Organización de Datos 75.06. Primer Cuatrimestre de 2020. 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	6	Entrega Hojas:
Corrección							Total:
Puntos	/15	/15	/20	/15	/15	/20	/100

Nombre:
Padrón:
Corregido por:

1) Estamos construyendo un árbol de decisión para clasificar si un conjunto de platos es adecuado para carnívoros (clase C) o veganos (clase V). Para un cierto split el algoritmo CART debe analizar si hacer o no un split por la columna "ingrediente 2 en gramos" que es una columna numérica. Dados los valores de la columna y la clase a predecir le pedimos que indique cuál sería el valor de la ganancia de información para la columna en cuestión.

23, V

80, C

85, C

65, V

100, V

11, C

(15 pts)

2) Un sitio de ventas de autos quiere agregar una nueva funcionalidad de forma que cuando un usuario cargue su auto para la venta, el sitio le recomiende un precio de venta del mismo. Para ello quiere generar un modelo con XGBoost que al momento de cargar un nuevo aviso prediga el precio de venta.

El set de datos sobre las publicaciones de autos posee estos atributos:

(fecha, marca, modelo, versión, año, cantidad puertas, segmento/tamaño, equipamiento, kilómetros, estado, transmisión, color, tipo de combustible, motor, potencia, unico dueño, provincia, dueño directo, precio)

Indique el proceso de feature engineering que realizaría, indicando el detalle de las features que probaría para armar el módelo de predicción mediante XGBoost. Se pide que de un ejemplo de un row del set de datos final.

(15pts)

3) Dado un stream compuesto por los siguientes números: 3, 1, 4, 1, 5, 9, 2, 6, 5. Queremos aplicar el algoritmo de Flajolet Martin para calcular el momento de orden 0 del stream usando una función de hashing de la familia: $h(x) = a*x + b \bmod 32$. El resultado debe tomarse como un número de 5 bits contando la cantidad de ceros a derecha del mismo. No todos los valores de a y b son adecuados, le pedimos que nos explique qué valores de a y b son los más adecuados y por qué motivo. A modo de ejemplo analice las funciones resultantes de usar: (a=2,b=1) (a=3,b=7) y (a=4, b=0). (20 pts)

4) En este ejercicio le pedimos que demuestre que el algoritmo de K-Means que vimos en el curso para la distancia euclídea siempre termina.Como esto puede parecer un poco confuso lo ayudamos con los siguientes pasos:

a. Demuestre que el punto que minimiza la función costo para un cierto cluster es el promedio de todos los puntos en el cluster. (5 pts)
b. Demuestre que en cada paso de K-Means la función costo solo puede disminuir. (5 pts)

Solo falta un detalle más para terminar de demostrar que K-means termina y le pedimos que nos diga cuál es. (5 pts)

5) Tenemos la siguiente información parcial sobre un grafo dirigido al cual hemos aplicado PageRank hasta la convergencia:

- El PageRank de A es 0.15
- A tiene 2 links uno hacia C y otro hacia B
- C recibe solo dos links, uno viene de B
- El PageRank de D es 0.1
- B recibe un solo link
- D solo recibe un link que viene de C
- Desde C solo sale un link.

¿Cuántos links salen desde B? Explique detalladamente todas las deducciones realizadas hasta llegar a la respuesta. (15 pts)

6) Tenemos la siguiente matriz representando usuarios (A,B,C,D) y productos comprados (1 a 8).

	P1	P2	P3	P4	P5	P6	P7	P8
A		1			1	1		
B			1			1	1	1
C	1				1	1		
D		1		1			1	

Usando la semejanza de Jaccard y collaborative filtering user-user con k (cantidad de vecinos) = 2. indique cuáles serían los 3 productos que le recomendaría al usuario "A" y en qué orden. En caso de que haya empates use collaborative-filtering item-item para desempatar. (20 pts)

Organización de Datos 75.06. Primer Cuatrimestre de 2019. Examen parcial, primera oportunidad: Resolución

Resolutividad:

Ejercicio 1: hay que ordenar por el valor de la columna por el cual se hace el split con lo cual nos queda algo tipo C,V,V,C,C,V. Ahora hay que ver cual es el split más conveniente siendo las opciones

C vs VVCCV

CV vs VCCV

CVV vs CCV

CVVC vs CV

CVVCC vs V

Con esos splits hay que calcular las entropias y ver cual es el mejor split.

Ejercicio 2: Features….

Ejercicio 3: 3, 1, 4, 1, 5, 9, 2, 6, 5.

para 2x + 1 : 7, 3, 9, 3, 11, 19, 5, 13, 11 (son todos valores impares asi que nunca hay 0 al final! por lo que si a es par no es bueno que b sea 1)

para 3x + 7: 16, …. etc (esta es decente)

para 4x: 12,4,16,4,20,4,8,24,20 (son todos pares, no es bueno que b sea 0 si a es par!)

De 1 y 3 resulta que si a es par la funcion devuelve todos pares o impares… Por lo que a debe ser impar.

ejercicio 4: para el punto a alcanza con plantear la funcion costo de k-meas para 1 cluster, derivar e igualar a 0 y queda que el resultado es el promedio de todos los puntos del cluster. para el punto b hay que observar que un punto solo se reasigna de cluster si el centroide esta más cerca que el anterior por lo que la funcion costo solo puede decrecer. Finalmente el punto que falta es explicar que la cantidad de posibles clusterizaciones es un numero finito.

Ejercicio 5 (sacado de la revista Cacumen, junio 1995). Si el PR de A es 0.15 y salen dos links entonces cada link lleva 0.075, A D solo llega un link de C y el PR de D es 0.1 por lo que el PR de C debe ser 0.1 y como a C llegan dos links uno de A que es 0.075 y otro de B sabemos que ese otro debe ser 0.025. A su vez B solo recibe un link con 0.075 y una de sus salidas es 0.025 por lo que deducimos que de B tienen que salir 3 links.

Ejercicio 6 .

El usuario más similar a A es C con Jaccard ½ y luego viene D con Jaccard ⅛ por lo que la primera recomendacion es el item 1 que fue comprado por C y no por A, luego vendrian los items 4 y 7 que fueron comprados por D pero no por A. Como hay un empate aplicando item-item vemos que para los ítem que compro A (2,5,6) los items más similares son el 4 (0.5 contra el 2) y el 7 (0.33 contra el 2) por lo tanto desempatamos entre 4 y 7 a favor del item 4. Si se tomara en promedio en vez de tomar el mas similar quedaría el 7 y se considera bien también.