University of Glasgow | School of Computing Science

CSC3005 Laboratory/Tutorial 1: Data Exploration and Data Quality

1. **Data exploration** is initial investigation of data to better understand the data specific characteristics. Two key reason for data exploration are the following:
    a) To guide the selection of the appropriate data preprocessing and data analysis technique.
    b) To leverage on humans' abilities to recognize data patterns.

    Quantitative measure in terms of summary statistics such as the mean and covariance will be used.

    The Iris flower plant sample data from UCL (T1Q1_iris.data) contains information on 150 Iris flowers, 50 each from one of three Iris plant species: Setosa, Versicolour, and Virginica. Each flower plant is characterized by five attributes/features:
    a. sepal length in centimeters
    b. sepal width in centimeters
    c. petal length in centimeters
    d. petal width in centimeters
    e. class (Iris Setosa, Iris Versicolour, Iris Virginica)

    Load the T1Q1_iris.data as a CSV data file into a Python Pandas DataFrame object and compute various summary multivariate statistics **with statistics library and without (write your own formulae, obtained from lecture note) statistics library** on the following for each quantitative attribute
    a. Mean length
    b. Standard deviation of the length
    c. Minimum length and Maximum length
    d. Covariance between pairs of attributes
    e. Correlation between pairs of attributes

Using Python matplotlib library, perform data visualization by plotting

    a. Histogram for the petal length attribute by discretizing it into 20 separate bins and counting the frequency for each bin.
    b. Boxplot to show the distribution of values for each attribute
    c. Scatterplot to visualize the joint distribution between pairs of attributes
    d. Parallel coordinates to display all data points simultaneously

Refer to https://pandas.pydata.org/pandas-docs/stable/reference/frame.html

2. Data Quality

Poor Data Quality will result poor effect on data analytics. We will explore three of these poor data qualities in terms of

a) Missing Data
b) Outlier Data
c) Duplicate data

Load clinical T1Q2_breast-cancer.data as CSV data file into a Python Pandas DataFrame object. It contains information on 699 cases of breast cancer cell characteristics. Each case is characterized by ten attributes/features and id number namely

a. Sample code number: id number
b. Clump thickness (1-10)
c. Uniformity of cell size (1-10)
d. Uniformity of cell shape (1-10)
e. Marginal adhesion (1-10)
f. Single Epithelial cell size (1-10)
g. Bare Nuclei (1-10)
h. Bland chromatin (1-10)
i. Normal Nucleoli (1-10)
j. Mitoses (1-10)
k. Class (2 for benign, 4 for malignant)

a) Missing Data

It is common to have missing data for attribute due to their information was not collected or some attributes are inapplicable to some data cases. In the breast cancer dataset (T1Q2_breast-cancer.data), the missing values are encoded as '?' in the dataset. Perform the following cleaning data task for missing data on the breast cancer dataset

i. Identity those attributes that have missing data and convert these missing data values in these attributes to NaNs.
ii. count the number of missing values in each of these attributes.

There are two approaches to deal on missing Data

Approach 1

i. Calculate the median value of that attribute that has missing data

ii. Replace those missing data in that attribute with the calculated median Comment why we choose median?

Approach 2

i. Drop those attributes that has missing data by using *dropna()* function.
ii. How many data cases have left after dropping the attributes that has missing data?

Which approach is better?

b) Outlier Data

Outlier data are those data with characteristics that are considerably different from the rest of the dataset. Perform the following cleaning data task for outlier data on the breast cancer dataset.

i. Ensure those missing data has been replaced by *NaNs* as done in part (a) as we are checking anomality based on integer (*int64*).
ii. Drop the class attribute in the data using *drop()* function as class attribute is qualitative and not quantitative
iii. Using *boxplot()* function to plot to visualize the outliers. You will observe five attributes will have abnormally high values in some of their dataset.
iv. Use Z score methodology for each attribute to remove such outlier
    a. Calculate z score of each data value of an attribute where $z = x - \bar{\mathbf{x}}/\mu$ where x is each data value for an attribute and $\bar{\mathbf{x}}$ is the mean value of an attributes. μ is the standard deviation value for an attribute
    b. Drop data that has $z > 3$ or $z < -3$. Why?
    c. How many data instances have left after remove outlier data?

c) Duplicate Data

Duplicate datasets are common especially those obtained by merging multiple data sources such as Facebook Big Data servers at various geographical locations around the globe may contain duplicates or near duplicate instances. The term deduplication is often used to refer to the process of dealing with duplicate data issues. Perform the following cleaning data task for duplicate data on the breast cancer dataset.

i. Use *duplicated()* function to determine number of duplicate data in the dataset
ii. Drop those dataset that are duplicate. How many data instances have left after remove outlier data?

3. Relationship cosine and correlation measures.
   a) What is the range of values that are possible for the cosine measure?
   b) If two objects have a cosine measure of 1, are they identical? Explain.
   c) What is the relationship of the cosine measure to correlation, if any?
   d) Figure 3(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?
   e) Figure 3(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?
   f) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L2 length of 1.
   g) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.

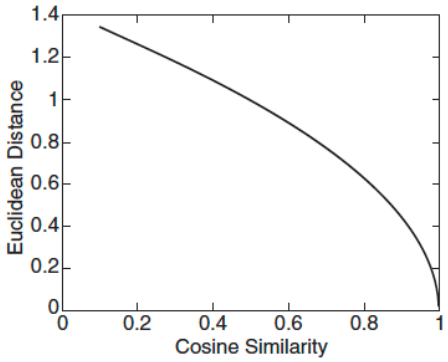

Fig 3(a) Relationship between Euclidean distance and Cosine Similarity

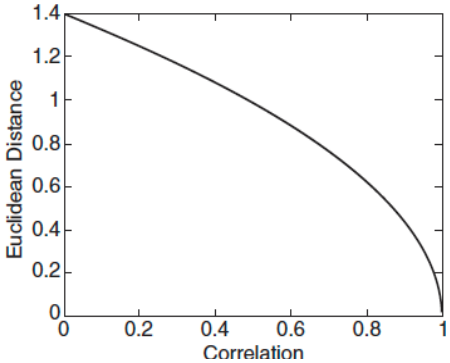

Fig 3(b) Relationship between Euclidean distance and Correlation