Worksheet 3: Distance and Similarity

1. Worksheet Questions

a. **In the minkowski distance, describe what the parameters p and d are.**

- In the Minkowski distance, the parameter p determines the order of the norm used to compute the distance. It controls how the differences in each dimension are aggregated. The parameter d represents the number of dimensions in the space where the distance is being measured.

- p **(order parameter)**: This controls how the individual differences in each dimension are aggregated. Different values of p give different types of distances:
  - $p = 1$ → Manhattan distance (sum of absolute differences)
  - $p = 2$ → Euclidean distance (root of sum of squared differences)
  - $p = \infty$ → Chebyshev distance (maximum absolute difference)

  As p increases, the largest difference in any single dimension has more influence on the final distance.
- d **(dimensionality)**: This represents the number of dimensions in the space where the distance is being measured. For example, in a 2D space, $d = 2$, meaning the points A and B each have two coordinates (e.g., $A = (x_1, y_1)$, $B = (x_2, y_2)$). In higher dimensions, Minkowski distance extends similarly but sums over all dimensions.

b. **In your own words describe the difference between the Euclidean distance and the Manhattan distance.**

- The Euclidean distance measures the straight-line (or shortest) distance between two points, calculated using the Pythagorean theorem. The Manhattan distance, on the other hand, measures the distance by summing the absolute differences along each dimension, resembling movement along a grid (like a taxi driving on city streets).

c. **Consider A = (0, 0) and B = (1, 1). When:**

   a. p = 1, d(A, B) = 2 = = | 1−0 | + | 1−0 | =2 (Manhattan distance)

b. p = 2, d(A, B) = 2 = $\sqrt{(1-0)^2-(1-0)^2} = \sqrt{2}$ ≈1.41 (Euclidean distance)

c. p = 3, d(A, B) = $2^{1/3}$=1.26

d. p = 4, d(A, B) = $2^{1/4}$=1.19

d. **Describe what you think distance would look like when p is very large.**

- As p approaches infinity, the Minkowski distance converges to the Chebyshev distance, which considers only the maximum absolute difference in any one dimension. In other words, the distance between two points will be dominated by the largest single-coordinate difference.

e. **Is the minkowski distance still a distance function when p < 1? Expain why / why not.**

- When p<1, the Minkowski distance does not satisfy the triangle inequality, a crucial property of a metric. This is because the sum of distances between two points may become smaller than expected, leading to an unrealistic measure of separation. As a result, it is not a valid distance function in a strict mathematical sense.

f. **When would you use cosine similarity over the Euclidan distance?**

- Cosine similarity is preferable over Euclidean distance when the magnitude of vectors is not as important as their orientation. This is common in high-dimensional spaces such as text analysis (e.g., comparing document similarity) or recommendation systems, where angle-based similarity provides more meaningful results than raw distance.

g. **What does the jaccard distance account for that the Manhattan distance doesn't?**

- The Jaccard distance accounts for the similarity between sets by measuring the ratio of shared elements to the total number of unique elements. In contrast, the Manhattan distance purely sums absolute differences in numerical values without considering shared characteristics. Jaccard distance is useful for categorical or binary data, while Manhattan distance is more appropriate for continuous numerical data.