Organización de Datos 75.06. Primer Cuatrimestre de 2021. Examen por promoción

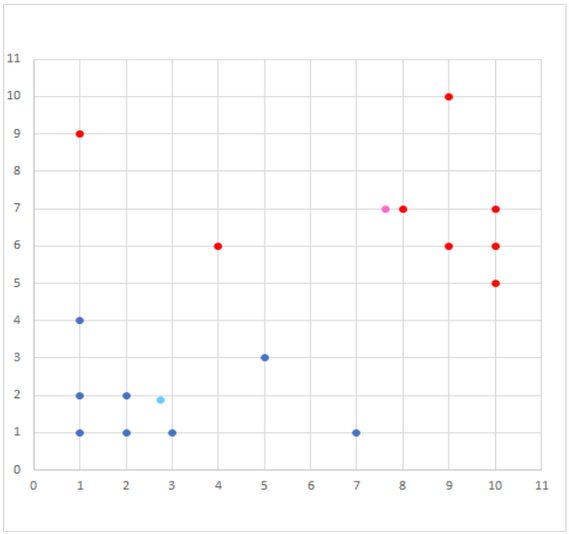
Importante: Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4). Si tiene dudas o consulta estaremos disponibles vía meet, pero tengan en cuenta que solo se contestarán dudas de enunciado, y no deben compartir por esa vía nada relacionado con la resolución. Está prohibido realizar cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen estará disponible en gradescope.

"It doesn't matter what we want. Once we get it, then we want something else ." — Lord Baelish – Game of Thrones

#	1	2	3	4	5	Entrega Hojas:
Corrección						Total:
Puntos	/20	/15	/20	/15	/30	/100

Nombre:
Padrón:
Corregido por:

1) En el marco de los JJOO Tokio 2020, se decidió analizar los datos de países tomando en cuenta dos datos: cantidad de miembros de la delegación y cantidad de medallas obtenidas. Ambas columnas fueron normalizadas en el rango (1,10). El objetivo era identificar grupos de países poderosos y débiles en los juegos. Inicialmente se usó K-Means con k=2 y se obtuvieron los resultados que vemos debajo (los puntos naranja y celeste son los centroides).

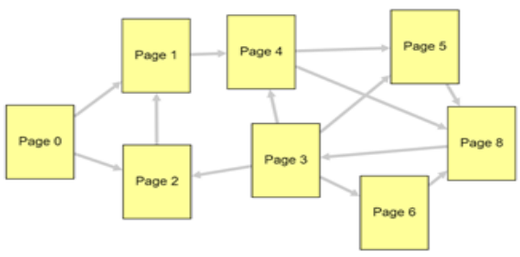


Al ver la representación observamos que el resultado de K-Means no es satisfactorio ya que algunos países quedan caracterizados como poderosos (rojos) o débiles (azul) pero en realidad no son realmente muy débiles ni muy poderosos. Le pedimos que proponga una solución a este problema, haga una clusterización alternativa, muestre la solución y explique por qué es mejor que el resultado de K-Means.

2) Dada la siguiente matriz con las calificaciones de usuarios de distintas series, utilizar collaborative filtering user-user (utilizando correlación de Pearson para obtener los 2 usuarios más similares) para estimar la calificación del ítem 3 para el usuario a:

	1	2	3	4	5	6	7	8
a	1			5	1		2	3
b	3	2		1	1		3	2
c	2		4		2		4	2
d	1	2			1			3
e		3	3		1	4		2
f	4		5	1			4	1

3) Dada las siguientes páginas y links entre ellas:



A- Determinar el valor de page rank de cada una (usando β=0,8).
B- Indicar cómo afecta al resultado detectar que la Page 0 no es una página confiable.

4) Dado el siguiente stream: {3,6,6,3,3,6,5,6} estamos usando la función h(x) = x mod 32 para aproximar el cálculo del momento de orden 0 del stream usando Flajolet Martin (con un único estimador). Para ello estamos usando los 5 bits considerando los bits a derecha para el algoritmo.

Se pide

A. Evaluar cómo está funcionando la función de hashing actual para estimar el momento de orden 0 para el stream, justificando su respuesta.
B. Plantear una función de hashing alternativa de la forma h(x) = (ax+b) mod 32 que mejore lo evaluado en 1, justificando su respuesta.
C. Plantear una función de hashing en este escenario que haga que el algoritmo Flajolet Martin no funcione.

5) Proponer datasets de train y/o test de al menos 4 puntos cada uno, y un modelo adecuado, tal que se cumpla lo pedido en cada ítem:

a) Una regresión cuyo MSE en test sea 0 pero en train sea 0,5.
b) Dos modelos de los que se conoce únicamente su probabilidad de salida Ŷ, uno cuyas predicciones tienen un accuracy 0,5 en train y el otro con accuracy 0,7 en train y que ensamblados con averaging de su probabilidad tienen un accuracy de 0,8 en train.
c) Un árbol de decisión (entrenado sobre el dataset de train) cuyas variables de entrada son el binary encoding de una variable categórica de al menos 3 clases y tiene 0.75 de accuracy en train.
d) Un KNN (entrenado sobre el dataset de train) con k=2 que clasifique mal más de la mitad de los puntos de test.

Justificar que se cumple lo pedido en cada ítem a partir de aplicar el modelo propuesto (hacer las cuentas para ver que cumple).