$. What is the new mean value μ' and the new variance v'?

Note that the result will be the same no matter what x_i $(1 \leq i \leq n)$ are. You are expected to derive functions of calculating $\mu' = f(\mu, n, x_{n+1})$ and $v' = g(v, \mu, n, x_{n+1})$. Making fake numbers to find the answer, not mathmatically proving the answer, will have only half of the points.

2 Data Integration - Correlation Analysis (10 points)

Suppose two stocks X_1 and X_2 have the following values in one week:
$(2, 5), (3, 8), (5, 10), (4, 11), (6, 14)$
Are their prices rising/falling together or in different trends?

3 Data Reduction - Regression (20 points)

Suppose we have K numerical features and 1 numerical label. A standard linear regression model M makes a coefficient of determination R. Now we add one more feature, that is the average of the original K features, and build a new linear regression model M'. Will the new coefficeint of determination R' be bigger than, or small than, or equal to R? Please mathmatically prove your answer.

CSE 40647/60647: Data Science **Fall 2019**

Homework 1: Programming Assignment

Handed Out: August 29, 2019 *Due: September 16, 2019 11:55pm*

Save your homework submission as *NETID-hw1-programming.zip*. The zip file has one pdf file *NETID-hw1-programming.pdf* and multiple code files.

You are allowed to use any programming language (Python recommended; R, C++, Java, etc.), however, the solutions will be in **Python**. You are allowed to use any public package (including Numpy and Scikit-learn) and any other kind of tools (Excel).

A Film Dataset

File name: data-film.csv

Introduction: Suppose we have 1,000,000 films. We sample 150 films, so our dataset has 150 films (as the data objects). Each film has an ID ("f\$DIGIT"; \$DIGIT $\in \{1,\ldots,150\}$) and a genre (or called category/class) from {"ACTION", "ROMANCE", "COMEDY"}. Each film has four attributes. Each attribute is the average rating of the film given by a specific website. The attribute name is "AVGRATING_WEBSITE_\$DIGIT" (\$DIGIT $\in \{1,2,3,4\}$). The attribute values are numerical. Note that the rating scales can be different.

The first line is the header of attribute and label names. The following lines (rows) are object ID, attribute values, and label values of the data objects (films). The columns are separated by comma.

Example: The third line is "f2,4.9,3,1.4,0.2,ACTION": The film "f2" is an ACTION film. It is graded as 4.9 on Website 1, 3.0 on Website 2, 1.4 on Website 3, and 0.2 on Website 4.

Data Description (10 points)

The object-feature data matrix, which is denoted as **D**, has $m = 150$ objects and $n = 4$ features (and therefore, it has 600 values). Please use *Z-score normalization* to normalize the data matrix by each feature. The normalized data matrix is denoted as **A**. What are the maximum/minimum Z score for each feature?

Output: Write down the maximum/minimum Z score in the pdf. Save your code as NETID-hw1-1.py.

Data Visualization (30 points)

With the original data matrix **D**:

(1) generate *boxplot* for the first attribute "AVGRATING_WEBSITE_1;"

Output: Show the figure in the pdf. Save your code as NETID-hw1-2-1.py.

(2) generate *histogram* for the third attribute "AVGRATING_WEBSITE_3". Set the number of histograms/bins as 10;

Output: Show the figure in the pdf. Save your code as NETID-hw1-2-2.py.

(3) generate *bar chart* where the X-axis is the genre and the Y-axis is the mean value of average ratings of films given in the first attribute "AVGRATING_WEBSITE_1;"

Output: Show the figure in the pdf. Save your code as NETID-hw1-2-3.py.

(4) generate 2-dimensional *scatter plot* using attributes "AVGRATING_WEBSITE_1" and "AVGRATING_WEBSITE_3", where the marker types and colors should be different for different genres (node labels);

Output: Show the figure in the pdf. Save your code as NETID-hw1-2-4.py.

(5) calculate the *KL divergence* between attributes "AVGRATING_WEBSITE_1" and "AVGRATING_WEBSITE_3." Hint: For each attribute, you need to generate a probability distribution. You can determine bins on the attribute value by yourself. For each bin, you have the frequency (which is the number of films that have an attribute value in the bin) can calculate the probability. For example, if the total number of films is 150, if there are 15 films whose value of attribute "AVGRATING_WEBSITE_1" is in the bin $[1.5, 2)$ (between 1.5 and 2), the probability is 0.1. Then you have two probability distributions: one for each attribute. Just call python libraries to calculate the KL divergence.

Output: Write down the KL divergence score in the pdf. Save your code as NETID-hw1-2-5.py.

Data Cleaning and Integration (10 points)

With the original data matrix $\mathbf{D}$:
(1) calculate *correlation coefficients* $\rho_{i,j}$, which is ρ("AVGRATING_WEBSITE_i", "AVGRATING_WEBSITE_j"), between every pair of the attributes using covariance analysis.

Output: Write down the correlation coefficients in the pdf. Save your code as NETID-hw1-3-1.py.

With the Z-score normalized data matrix $\mathbf{A}$:

(2) (the same as (1)) calculate *correlation coefficients* between every pair of the attributes using covariance analysis.

Output: Write down the correlation coefficients in the pdf. Save your code as NETID-hw1-3-2.py.

(3) Are the above results the same? And why?

Output: Write down your reasoning in the pdf.