TICS411- Prueba 1

MINERÍA DE DATOS

Universidad Adolfo Ibáñez

2024-1

Profesores: Claudio Díaz - Miguel Carrasco - Alfonso Tobar

Fecha: 30/04/24

Nombre: _____ **Rut**: _____ **Sección**: _____

---

*Esta prueba contiene **11 páginas** y **12 preguntas** totalizando 30 puntos. Además, en la parte final dispone de un **Formulario**. Recuerde responder con letra clara y legible.*
¡Buena suerte!

Parte I: Preguntas de Selección Múltiple (20 min)

1. (1 punto) Indique cuál(es) de las siguiente(s) afirmación(es) son verdaderas. "Si es que existe un valor faltante en mi base de datos, uno puede _____."

   I. Eliminar los atributos con datos faltantes

   II. Ignorar los datos faltantes

   III. Reemplazar con el promedio de datos los posibles datos

   A. Sólo I

   B. Sólo II

   C. Sólo III

   D. I y III

   E. I, II y III

2. (1 punto) Cuál de los siguientes algoritmos de agrupación entregaría el **peor coeficiente de silueta** para los datos en la Figura 1:

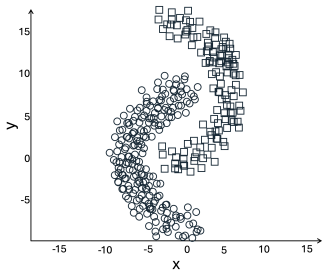

Figura 1: Datos Bidimensionales

   A. K-Means.

   B. Apriori.

   C. Jerárquico.

   D. DBSCAN.

3. (1 punto) ¿Cuál es uno de los principales roles de **"epsilon"** en el algoritmo DBSCAN?
   A. Define el número mínimo de puntos requeridos para formar un Core Point.
   B. Especifica la distancia máxima entre dos puntos para que se consideren parte del mismo grupo.
   C. Es el número de clústeres que formará el Algoritmo.
   D. Representa la distancia entre los centroides en los clústeres resultantes.

4. (1 punto) Con respecto a la evaluación de un proceso de clustering. ¿Qué afirmación es más correcta con respecto a la interpretación del Coeficiente de Silueta?
   A. Valores de Silueta más altos sugieren una mejor cohesión y separación de los clústeres, lo que indica clústeres distintos y definidos.
   B. Valores Silueta más bajos indican distancias grandes entre clústeres y una cohesión intracluster más débil, teniendo entonces un modelo de clústers menos confiable.
   C. Valores silueta más altos indican que hay presencia de datos atípicos.
   D. Valores de silueta más bajos son indicativos de clústeres de baja cohesión, compuestos por una baja cantidad de puntos.

5. (1 punto) La imagen de la Fig.2 muestra un agrupamiento jerárquico. En particular, ¿qué tipo de linkage se está usando?
   A. Single. B. Complete. C. Manhattan D. Centroid.

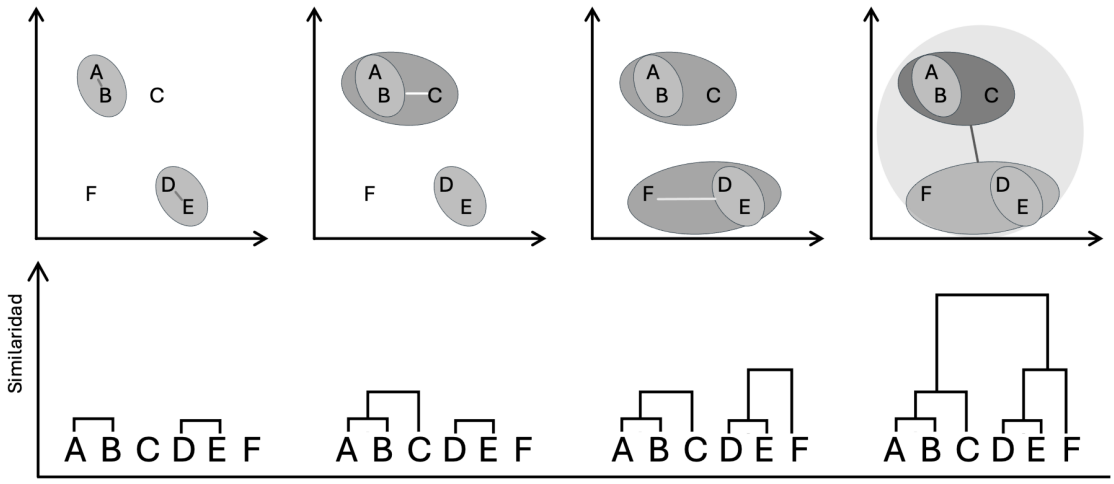

Figura 2: Clustering Jerárquico

Parte II: Preguntas de Desarrollo (30 min)

6. (2 puntos) Indique qué características tiene el Histograma y el Gráfico de Barras. Compárelos.

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

7. (1½ puntos) En la Figure 3, indique sobre las líneas correspondientes las Distancias de Minkowski para $r = 1, 2$ e ∞. Es importante que la respuesta sea precisa y fácil de entender para el corrector para evitar confusión (Si la respuesta no es clara podría resultar en una pérdida de puntaje).

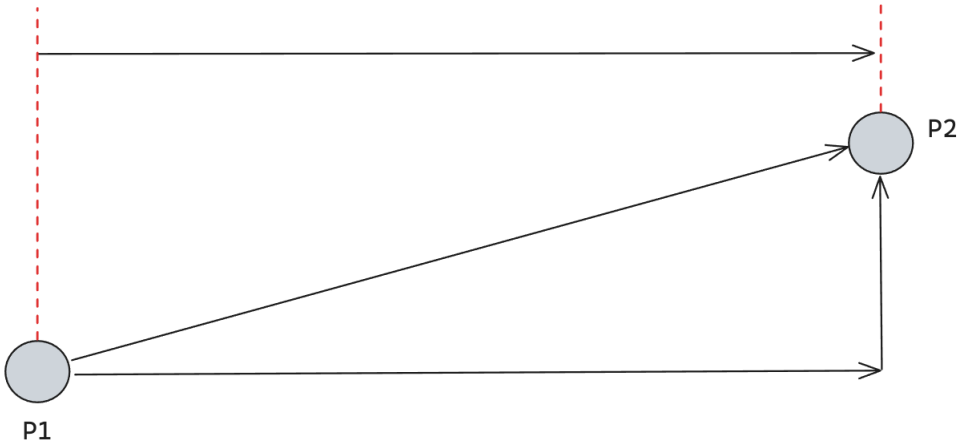

Figura 3: Distancias entre P1 y P2. Las líneas representan diferentes modelos de distancias.

8. (4 puntos) Nombre cuales son las etapas que componen la Minería de Datos. Explique brevemente cada una de ellas.

Parte III: Ejercicios Prácticos (60 min)

9. Suponga la siguiente Matriz de Distancias (ver Tabla 1) entre 10 puntos existentes en el espacio.

	1	2	3	4	5	6	7	8	9	10
1	0.00	6.00	11.00	14.00	17.00	4.00	7.21	11.18	14.14	17.12
2	6.00	0.00	5.00	8.00	11.00	7.21	4.00	5.39	8.25	11.18
3	11.00	5.00	0.00	3.00	6.00	11.70	6.40	2.00	3.61	6.32
4	14.00	8.00	3.00	0.00	3.00	14.56	8.94	3.61	2.00	3.61
5	17.00	11.00	6.00	3.00	0.00	17.46	11.70	6.32	3.61	2.00
6	4.00	7.21	11.70	14.56	17.46	0.00	6.00	11.18	14.14	17.12
7	7.21	4.00	6.40	8.94	11.70	6.00	0.00	5.39	8.25	11.18
8	11.18	5.39	2.00	3.61	6.32	11.18	5.39	0.00	3.00	6.00
9	14.14	8.25	3.61	2.00	3.61	14.14	8.25	3.00	0.00	3.00
10	17.12	11.18	6.32	3.61	2.00	17.12	11.18	6.00	3.00	0.00

Tabla 1: Matriz de Distancias entre 10 puntos.

Asumiendo que el **Punto 1** y el **Punto 10** son los centroides iniciales de un modelo **K-Means**. Responda las siguientes preguntas:

(a) (1 punto) ¿Cuántos Clusters se están buscando?

(b) (2 puntos) Luego de la **primera iteración** del proceso de K-Means. ¿Qué puntos pertenecen a cada cluster al final de la iteración? Muestre su desarrollo.

10. Usted dispone de los siguientes datos tabulados de un juego con formas y colores. El dataset está compuesto de 8 registros y 2 características como se ve en la Tabla 2:

Item	Forma	Color
1	cuadrado	rojo
2	cuadrado	azul
3	circulo	rojo
4	circulo	verde
5	triangulo	rojo
6	cuadrado	amarillo
7	circulo	amarillo
8	circulo	verde

Tabla 2: Dataset

(a) (1 punto) ¿A qué tipo de variables corresponden los datos de **Forma** y **Color**?

(b) (2 puntos) ¿Cuál es la matriz de frecuencia de atributos asociada a la tabla anterior?

(c) (2 puntos) Para la combinación de items 1-3 y 3-7. ¿Qué pares son **más similares** según la métrica de Goodall? Justifique su respuesta.

(d) (2 puntos) ¿Cuál sería la matriz de transformación binaria del atributo **Forma** para la matriz anterior.

11. Una fábrica de frutas ha decidido contratarlo a usted para implementar un sistema que automáticamente agrupe dos tipos de fruta que actualmente procesa. La empresa dispone de un conjunto de datos que ha obtenido mediante una cámara de inspección industrial instalada en su línea de producción. A través de ella ha podido obtener dos característias: ***tamaño*** y ***color***. Debido a su experiencia en Algoritmos de Minería de Datos, se le ha asignado a usted emplear el algoritmo DBSCAN sobre el conjunto de datos y una matriz de distancias Euclideanas en la Tabla 3 (ver puntos en Fig.4). La empresa no está segura de la aplicación del algoritmo, es por ello que realiza las siguientes preguntas técnicas sobre el algoritmo que usted debe responder justificadamente.

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.0	0.5	1.0	1.6	0.8	1.3	3.6	3.6	4.3	4.5	5.0	4.0
2	0.5	0.0	0.5	1.1	0.8	1.4	3.5	3.3	4.0	4.2	4.7	4.2
3	1.0	0.5	0.0	0.8	0.8	1.3	3.0	2.8	3.5	3.7	4.2	4.1
4	1.6	1.1	0.8	0.0	1.6	2.1	3.4	3.0	3.7	3.8	4.4	4.8
5	0.8	0.8	0.8	1.6	0.0	0.6	2.9	2.9	3.6	3.8	4.3	3.4
6	1.3	1.4	1.3	2.1	0.6	0.0	2.5	2.7	3.3	3.6	4.0	2.8
7	3.6	3.5	3.0	3.4	2.9	2.5	0.0	0.9	1.0	1.4	1.6	2.7
8	3.6	3.3	2.8	3.0	2.9	2.7	0.9	0.0	0.7	0.9	1.4	3.6
9	4.3	4.0	3.5	3.7	3.6	3.3	1.0	0.7	0.0	0.4	0.7	3.7
10	4.5	4.2	3.7	3.8	3.8	3.6	1.4	0.9	0.4	0.0	0.6	4.1
11	5.0	4.7	4.2	4.4	4.3	4.0	1.6	1.4	0.7	0.6	0.0	4.1
12	4.0	4.2	4.1	4.8	3.4	2.8	2.7	3.6	3.7	4.1	4.1	0.0

Tabla 3: Distancia entre cada punto de la Fig.4

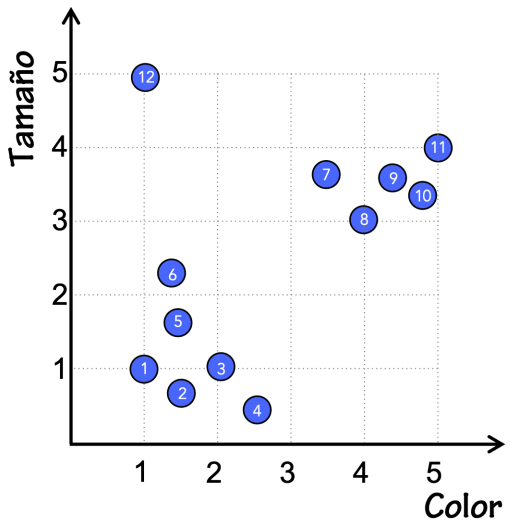

Figura 4: Gráfico de puntos de las variables tamaño y color

(a) (1 punto) La empresa sugiere usar $eps = 5$ y $minPts = 4$. ¿Cuántos clústers encontraría con estos parámetros? Justifique su respuesta.

(b) (2 puntos) Dado los parámetros $eps = 1$ y $minPts = 2$, dibuje la clusterización resultante. Para ello indique la etiqueta de cada punto en el gráfico y si el punto es ruido, border o Core.

12. Suponga que usted dispone de dos grupos de datos definidos como Cluster A y Cluster B (ver Figura 5). A continuación, responda las siguientes preguntas.

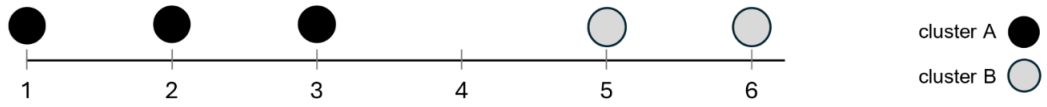

Figura 5: Coeficiente Silueta

(a) (2½ puntos) Complete los valores de la Tabla 4 y calcule el coeficiente de Silhouette para cada punto empleando la distancia L_1.

	a_i	b_{iA}	b_{iB}	b_i	s_i
1					
2					
3					
5					
6					

Tabla 4: Tabla Silhouette

(b) (1 punto) Calcule el Coeficiente de Silhouette Promedio.

(c) (1 punto) ¿Qué significa el valor anterior?

Formulario

Similaridad
Sea p y q los valores de un atributo para dos puntos.

Tipo atributo	Disimilaridad	Similitud
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	d=|p-q| /(n-1)	s=1-|p-q| /(n-1)
Interval o ratio	d=|p-q|	s=-d; s=1/(1+d)

Overlap

$$S(p_i, q_i) = \begin{cases} 1, & \text{if } p_i = q_i \\ 0, & \text{if } p_i \neq q_i \end{cases}$$

Ocurrencia Inversa

$$S(p_i, q_i) = \frac{1}{p_k(p_i)^2}$$

Goodall

$$S(p_i, q_i) = 1 - p_k(p_i)^2$$

Similitud entre vectores Binarios
Sea p y q vectores de atributos binarios y M_{XY} = El número de atributos donde p es X y q es Y.

$$\boldsymbol{SMC} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} \qquad \boldsymbol{JC} = \qquad = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

Estadísticos

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x(i) - \mu)^2$$

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n S_x S_y}$$

Distancia de Minkowski: Sea p y q vectores m dimensionales:

$$d(p, q) = \left(\sum_{k=1} (p_k - q_k)^r \right)^{1/r}$$

- Para $r = 1 \rightarrow$ Distancia Manhattan.
- Para $r = 2 \rightarrow$ Distancia Euclideana.
- Para r = $\infty \rightarrow$ Distancia Chebyshev.

Distancia de Mahalanobis: $d(p, q) = \sqrt{(p - q)^T \Sigma^{-1} (p - q)}$.

Centroide de un Cluster:

$$r_k = \frac{1}{n_k} \sum_{x(i) \in C_k} x(i)$$

Within-Cluster-Distance

$$wc(C) = \sum_{k=1} wc(C_k) = \sum_{k=1}^{K} \sum_{x(i) \in C_k} d(x(i), r_k)$$

Between-Cluster-Distance

$$bc(C) = \sum_{1 \leq j \leq k \leq K} d(r_j, r_k)$$

- **Single Linkage**: $D(C_i, C_j) = \mathbf{min}\{d(x, y) | x \in C_i, y \in C_j\}$
- **Complete Linkage**: $D(C_i, C_j) = \mathbf{max}\{d(x, y) | x \in C_i, y \in C_j\}$
- **Average Linkage**: $D(C_i, C_j) = \mathbf{avg}\{d(x, y) | x \in C_i, y \in C_j\}$

Hopkins

$$\mathrm{H} = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

- p: Número de Puntos de muestra.
- w_i: Distancia desde un punto aleatorio al vecino más cercano en los datos.
- u_i: distancia de un punto simulado al vecino más cercano en los datos.

Cohesión y Separación

$$SSE = \sum_{k=1}^{K} \sum_{x \in C_k} (X - \bar{C_k})^2$$

$$SSB = \sum_{k=1}^{K} |C_k| (\bar{C_k} - \bar{X})^2$$

Coeficiente de Silhouette

- a_i: Distancia promedio del punto i al los otros puntos del mismo cluster.
- b_{ij}: Distancia promedio del punto i a todos los puntos del cluster j.
- b_i: Mínimo b_{ij} tal que el punto i no pertenezca al punto j.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

$$S = \frac{1}{n} \sum_{i=1}^{n} s_i$$