Experiment No.5

Title: Applying similarity measures on the numeric datasets

Batch: **Roll No.:** **Experiment No.: 5**

Aim: Applying similarity measures on the numeric datasets and textual 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 $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by *p* numeric attributes. The Euclidean distance between objects *i* and *j* is defined as,

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

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,

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

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,

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

Where *h* is a real number such that *h* >= 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

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

Procedure / Approach /Algorithm / Activity Diagram:

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)

Questions:

1. What is distance in Data Science and what are its importance?
2. What are the different applications of Numeric similarity measure?
3. Why to use Mahalonobis distance if Euclidian distances available? Give suitable example with justification.

Outcomes:

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

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.