UNIVERSIDAD ADOLFO IBÁÑEZ

TICS411- Examen

MINERÍA DE DATOS

Universidad Adolfo Ibáñez

2024-1

Profesores: Claudio Díaz - Miguel Carrasco - Alfonso Tobar
Fecha: 10/07/24

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

---

*Esta prueba contiene **14 páginas** y **10 preguntas** totalizando 20 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) Suponga que usted dispone de un dataset con **n registros** y **m características** de distinta escala. Además, todas las características son del tipo intervalo. A continuación, se presenta una serie de afirmaciones las cuales algunas son verdaderas y otras falsas. **Lea atentamente cada una de ellas.**

   I. No es posible aplicar la distancia de Mahalanobis ya que esta solo funciona para datos del tipo ratio.

   II. El máximo número teórico de clusters es n ya que no es posible generar más clusters que registros en el dataset.

   III. La normalización siempre se aplica a cada registro (por filas) y no por sus características (columnas)

   IV. Mientras más características tenga el data set, mejor es su normalización.

   V. La distancia de Mahalanobis pondera la escala entre las variables a través de la matriz inversa de covarianza.

   Responda cuáles de las siguientes alternativas son verdaderas.

   A. III y IV

   B. III y V.

   C. I y IV

   D. II y V

   E. Todas son verdaderas

2. (1 punto) Una empresa de análisis de datos se encuentra con un grave problema. El proceso de recolección de datos posee errores de digitación debido a que los encuestadores ingresaron mal algunos campos. A continuación se presenta una serie de afirmaciones las cuales algunas son verdaderas y otras falsas. Lea atentamente cada una de ellas.

   I. Un algoritmo apropiado para reducir el número de datos atípicos (outliers) es el algoritmo K-Means ya que utiliza la distancia promedio hacia los k-ésimos puntos más alejados

   II. El algoritmo DBSCAN resulta apropiado siempre y cuando sean seleccionados en forma adecuada los parámetros del algoritmo

   III. Una forma adecuada corresponde al cálculo de las distancias entre los vecinos. Para ello calculamos el promedio de distancia al k-esimo vecino más cercano, y según los valores obtenidos podemos aplicar un umbral de decisión.

   IV. Debido a que los datos fueron mal ingresados, lo apropiado en estos casos es normalizar los datos y luego aplicar una técnica estadística como el test de Chi-cuadrado.

   Responda cuáles de las siguientes alternativas es verdadera.

   A. I y II

   B. III y IV

   C. I y III

   D. II y III

   E. Todas son correctas

Parte II: Preguntas de Desarrollo (30 min)

3. (1 punto) Explique que es la maldición de dimensionalidad (en inglés ***curse of dimensionality***)

4. (1 punto) La cadena de supermercado Jumbox quiere utilizar el **algoritmo Apriori** para identificar conjuntos de artículos frecuentes para generar promociones. Al fijar el umbral mínimo de soporte en 10 %, solo genera conjuntos frecuentes con una cantidad masiva de artículos individuales frecuentes. ¿Por qué se genera este resultado y **cómo podría solucionarse**?

5. (1 punto) Considere los siguientes grupos de datos de acuerdo a la Figura 1.
   - Grupo A (Alta densidad): 50 puntos dentro de un radio de 1 unidad.
   - Grupo B (densidad media): 30 puntos dentro de un radio de 2 unidades.
   - Grupo C (Baja densidad): 20 puntos dentro de un radio de 3 unidades.
   - Valores atípicos: 5 puntos dispersos lejos de estos grupos.

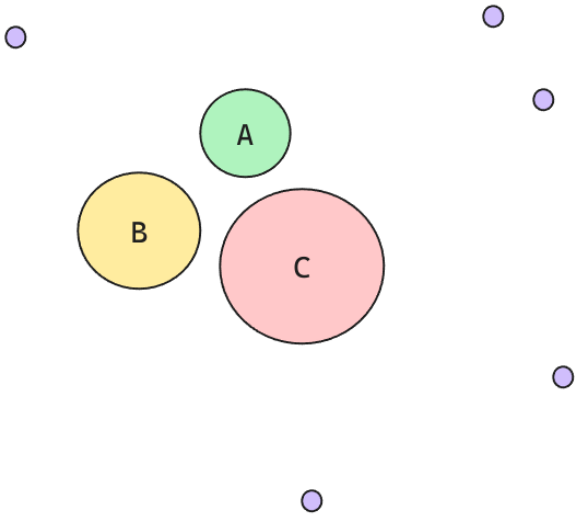


Figura 1: Grupos de Datos

¿Qué método es mejor para detectar valores atípicos en este caso: Métodos basado en distancia (Mahalanobis) o Métodos basados en densidad (LOF)? **Justifique su respuesta con lo visto en el curso.**

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

6. (1 punto) El banco Mundial decidió segmentar países Europeos de acuerdo a variables de relevancia socio-económicas y demográficas utilizando un proceso de Clustering Jerárquico. Se le pide a usted que pueda separar las naciones Europeas en estudio, en regiones donde se podrían aplicar políticas similares para el bienestar de dichos países. Para ello se le hace entrega del siguiente Dendograma:

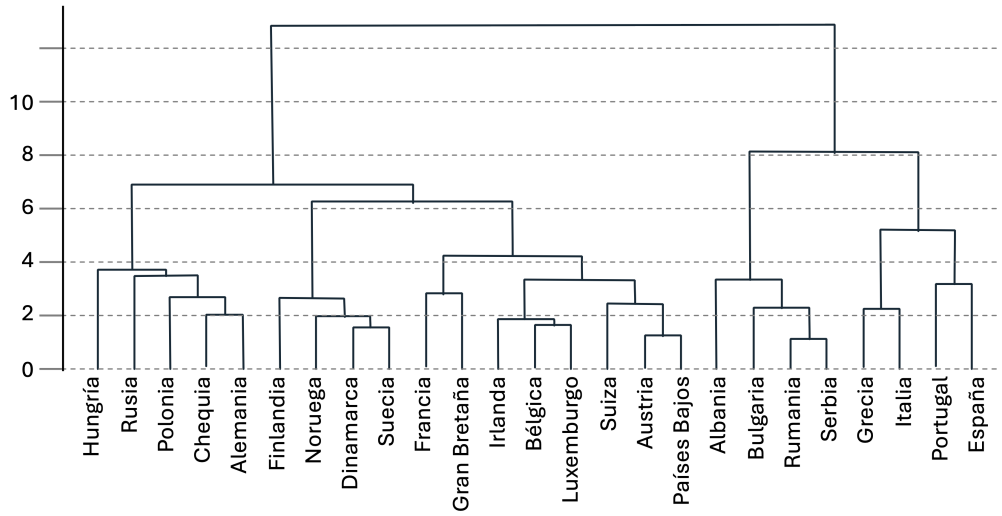


Figura 2: Dendograma

(a) (1 punto) ¿Indique gráficamente qué umbral de distancia debería utilizar para obtener 7 regiones?

(b) (1 punto) Si se utiliza el umbral de distancia 8, ¿Cuántas regiones se obtienen? **Indique qué países pertenecen a cada región.**

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

7. (1 punto) Suponga que usted dispone de una base de datos con transacciones comerciales del banco PrestaMoney. La Unidad de Fraude Financiero (UFF) del banco ha detectado que un porcentaje muy pequeño de las transacciones son de origen fraudulento ¿Qué acción o acciones realizaría usted previo al diseño de un clasificador?

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

Parte III: Ejercicios Prácticos (60 min)

8. Suponga el siguiente Dataset en el cual queremos modelar la variable C mediante el uso de los predictores X1 y X2 utilizando el Algoritmo de Naive Bayes.

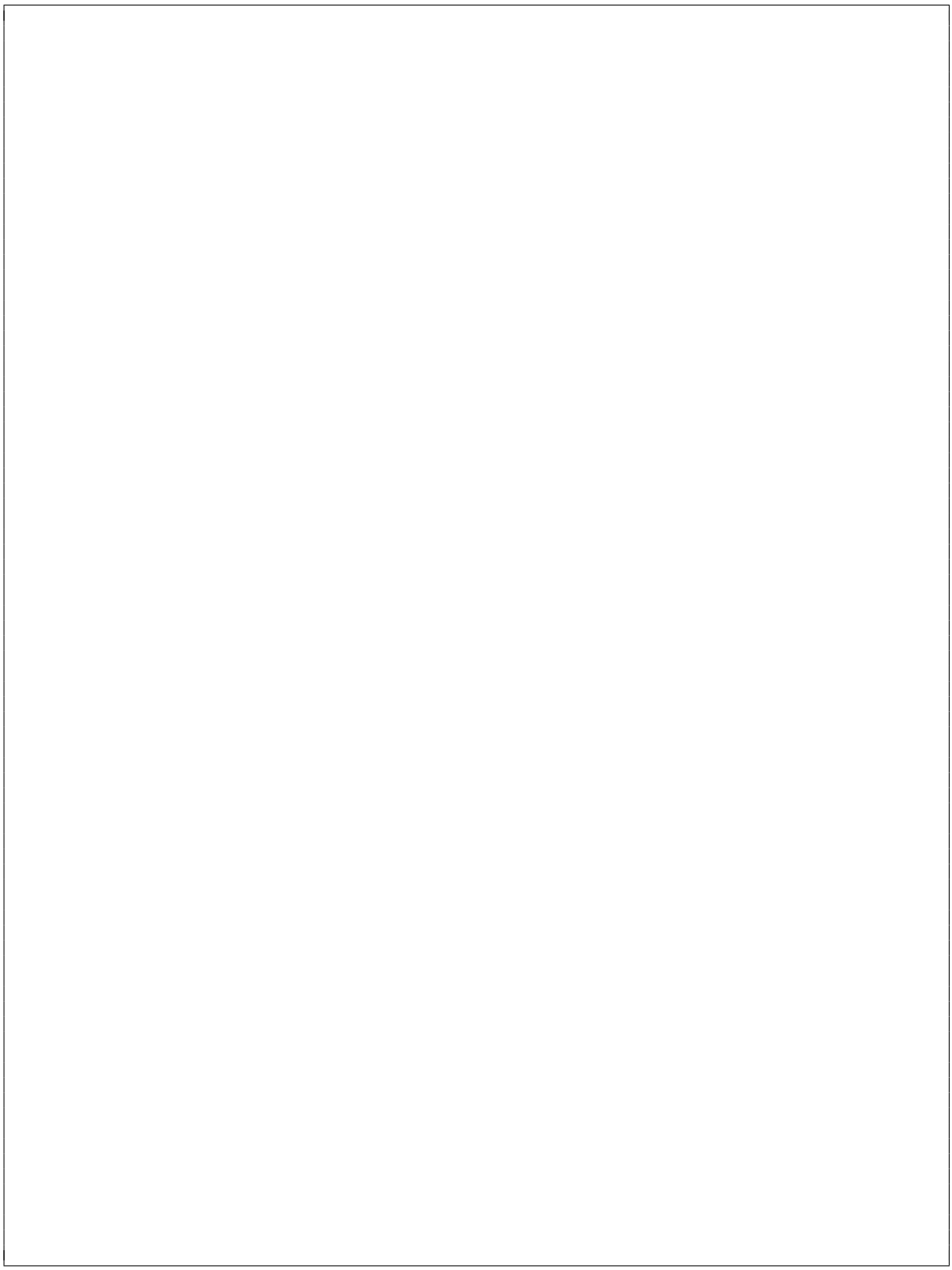
	X1	X2	C
1	Mucho	Alto	1
2	Poco	Bajo	1
3	Mucho	Alto	0
4	Mucho	Medio	1
5	Mucho	Medio	1
6	Mucho	Bajo	1
7	Mucho	Bajo	1
8	Poco	Medio	0
9	Mucho	Alto	0
10	Mucho	Alto	1

Tabla 1: Dataset Naive Bayes

(a) (1 punto) Utilizando una Estrategia de **5-Fold sin Shuffle**. ¿Qué registros pertenecen al set de validación de cada Fold?

(b) (2 puntos) Calcule las predicciones del modelo para el set de Validación del Fold 3. **Hint: Recuerde que los Datos de validación no se utilizan en el entrenamiento del Modelo.**

(c) (1 punto) ¿Cuál sería el Accuracy con las predicciones generadas en b)?



9. Se sabe que un **Modelo de Clasificación Binaria** fue evaluado con 100 registros de los cuales se obtuvieron las siguientes métricas:

   - **Accuracy**: 0.5
   - **Precision**: 0.5
   - **Recall**: 0.4

   (a) (3 puntos) A partir de estos valores, ¿Cómo quedaría la Matriz de Confusión para este problema? **Muestre claramente cómo obtuvo sus resultados.**

10. La unidad de analítica del supermercado LeroMart se encuentra estudiando la posibilidad de contratar a un analista de datos para su área de analítica. Le han mencionado que existe una herramienta denominada Reglas de Asociación, y les gustaría conocer algunos resultados relevantes de esta herramienta. En vista de su experto conocimiento en Minería de Datos, se le ha contratado a usted como consultor externo para demostrar su experiencia con esta herramienta.

    Basado en la información disponible en la tabla con productos (donde cada fila representa una transacción), responda las siguientes preguntas.

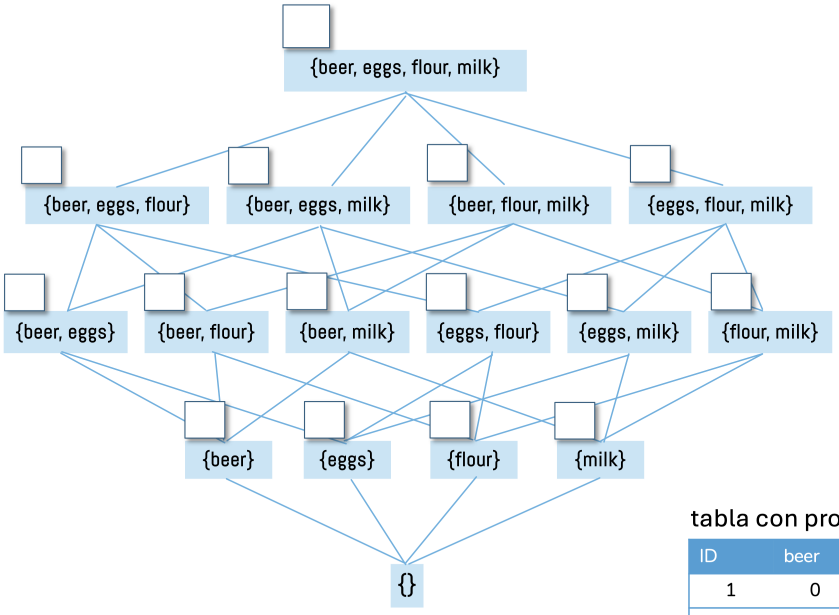


tabla con productos

ID	beer	eggs	flour	milk
1	0	1	1	1
2	1	1	0	0
3	0	0	1	1
4	0	1	1	1
5	0	0	0	1

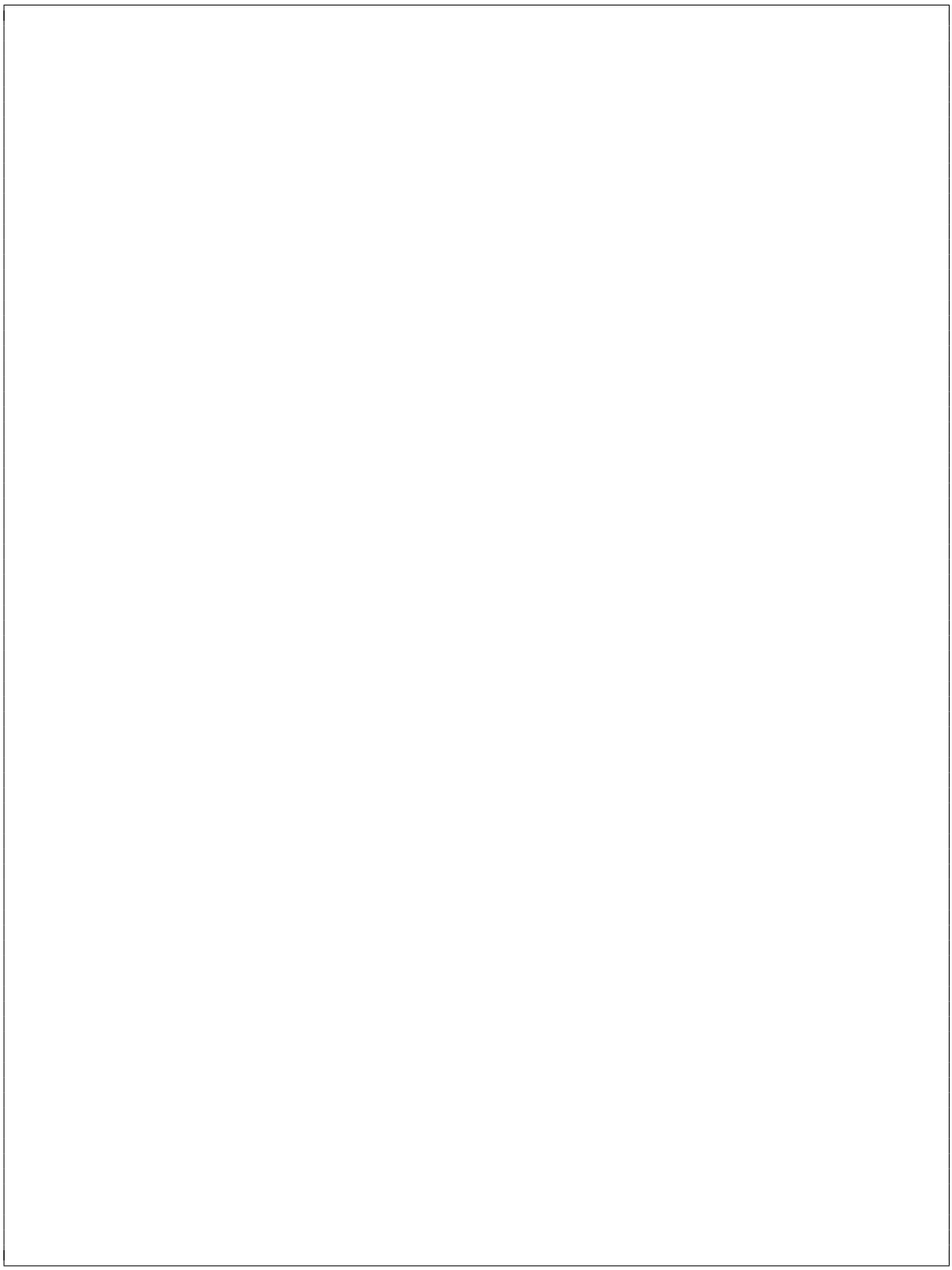
Figura 3: Frequent Itemset

(a) (1 punto) Complete el soporte de cada regla en el espacio en blanco de las combinaciones presentes en la Figura 3.

(b) (1 punto) Si el Soporte es mayor a 3/10, cuáles de los nodos se eliminan y cuáles quedan.

(c) (1 punto) Escriba todas las reglas posibles de construir a partir de los nodos restantes de la pregunta (b).

(d) (1 punto) ¿Cuántas reglas de asociación se pueden generar a partir una lista de objetos de un supermercado con k elementos? **Justique su respuesta.**



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.

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

Apriori

$$Supp(X \to Z) = P(X \cup Z)$$

$$Conf(X \implies Z) = \frac{Supp(X \cup Z)}{Supp(X)}$$

$$Transacciones = \sum_{i=0}^{2^k} \sum_{j=0}^{|U_i|} \binom{|U_i|}{j}$$

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

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

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.

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

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

Scoring Functions

$$\textbf{GeneralScoring(M)} = \sum_{i=1}^{N} d\left[f(x_i); M), y(i)\right]$$

$$\textbf{ZeroOneLoss(M)} = \frac{1}{N}\sum_{i=1}^{N} I\left[f(x_i); M), y(i)\right]$$

$$\textbf{SquaredLoss(M)} = \frac{1}{N}\sum_{i=1}^{N} \left[f(x_i); M) - y(i)\right]^2$$

Confusion matrix		Predicted Class: No	Predicted Class: Yes
Actual Class	No	True Negative	False Positive
	Yes	False Negative	True Positive

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Decision Trees

Gini

$$G(X) = 1 - \sum_{x_i} p(x_i)^2$$

Entropía

$$H(X) = -\sum_{x_i} p(x_i) log_2 p(x_i)$$

Naive Bayes

$$P(C|X_1, X_2, ..., X_k) \propto \prod_{i=1}^{k} P(X_i|C)P(C)$$