Assignment 1 – DATA 240

Q1 (30%)) Assume each object is characterized by a set of continuous- valued attributes and answer the following questions. Please explain your answers. Simple yes or no would not be accepted.

1. a) If two objects have a cosine similarity of 1, must their attribute values be identical? Explain.
2. b) If two objects have a correlation value of 1, must their attribute values be identical? Explain.
3. c) If two objects have a Euclidean distance of 0, must their attribute values be identical? Explain.
4. d) Let x and y be the attribute vectors of two objects. State whether the following proximity measures—cosine, correlation, and Euclidean distance—are invariant (unchanged) under the following transformations. Specifically, if $x \rightarrow x'$ and $y \rightarrow y'$, would $\text{cosine}(x, y) = \text{cosine}(x', y')$, $\text{correlation}(x, y) = \text{correlation}(x', y')$, and $\text{Euclidean}(x, y) = \text{Euclidean}(x', y')$. Please explain your answers by referring to the formula for each proximity measures. Simple yes or no would not be accepted.
   I. Translation: $x \rightarrow x + c$ and $y \rightarrow y + c$, where c is a constant added to each attribute value in x and y.
   II. Scaling: $x \rightarrow cx$ and $y \rightarrow cy$, where c is a constant multiplied to each attribute value in x and y.
   III. Standardization: $x \rightarrow \frac{x - \bar{x}}{\sigma_x}$ and, $y \rightarrow \frac{y - \bar{y}}{\sigma_y}$ where c and d are constants.

Ans1(a):

If two objects have a cosine similarity of 1, it doesn't necessarily mean their attribute values are identical because cosine similarity measures the cosine angle between two vectors in a multidimensional space, indicating the similarity of their directions. If the cosine similarity between two objects is 1, it means their attribute vectors are pointing in the same direction, but not necessarily with the same magnitude. Therefore, while their attribute values may have the same relative direction, they could still have different magnitudes.

Ans1(b):

If two objects have a correlation value of 1, it means that there is a perfect linear relationship between their attribute values. However, this does not necessarily imply that their attribute values are identical.

For example: One object's attribute value might be twice the magnitude of the other object's attribute values, or they might have different intercepts. Despite these differences in scale or

offset, their attribute values would still exhibit a perfect linear relationship, resulting in a correlation coefficient of 1.

Ans1(c):

If two objects have a Euclidean distance of 0, it means that they are located at the same point or location in the attribute space. However, this does not necessarily imply that their attribute values are identical.

Euclidean distance measures the straight-line distance between two points in a multidimensional space. When the Euclidean distance between two objects is 0, it indicates that their attribute values are the same in all dimensions, it does not guarantee that the attribute values are identical, as there could still be differences in scale or other characteristics.

d.

I. Translation: x -> x+c and y -> y+c, where c is a constant added to each attribute value in x and y.

Ans :

Ans (d)

I Translation: $x \to x+c$
and $y \to y+c$

where c is a constant added to each attribute value in x and y.

The impact of the translation transformation

$x \to x+c$ and $y \to y+c$ where c is a constant added to each attribute value in x and y on the three proximity measures:

i) Cosine similarity :- The cosine similarity between two vectors x and y is defined as

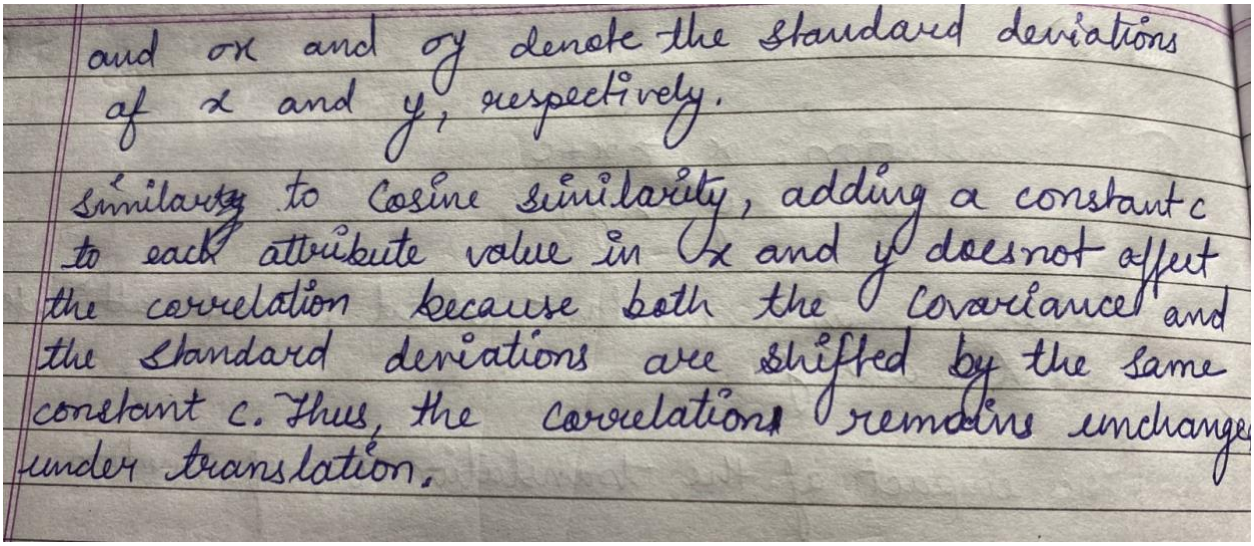
$$\text{cosine}(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

If we add a constant c to each attribute value in x and y does not affect the cosine similarity because both the numerator (dot product) and the denominators (magnitude of vectors) are shifted by the same constant c. Thus, the cosine similarity remains unchanged under Translations.

ii) Correlation: The correlation between two vectors x and y is defined as:

$$\text{correlation}(x,y) = \frac{\text{cov}(x,y)}{\sigma_x, \sigma_y}$$

where cov(x,y) denotes the covariance between x and y



III. Euclidean Distance: Euclidean distance is invariant to translation because it measures the straight-line distance between two points in space and is unaffected by shifts in their position.

The formula for Euclidean distance between vectors x and y is:

Euclidean(x,y)=‖x−y‖

Adding a constant c to each attribute value in x and y does not affect the Euclidean distance because it shifts both vectors by the same amount, preserving their distance.

II. Scaling:

Cosine Similarity: Cosine similarity is invariant to scaling because it measures the cosine of the angle between two vectors and is unaffected by changes in their length. Scaling both vectors x and y by a constant c does not change their direction, only their magnitude, which does not affect cosine similarity.

1. Correlation: Correlation is also invariant to scaling because it measures the linear relationship between two variables and is unaffected by changes in their scale. Scaling both vectors x and y by a constant c does not change their linear relationship, if the scaling factor is applied uniformly to both variables.

2. Euclidean Distance: Euclidean distance is not invariant to scaling because it measures the straight-line distance between two points in space, which changes with scale. Scaling both vectors x and y by a constant c multiplies their distance by the same factor, affecting the Euclidean distance.

III Standardization: x-> x-c/d and y -> y-c/d where c and d are constants.

1. Cosine Similarity: Cosine similarity is invariant to standardization because it measures the cosine of the angle between two vectors, which is unaffected by shifts and scaling. Standardizing both vectors x and y by subtracting their mean and dividing by their standard deviation does not change their direction, only their scale, which does not affect cosine similarity.
2. Correlation: Correlation is invariant to standardization because it measures the linear relationship between two variables, which is unaffected by shifts and scaling. Standardizing both vectors x and y preserves their linear relationship by removing shifts and scaling effects.
3. Euclidean Distance: Euclidean distance is not invariant to standardization because it measures the straight-line distance between two points in space, which changes with standardization. Standardizing both vectors x and y by subtracting their mean and dividing by their standard deviation changes their distance, as it alters their scale and position.

Q2 (15%)) Consider the following distance measure D between two clusters of data points, X and Y: $D(X, Y) = \min\{d(x, y) : x \epsilon X, y \epsilon Y\}$
where $d(x, y)$ is the Euclidean distance between two data points, x and y. Intuitively, D measures the distance between clusters in terms of the closest two points from each cluster. Does the distance measure satisfy the positivity, symmetry, and triangle inequality properties? For each property, show your proof clearly or give a counter- example if the property is not satisfied.

Ans: Let's analyze each property:

a) Positivity: For any two clusters X and Y, the distance measure D(X, Y) should be non-negative.
   Proof: The Euclidean distance d(x, y) between any two points x and y is always non-negative. Taking the minimum over non-negative values will still yield a non-negative value.
   Thus, D(X, Y) = min{d(x, y) : x ∈ X, y ∈ Y} will also be non-negative.

b) Symmetry: The distance measure D(X, Y) should be equal to D(Y, X) for any clusters X and Y.
   Proof: The minimum operation is symmetric; that is, min{a, b} = min{b, a}.
   Therefore, D(X, Y) = min{d(x, y) : x ∈ X, y ∈ Y} = min{d(y, x) : y ∈ Y, x ∈ X} = D(Y, X).

c) Triangle Inequality: For any three clusters X, Y, and Z, the distance measure satisfies: D(X, Z) ≤ D(X, Y) + D(Y, Z). Proof: Let's consider three clusters X, Y, and Z. For any point x ∈ X and z ∈ Z, we have: d(x, z) ≤ d(x, y) + d(y, z) for some point y ∈ Y (by the triangle inequality for Euclidean distance).

Thus, $\min\{d(x, z) : x \in X, z \in Z\} \leq \min\{d(x, y) + d(y, z) : x \in X, y \in Y, z \in Z\}$. By the definition of D, $D(X, Z) \leq D(X, Y) + D(Y, Z)$.
Therefore, the distance measure D satisfies the positivity, symmetry, and triangle inequality properties.

Q3 (20%)) Suppose you are given a census data, where every object corresponds to a household and the following continuous attributes are used to characterize each household: total household income, number of house- hold residents, property value, number of bedrooms, and number of vehicles owned. Suppose we are interested in clustering the households based on these attributes.

a) Explain why cosine is not a good measure for clustering the data.
b) Explain why correlation is not a good measure for clustering the data.
c) Explain what preprocessing steps and corresponding proximity measure you should use cluster the data.

Ans : a) Cosine similarity measures the cosine of the angle between two vectors and is commonly used for text data or high-dimensional sparse data where the magnitude of the vectors is not as important as the orientation. In the context of household attributes such as total household income, number of residents, property value, etc., cosine similarity may not be appropriate because it does not take into account the magnitude of the vectors, only their orientation. For example, two households could have very similar attribute values but vastly different magnitudes, leading to misleading similarity scores.

b) Correlation measures the linear relationship between two variables and is commonly used to determine how much one variable changes when another variable changes. However, correlation may not be suitable for clustering household data because it assumes a linear relationship between variables, which may not hold true for all attributes in the dataset. Additionally, correlation measures the strength and direction of the relationship between two variables but may not capture complex relationships or interactions among multiple variables simultaneously.

d) To preprocess the data for clustering, it's important to standardize or normalize the continuous attributes to ensure that they are on the same scale and have comparable influence on the clustering process. This can be achieved by subtracting the mean and dividing by the standard deviation (standardization) or by scaling the values to a range such as [0, 1] (normalization). Once the data is preprocessed, a suitable proximity measure for clustering could be Euclidean distance. Euclidean distance considers both the direction and magnitude of the vectors, making it suitable for continuous attribute data. Additionally, other distance measures such as Manhattan distance or Mahalanobis

distance could also be considered depending on the distribution and characteristics of the data.

Q4 (15%)) Consider a weighted, undirected, graph G. Let $e(u, v)$ be the weight of the edge between nodes u and v, where e$(u, u) = 0$ and $e(u, v) = \infty$ if u and v is disconnected.

Assume the graph is a connected component, i.e., there exists a path between every two nodes. Suppose the path length, $d(u, v)$, is defined as follows:

$0, if\ u = v$;
$d(u, v) = f(x) =$ I$e(u, v), if\ there\ is\ an\ edge\ between\ u\ and\ v$;

$min_{\&'(')} d(u, w) + d(w, v), otherwise$;
Is $d(u, v)$ a metric? State your reasons clearly. (Check whether the positivity, symmetry, and

triangle inequality properties are preserved).

Ans 4:

Ans 4

$$d(u,v) = f(x) = \begin{cases} 0, & \text{if } u = v, \\ e(u,v), & \text{if there is an edge between } u \text{ and } v; \\ \min_{w \neq u \neq v} d(u,w) + d(w,v), & \text{otherwise} \end{cases}$$

Checking : Positivity, Symmetry and the triangle equality To determine whether $d(u,v)$ is a metric,

① Positivity : $d(u,v)$ should be non negative for all u and v, and it should be zero if and only if $u = v$

- $d(u,v) = 0$ if and only if $u = v$ (directly from the definition)
- $d(u,v) \geq 0$ since its a sum of non-negative values (edge weights or the minimum of sums).

② Symmetry :- $d(u,v) = d(u,v)$ for all u and v.

- $d(u,v) = \min(e(u,v), \min_w (d(u,w) + d(w,v)))$
- $d(v,u) = \min(e(v,u), \min_w (d(v,w) + d(w,u)))$
- Since $(u,v) = e(v,u)$ (since the graph is undirected), and the minimum operation is commutative, $d(u,v) = d(v,u)$

(3) Triangle inequality :- for all u, v and w,

- $d(u,v) \leq d(u,w) + d(w,v)$
- $d(u,v) = \min(e(u,v), \min_w (d(u,w) + d(w,v)))$
- $d(u,w) + d(w,v) \geq \min(e(u,w) + e(w,v), \min_x (d(u,x) + d(x,w)) + \min_y (d(w,y) + d(y,v)))$
- By the triangle inequality for edge weights,

  $e(u,v) \leq e(u,w) + e(w,v)$
- Therefore, $d(u,v)$ satisfies the triangle inequality.

Since $d(u,v)$ satisfies all three properties it is indeed a metric.

Ans5a.

Q5 Consider the following data matrix on the right, in which two of its values are missing (the matrix on the left shows its true values).

-0.2326	0.2270
-0.0847	0.7125
0.1275	0.3902
0.1329	-0.1461
0.3724	0.1756
0.4975	0.8536
0.6926	0.7834
0.7933	0.7375
0.8229	0.2147
0.8497	0.4980
1.0592	0.7600
1.5028	1.0122

→

-0.2326	0.2270
-0.0847	0.7125
0.1275	0.3902
?	-0.1461
0.3724	-0.1756
0.4975	0.8536
0.6926	0.7834
0.7933	0.7375
0.8229	0.2147
0.8497	0.4980
1.0592	?
1.5028	1.0122

Ans: a Let's impute missing values for the matrix on the right by their respective column averages

Average column 1

$$= \frac{\{-0.2326 + (-0.0847) + 0.1275 + 0.3724 + 0.4975 + 0.6926 + 0.7933 + .8229 + .8497 + 1.0592 + 1.5028\}}{11}$$

$$= \frac{\{6.4006\}}{11} \Rightarrow 0.5818$$

Average column 2

$$= \frac{\{0.2270 + 0.7125 + 0.3902 + (-0.1461) + 0.1756 + 0.8536 + 0.7834 + 0.7375 + 0.2147 + 0.4980 + 0.4962 + 1.0122\}}{11}$$

$$= 5.4586 / 11 = 0.4962$$

RMSE for column 1 imputed value

$$RMSE = \sqrt{\frac{(A_{4,1} - \hat{A}_{4,1})^2 + (A_{11,2} - \hat{A}_{11,2})^2}{2}}$$

$A_{4,1} = 0.1329$, $\hat{A}_{4,1} = 0.5818$

$A_{11,2} = 0.7600$, $\hat{A}_{11,2} = 0.4962$

Putting values of eqn ① and ② in formula for RMSE

$$RMSE = \sqrt{\frac{(0.1329 - 0.5818)^2 + (0.7600 - 0.4962)^2}{2}}$$

$$RMSE = \sqrt{\frac{(-0.4489)^2 + (0.2638)^2}{2}} \Rightarrow \sqrt{\frac{0.20151121 + 0.01077444}{2}}$$

$$RMSE = \sqrt{\frac{0.21228565}{2}} \Rightarrow \sqrt{0.10614282} \Rightarrow 0.32579$$

$$RMSE = \sqrt{\frac{0.2015 + 0.0695}{2}} = \sqrt{0.1355} = 0.3681$$

Q5(b)

Ans b $$\mu_{i|j} = \hat{\mu}_i + \Sigma_{ij} \Sigma_{jj}^{-1} (x_j - \hat{\mu}_j)$$

$$\Sigma_{ij} \Sigma_{jj}^{-1} (x_j - \hat{\mu}_j)$$

First Iteration ⇒ mean value without missing value column 1

for col 1 is 0.5818 for col 2 is 0.4962

mean value for each column using both the non missing and imputed values.

for col 1 = $\frac{6.4006 + 0.5818}{12} \Rightarrow 0.5818$

for col 2 $\frac{5.4586 + 0.4962}{12} \Rightarrow 0.4962$

Given $\Sigma = \begin{bmatrix} 0.25 & 0.1 \\ 0.1 & 0.15 \end{bmatrix}$.

Inverse matrix of Σ

$$\Sigma^{-1} = \begin{bmatrix} 5.45 & -3.63 \\ -3.63 & 9.09 \end{bmatrix}$$

I

$i \to 1$
$j \to 2$

Iteration 1

$$\mu_{1/2} = \hat{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \hat{\mu}_2)$$

$$= 0.5818 + (0.1)(909)\left[(-0.1461) - 0.4962\right]$$

$$= 0.5818 + (0.909)(0.1)(-0.6423)$$

$$\mu_{1/2} = -0.0020507$$

$$\mu_{2/1} = \hat{\mu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \hat{\mu}_1)$$

$$= 0.4962 + (0.1)(5.45)(1.0592 - 0.5818)$$

$$\boxed{\mu_{2/1} = 0.7563}$$

Iteration 2 Imputing the calculated value and calculating mean

~~$\hat{\mu}_1 = 0.5$~~

for $\hat{\mu}_1 = 0.5332$

$\hat{\mu}_2 = 0.5178$

$$\mu_{1/2} = 0.5332 + (0.909)\left[(-0.1461) - (0.5178)\right]$$

$$\boxed{\mu_{1/2} = -0.0702}$$

$$\mu_{2/1} = 0.5178 + (0.545)(1.0592 - 0.5332)$$

$$\boxed{\mu_{2/1} = 0.8044}$$

Iteration 3: $\hat{\mu}_1 = 0.5274$

$\hat{\mu}_2 = 0.5218$

$\mu_{1/2} = 0.5274 + 0.909(-0.1461 - 0.5218)$

$\mu_{1/2} = -0.079$

$\mu_{2/1} = 0.5218 + 0.545(1.0592 - 0.5274)$

$\mu_{2/1} = 0.8116$

Iteration 4: $\hat{\mu}_1 = 0.5267$

$\hat{\mu}_2 = 0.5224$

$\mu_{1/2} = (0.5267) + 0.909(-0.1461 - 0.5224)$

$\mu_{1/2} = -0.055$

$\mu_{2/1} = 0.5224 + 0.545(1.0542 - 0.5267)$

$\mu_{2/1} = 0.812$

Iteration 5: $\hat{\mu}_1 = 0.5262$

$\hat{\mu}_2 = 0.5225$

$\mu_{1/2} = 0.5262 + 0.909(-0.1461 - 0.5225)$

$= -0.0815$

$\mu_{2/1} = 0.5225 + 0.545(1.0592 - 0.5262)$

$= 0.813$

RMSE → Iteration 1

$$RMSE = \sqrt{\frac{[0.1329 - (-0.00205)]^2 + (0.76 - 0.7563)^2}{2}}$$

$$= 0.0949$$

Iteration 2

$$RMSE \rightarrow \sqrt{\frac{(0.1329 - (-0.0702))^2 + (0.76 - 0.8044)^2}{2}}$$

$$= 0.147$$

Iteration 3

$$RMSE \rightarrow \sqrt{\frac{(0.1329 + 0.079)^2 + (0.76 - 0.8116)^2}{2}}$$

Iteration 4

$$RMSE = \sqrt{\frac{(0.1329 - (-0.085))^2 + (.76 - 0.812)^2}{2}}$$

$$= 0.154$$

Iteration 5

$$RMSE = \sqrt{\frac{(0.1329 + 0.0815)^2 + (0.76 - 0.813)^2}{2}}$$