MINERIA DE DATOS UTILIZANDO SISTEMAS INTELIGENTES

PRACTICA **2**

AGRUPAMIENTO (CLUSTERING)

Material de Lectura: Capítulo 16 del Libro Introducción a la Minería de Datos de Hernández Orallo

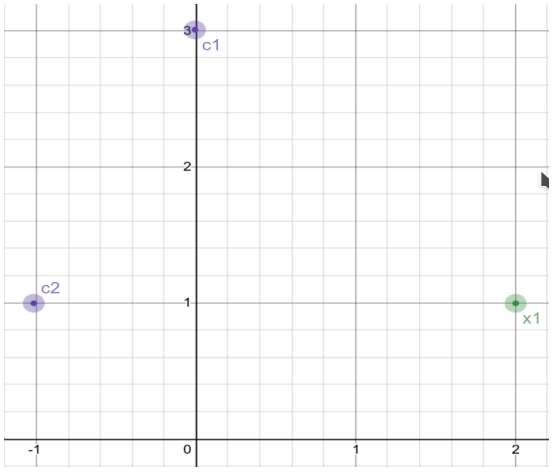
Ejercicio a entregar: 2 (todos los incisos).

*Nota: para realizar esta práctica deberá instalar el operador **Silhouette**[1] copiando el archivo CPPlugin-0.3.jar al directorio **lib/plugins** dentro del directorio de instalación de Rapidminer. Si tiene Rapidminer abierto, deberá reiniciarlo para que el operador pueda ser utilizado.*

1. Funciones de distancia y cálculo de centroides

a) Calcular la distancia euclídea entre los puntos **p1 = (2,-2,2,-4)** y **p2 = (3,0,6,-2)**. Calcular también la distancia euclídea al cuadrado.

b) Dados los centroides **c1** y **c2** de la **Figura 2**, calcular las distancias euclídea del punto **x1** a **c1** y **c2**. Luego indicar el centroide más cercano a **x1**. Calcular también la distancia euclídea al cuadrado.



c) La distancia de Manhattan entre dos puntos **A** y **B** se define como **d(A,B) = Σi |Ai-Bi|**, donde Ai y Bi son los valores de los puntos. Por ejemplo, si **A=(3,0,1)** y **B=(-2,4,7)**, entonces d(A,B)= |3-(-2)| + |0-4| + |1-7| = |5| + |-4| + |-6| = 5 + 4 + 6 = 15

Repetir los puntos **a-b** con la distancia de Manhattan.

---

[1] Tenga en cuenta que por un problema de implementación, dicho operador NO funciona si el conjunto de datos contiene atributos nominales, aún cuando han sido marcados como clase o id.

Puntos	Distancia Euclídea	Distancia Euclídea al cuadrado	Distancia de Manhattan
p1,p2			
x1,c1			
x1,c2			

d) ¿Qué relación encuentra entre la distancia Euclídea y la Euclídea al cuadrado, para las distancias entre x1 y c1 y entre x1 y c2?

e) Dados los datos x1 = (2,3), x2 = (-2,2) y x3 = (6,-14) calcular el centroide C de los mismos. Dibujar los datos y el centroide C en un plano cartesiano.

2. Algoritmo K-medias

Nota: Es obligatorio utilizar el formato de tabla indicado ejercicio para la entrega. Incluir las tablas y respuestas al principio, y al final un anexo con las cuentas de cada inciso. Utilizar una precisión de a lo sumo 2 decimales cuando sea necesario. No utilizar fracciones para expresar los resultados. Si las distancias de un ejemplo a los centroides son iguales, elegir un centroide de forma aleatoria.

a) Dados los datos x1, x2, x3 y x4, y los centroides iniciales c1 y c2 (ver gráfico) realizar tres iteraciones del algoritmo k-medias visto en la teoría para obtener un clustering con **k=2** grupos, y calcular el valor final de los centroides. Utilizar la medida de **distancia euclídea al cuadrado**. En cada iteración, graficar los puntos y centroides (recomendamos GeoGebra para graficar). Graficar también el resultado final.

	Iteración 1					Iteración 2					Iteración 3				
	c1	(3,3)	c2	(4,3)		c1		c2			c1		c2		
Puntos	Dist. a **c1**		Dist a **c2**		Centroide Asignado	Dist a **c1**		Dist a **c2**		Centroide Asignado	Dist a **c1**		Dist a **c2**		Centroide Asignado
x1 = (2,0)															
x2 = (1,2)															
x3 = (3,4)															
x4 = (4,4)															

b) Si realizara más iteraciones del algoritmo ¿cambiarían las asignaciones? ¿y los centros? ¿por qué?

c) Realice el mismo ejercicio que a), pero ahora con otros centroides iniciales. ¿Obtuvo el mismo resultado final? ¿Por qué en este caso se requieren menos iteraciones?

	Iteración 1					Iteración 2				
	c1	(1,2)	c2	(3,4)		c1		c2		
Puntos	Dist. a **c1**		Dist a **c2**		Centroide Asignado	Dist a **c1**		Dist a **c2**		Centroide Asignado
x1 = (2,0)										
x2 = (1,2)										
x3 = (3,4)										
x4 =(4,4)										

d) Realice el mismo ejercicio que a), pero ahora con un nuevo conjunto de centroides. ¿Obtuvo el mismo resultado final? Este resultado ¿es correcto?

	Iteración 1					Iteración 2				
	c1	(2,3)	c2	(2,-1)		c1		c2		
Puntos	Dist. a **c1**		Dist a **c2**		Centroide Asignado	Dist a **c1**		Dist a **c2**		Centroide Asignado
x1 = (2,0)										
x2 = (1,2)										
x3 = (3,4)										
x4 = (4,4)										

e) Indique cuáles de estas cosas pueden variar si se corre el algoritmo K-Medias con K=4 en un conjunto de datos fijo y se varían los valores de los centroides iniciales:

- Cantidad de iteraciones
- Valores de los centroides finales
- Cantidad de centroides
- Cantidad de grupos encontrados
- Cantidad de ejemplos por centroide/grupo

3. Clustering con Rapidminer. Validación del valor de k. Uso de métricas para evaluar agrupamientos.

En este ejercicio se busca utilizar dos índices distintos que pueden ayudar a elegir un buen valor de K para el agrupamiento. El archivo **2d_simple.csv** contiene 5000 filas, cada una con dos atributos numéricos, **x,y,** junto con el cluster original. Estos datos son sintéticos (generados artificialmente).

1. Cargar el conjunto de datos en RapidMiner.

2. Dado que este conjunto de datos puede visualizarse en 2D, determinar **visualmente** la cantidad de clusters utilizando algún gráfico de Rapidminer.

3. Utilizando los operadores ***Clustering (K-Means)*** y ***Performance: (Average) Silhouette***, realice un clustering de los datos y evalúe el resultado con el índice promedio *Silhouette*. El clustering debe realizarse con **k=2** y distancia **Euclídea**.

4. Repita el punto 3 variando el valor de **k** (al menos con k=2,3,..,9)**.** Haga una tabla con el valor del índice *Silhouette* para cada valor de **k**. Con esa tabla, determine el valor de **k** óptimo ¿coincide con el valor que determinó en el punto 2? Recuerde que en el índice *Silhouette* es mejor si el valor es más alto.

5. Repita el punto 4, pero ahora puntuando el resultado del clustering con *Davies-Bouldin* en lugar de *Silhouette*. Para ello utilizar el operador **Performance (Cluster Distance Performance)**, con *main criterion="Davies Bouldin",* y tildar "main criterion only", "normalize" y "maximize". Recuerde que el criterio Davies Bouldin es mejor cuando el valor es más bajo.

k	Silhouette	Davies-Bouldin

1		
2		
3		
4		
5		
6		
7		
8		
9		

6. Para el mejor valor de **k**, visualizar los puntos con un gráfico de dispersión, donde los ejes **x** e **y** están dados por los atributos del mismo nombre, y el color es el atributo **cluster** generado.

7. Repita los puntos anteriores para el conjunto de datos **2d_complex.csv.** Explicar por qué en este conjunto de datos el valor óptimo de **k** calculado con los índices *Silhouette* y *Davies Bouldin* no coincide con el óptimo determinado visualmente. ¿Qué está mal? ¿Nuestra lógica visual, el cálculo de los indices, o ninguna?

4. Características de Jugadores de Fútbol. Análisis de posiciones.

El archivo **jugadores_2023.xlsx** contiene datos de atributos físicas y habilidades de jugadores de fútbol en varias posiciones. Este conjunto de datos es complejo, y por lo tanto no es trivial entender sus atributos, valores y las relaciones entre los mismos. El atributo “Best position” indica la posición en que el jugador se desempeña usualmente. Sus valores son:

- **LB -** Defensor Izquierdo
- **CB**– Defensor Central
- **RB** – Defensor Derecho

- **CDM**– Mediocampista Defensivo Central
- **LM -** Mediocampista Izquierdo
- **CM** –Mediocampista Central
- **RM** – Mediocampista Derecho
- **CAM –**Mediocampista Atacante Central
- **LW** – Atacante Izquierdo
- **CF** –Atacante Central
- **RW** – Atacante Derecho
- **ST** –Atacante Goleador

Si bien los jugadores se agrupan tradicionalmente en base a su **posición**, queremos buscar un agrupamiento alternativo que surja naturalmente de sus características para entender mejor al conjunto de datos. Además, de este modo podríamos motivar nuevas estrategias de juego o formas de organizar tácticas en base a roles que surjan espontáneamente.

Utilice RapidMiner para agrupar los datos utilizando el algoritmo *K-medias*. Al cargar los datos, asigne el tipo *label* al atributo **posiciones** y el tipo **id** al atributo ***nombre*** para no incluirlos en el agrupamiento. Además,

a) El conjunto de datos contiene más de 10000 ejemplos (el límite de RapidMiner en la versión gratuita). Utilice el operador **Sample** para obtener una muestra aleatoria de 10000 ejemplos[2]. Tenga en cuenta que el conjunto de datos viene ordenado por el atributo **overall**, que indica el nivel de habilidad promedio del jugador. Verifique que la muestra obtenida tenga jugadores de todos los niveles.
b) Decida qué atributos dejar para el agrupamiento y cuales quitar. Por ejemplo, el atributo **Club Name** es nominal, con lo cual será difícil de usar en el agrupamiento, pero además dará poca información, ya que hay muchísimos clubes distintos, y relativamente pocos jugadores por cada club.
c) Transforme los atributos para poder aplicar K-Medias. Por ejemplo, el atributo **preferred foot (pie preferido)** es nominal, y debe ser convertido a un número para ser utilizado.
d) Tenga en cuenta los rangos de valores de los atributos. ¿Es necesario normalizarlos?

Pruebe distintos valores de **k** (2,3,4,5). En base a cada valor:

e) Analice los valores de los centroides ¿cómo podría caracterizar los grupos conseguidos? Tenga en cuenta características en común o diferencias entre ellos.
Recuerde que no hay una respuesta *correcta*, ni tampoco un valor de **k** *correcto.* Cada agrupamiento nos da una perspectiva distinta de los datos.

[2] Según la capacidad de su computadora, utilizar 10000 ejemplos puede hacer que el proceso tarde un tiempo considerable. En tal caso, sugerimos reducir la cantidad de ejemplos utilizados a 5000 o 2000.

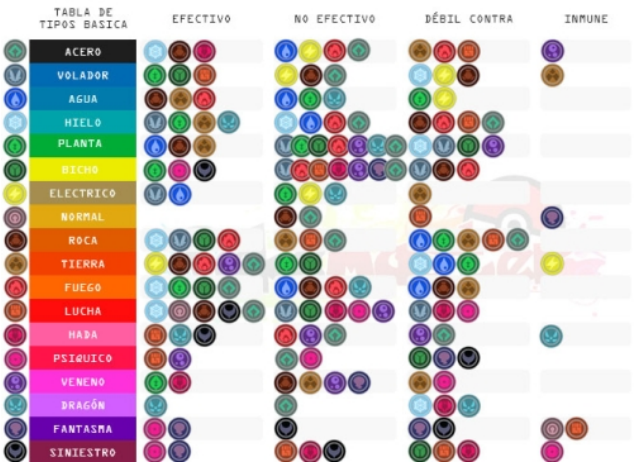
k	Descripción del agrupamiento
2	
3	
4	
5	

f) Observe el valor del atributo **posiciones** de los jugadores. ¿Cómo se compara con el agrupamiento obtenido en base a los datos? Realizar un gráfico de dispersión de la clase original (**posiciones)** vs el nro de cluster asignado para ver la cantidad de veces que se corresponden.

g) Una hipótesis razonable es que la distribución de habilidades varíe con las generaciones, es decir, que los jugadores jóvenes tengan conjuntos de habilidades distintas que los jugadores veteranos. Para verificarla, realice dos agrupamientos por separado, uno de jugadores jóvenes y otro de veteranos, y compare los grupos obtenidos en ambos casos.

h) En el inciso anterior utilizamos el agrupamiento como herramienta para explorar una hipótesis. Piense en **otra** hipótesis como la propuesta que le interesaría explorar en este conjunto de datos y que pueda verificarse mediante agrupamiento, y realice el análisis.

5. Agrupamiento de Pokemon



El archivo **pokemon_attributes.csv** contiene información de distintos Pokemon del juego del mismo nombre. Cada Pokemon tiene varios atributos básicos: puntos de vida (HP), habilidad de ataque, defensa, ataque especial y defensa especial, y velocidad. También se conoce si el Pokemon es legendario, en qué generación fue introducido (8 distintas), su nombre y su identificador. Por último, también tiene hasta 2 tipos característicos, como Agua, Tierra, Aire, etc. El orden de los tipos no es lo importante, y la mayoría de los Pokemon tiene un solo tipo. Si bien el tipo clasifica de alguna forma a los Pokemon, podemos buscar agrupamientos que también consideren otros atributos.

a) Para comenzar a entender el conjunto de datos, realice un gráfico de barras con la cantidad de Pokemon de cada tipo. Realice también una matriz de dispersión con todos los atributos numéricos para entender la relación entre los mismos y su distribución básica.

b) Decida qué atributos deberá quitar previo a realizar un agrupamiento, y cuáles deberá recodificar. Tenga en cuenta que los atributos “Type 1” y “Type 2” no indican un orden, y codifique entonces el tipo de manera acorde. Decida también si la escala de los atributos puede afectar el agrupamiento.

c) Realice un agrupamiento de los Pokemon basado en los diversos atributos que tiene, incluyendo los tipos, para encontrar grupos de Pokemon similares y entender a grandes rasgos qué tipos pueden ser similares o distintos. Elegir un valor de K de 3, para que los grupos sean fáciles de interpretar.

d) Prueba con valores de K mayores, de 3 a 8. ¿Por qué los grupos comienzan a coincidir con ciertos tipos? En base a estos agrupamientos ¿qué tipos de Pokemon consideraría similares?

6. Agrupamiento de especies de animales basada en datos.

Las clasificaciones[3] de animales en grupos o tipos se realiza en base a sus características genéticas y fenotípicas. Estas clasificaciones pueden variar según algunos criterios, muchas veces subjetivos y varía con el tiempo. Por ejemplo, la clasificación principal en vertebrados e invertebrados ya no se considera como la única, y ha sido reemplazada por una clasificación en base a otras características.

Un grupo de biológos desea explorar la posibilidad de proveer una clasificación alternativa de 101 especies basada en datos de sus características fenotípicas. El archivo **Zoo.xls** contiene información de 16 características fenotípicas de 101 animales distintos.

a) Cargue los datos y observe sus características ¿Puede visualizar estos datos? ¿Y determinar visualmente sus grupos?

b) Utilice RapidMiner para agrupar los datos utilizando el algoritmo *K-medias,* con k=3 y así generar una clasificación en 3 nuevos grupos o tipos posibles.

El dataset contiene los atributos **animal** y **clase.** El atributo **clase** corresponde al grupo asignado bajo la clasificación actual de los animales. Al cargar los datos en RapidMiner, marque como *id* el atributo **animal** y como *label* el atributo **clase** para que no sean utilizados en el clustering.

---

[3] En este caso, la palabra clasificación se utiliza en el sentido biológico, y no nos estamos refieriendo a un problema de clasificación en minería de datos.

Luego, realizar un gráfico de dispersión de la clase original (**clase)** vs el nro de cluster asignado para ver la cantidad de veces que se corresponden.

c) Obtenga el índice Davies Bouldin del agrupamiento con distintos valores de **k** (2 a 7).

k	2	3	4	5	6	7
Davies- Bouldin						

d) En base al mejor clustering encontrado en c), analizar los centroides del mismo. Cada centroide correspondería a un nuevo "grupo" según la clasificación generada. Describa en palabras las características de cada grupo en base a los valores de los atributos del centroide. Recuerde que cada centroide es un vector con 16 dimensiones, una por cada atributo.

e) Obtenga la cantidad de ejemplos por cada grupo/centroide que encuentra.

Tenga en cuenta que no habrá respuestas *correctas* para este tipo de análisis de datos; cada agrupamiento nos ofrece una perspectiva distinta de los datos, y la interpretación depende del dominio.

Cluster o Grupo	Descripción	Animales
1	Contiene aquellos animales cuyo...	
2	….	

f) Comparando el cluster asignado con el atributo "clase", que representa la clasificación tradicional y no fue tenido en cuenta al momento de agrupar. ¿La clasificación basada en datos es similar a la tradicional?