

Experiment No.5

Title: Applying similarity measures on the numeric datasets

Batch:A2

Roll No.:16010421063

Experiment No.: 5

Aim: Applying similarity measures on the numeric datasets

Resources needed: Any programming language, any data source (RDBMS/Excel/CSV)

Theory:

Similarity measures:

Similarity measures for numeric attributes include the *Euclidean*, *Manhattan*, and *Minkowski distances*.

The most popular distance measure is Euclidean distance (i.e., straight line or “as the crow flies”). Let $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as,

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}. \quad \dots \quad (1)$$

Another well-known measure is the Manhattan (or city block) distance, named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is defined as,

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|. \quad \dots \quad (2)$$

Both the Euclidean and the Manhattan distance satisfy the following mathematical properties:

Non-negativity: Distance is a non-negative number.

Identity of indiscernible: The distance of an object to itself is 0.

Minkowski distance is a generalization of the Euclidean and Manhattan distances. It is defined as,

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h} \dots\dots\dots(3)$$

Where h is a real number such that $h \geq 1$. It represents the Manhattan distance when $h = 1$ and Euclidean distance when $h = 2$.

When $h \rightarrow \infty$, its a “supremum” (L_{\max} norm, L_{∞} norm) distance.

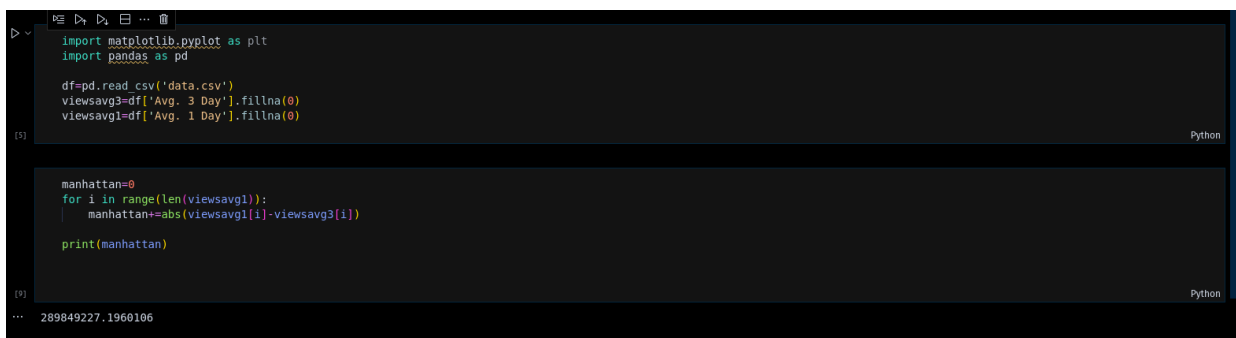
- This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Procedure / Approach /Algorithm / Activity Diagram:

1. Identify the suitable attributes to apply the numeric similarity measures and write python code to calculate Euclidean, Manhattan similarity measures on it.
-

Results: (Program printout with output / Document printout as per the format)



```

import matplotlib.pyplot as plt
import pandas as pd

df=pd.read_csv('data.csv')
viewsavg3=df[['Avg. 3 Day']].fillna(0)
viewsavg1=df[['Avg. 1 Day']].fillna(0)

manhattan=0
for i in range(len(viewsavg1)):
    manhattan+=abs(viewsavg1[i]-viewsavg3[i])

print(manhattan)

```

289849227.1960106

```

from math import sqrt

euclidian=0
for i in range(len(viewsavg3)):
    euclidian=(abs(viewsavg3[i]-viewsavg1[i]))**2
euclidian=sqrt(euclidian)
euclidian

[11]
... 64542.5

supremum=0

for i in range(len(viewsavg1)):
    supremum=max(supremum,abs(viewsavg1[i]-viewsavg3[i]))

supremum

[12]
... 6596001.0

```

Questions:

1. What are the different applications of Numeric similarity measure?

Euclidean distance measurement and principal component analysis methods are applied on such databases to identify the genes. In both methods, prediction algorithm is based on homology search approach. Digital Signal Processing technique along with statistical method is used for analysis of genes in both cases. Regression analysis: Manhattan distance is used in linear regression to find a straight line that fits a given set of points. Compressed sensing: In solving an underdetermined system of linear equations, the regularisation term for the parameter vector is expressed in terms of Manhattan distance.

2. What are the different applications of finding similarity between textual attributes?

Text similarity measures play an increasingly vital role in text related research and applications in several tasks such as text classification, information retrieval, topic tracking, document

clustering, questions generation, question answering, short answer scoring, machine translation, essay scoring, text summarization, topic detection and others.

Outcomes:

CO2: Comprehend descriptive and proximity measures of data.

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Successfully implemented Manhattan and Euclidian distance similarity measures on two columns of a dataset

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition
2. Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.