7506-2020-2 Examen por promocion

Santiago Locatelli

TOTAL POINTS

62 / 100

QUESTION 1

1 Clustering 5 / 20

✓ **- 5 pts** b) mal.

✓ **- 10 pts** c) mal.

QUESTION 2

2 Recomendaciones 20 / 20

✓ **- 0 pts** Correct

QUESTION 3

3 PageRank 8 / 20

✓ **- 12 pts** Plantea pero no resuelve

QUESTION 4

4 Streaming 20 / 20

✓ **- 0 pts** Correct

QUESTION 5

5 Feature Engineering 9 / 20

Punto a

✓ **- 2 pts** Elimina datos importantes, como la fecha de publicación o la ciudad.

✓ **- 3 pts** Usa los features que vienen pero no considera agregar nuevos features.

✓ **- 2 pts** No procesa el campo Servicios.

✓ **- 1 pts** No aprovecha que el estado de la propiedad tiene un orden implicito para codificarlo con label encoding.

✓ **- 2 pts** Utiliza binary encoding para la ciudad sin evaluar codificarlo con mean encoding.

Punto b

✓ **- 1 pts** Falta detalle de qué campos deberia normalizar y cómo.

1. Es la fecha de publicación
2. Es la fecha de publicación. Ud cree que el precio es el mismo si la propiedad se vendió el año pasado o hace 20 años?
3. Por?
4. Por?
5. Por qué binary encoding y no mean encoding?
6. Por qué?
7. El estado quizas sea mejor encodearlo con label encoding ya que hay un orden implicito en los datos.

Iago Locatelli 109107

a) Sabiendo que el algoritmo de Lloyd agarra 3(k) puntos del set de datos al azar y que hay 6 puntos en el set datos, podemos calcular las configuraciones posibles

La primera vez que se elija pueden ser 6 opciones.
La segunda vez que se elija pueden ser 5 opciones
" Tercera " " " " " 4 opciones

Por lo que por regla de la cadena tenemos que las configuraciones posibles son: $6 \cdot 5 \cdot 4 = \boxed{120}$

b) Solo 1. Analizando los puntos dados, vemos que hay 3 pares de puntos, los cuales estos pares tienen distancia mínima entre ellos. Estos pares son:

Par 1: x_1 y x_4 Par 2: x_2 y x_5 Par 3: x_3 y x_6

El hecho de que tome de a pares se debe a la proporción entre la cantidad de puntos y la cantidad clusters.

Entonces, tenemos como conclusión que los clusters van a converger al promedio de cada par.

Calculamos los puntos.

- $(0, 3)$
- $\left(\frac{8+8}{2}, \frac{6+0}{2}\right) = (8; 3)$
- $\left(\frac{16+16}{2}, \frac{0+6}{2}\right) = (16, 3)$

1 Clustering 5 / 20

✓ **- 5 pts** b) mal.

✓ **- 10 pts** c) mal.

Santiago Locatelli 104107.

2.

Estimar la probabilidad de que al usuario 6 le guste la serie 'D'.

Lo primero que hacemos es buscar los 3 vecinos más cercanos al usuario 6.

Calculamos las semejanzas de Jaccard:

$$J(6,1) = \frac{0}{6} = 0 \qquad J(6,2) = \frac{2}{6} = \frac{1}{3}$$

$$J(6,3) = \frac{2}{6} = \frac{1}{3} \qquad J(6,4) = \frac{2}{6} = \frac{1}{3}$$

$$J(6,5) = \frac{1}{5}$$

Por lo tanto los 3 usuarios más semejantes a 6 son los usuarios: 2, 3, 4

Como solamente se trata de upvotes y down votes, la probabilidad de que le guste la serie está definida como el porcentaje de upvotes sobre el total, es decir, que en este caso la probabilidad de que le guste la serie 'D' al usuario 6 es $\frac{1}{3}$.

Observación: si se usara la formula de ponderación, para obtener la probabilidad, usando como 'Si' = 1 y 'No' = 0 da el mismo resultado.

2 Recomendaciones 20 / 20

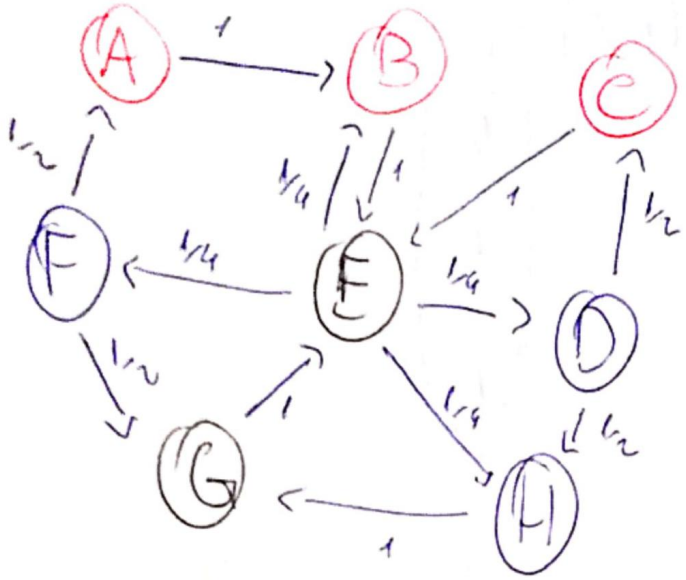
✓ **- 0 pts** **Correct**

Santiago Locatelli 104107

3. [redacted]

Teniendo el grafo:

Asumo un $\beta = 0.8$.



Podemos armar la matriz: A

	A	B	C	D	E	F	G	H
A	0	0	0	0	0	1/2	0	0
B	1	0	0	0	1/4	0	0	0
C	0	0	0	1/2	0	0	0	0
D	0	0	0	0	1/4	0	0	0
E	0	1	1	0	0	0	1	0
F	0	0	0	0	1/4	0	0	0
G	0	0	0	0	0	1/2	0	1
H	0	0	0	1/2	1/4	0	0	0

[redacted]

Entonces para calcular el PageRank se debe calcular:

$$PageRank = \beta \cdot A \circ \begin{pmatrix} 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \\ 1/8 \end{pmatrix} + (1-\beta) \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/3 \\ 0 \\ 1/3 \\ 0 \\ 1/3 \end{pmatrix}$$

No Resuelvo la cuenta por falta de tiempo.

3 PageRank 8 / 20

✓ **- 12 pts** **Plantea pero no resuelve**

Santiago Locatelli 104107

5

4\. a) Como por cada dato del stream se le aplica la función de hashing y se suma 1 en el número de bucket que dio como resultado, la suma de todos los valores en los filtros debe ser la misma.

Notando que $F_1 = F_2 = 20$, tenemos que pedir que A o B sea igual a 20.

$A = 20$ y $B = 21$, por lo tanto el filtro correcto es el A.

b) Los filtros nos quedan:

$F_1 = [4, 0, 8, 0, 0, 0, 5, 3)$

$F_2 = [0, 0, 3, 3, 4, 4, 6, 0]$

$F_3 = [5, 0, 1, 3, 2, 4, 4, 1]$
0 1 2 3 4 5 6 7

Para lograr la mayor estimación posible, los resultados de las funciones de hash deberían caer sobre los buckets que almacenan los valores más grandes.

Para lograr lo mencionado los valores de las funciones de hash deberían ser:

$h_1 = 2$ $h_2 = 6$ $h_3 = 0$

Entonces la estimación es $min(\{8, 6, 5) = 5.$

4 Streaming 20 / 20

✓ - 0 pts Correct

Santiago Locatelli 104104

6

5.

a) Asumo que el Feature "Fecha" es la fecha actual. (1)

- Crearía una nueva Feature que represente el año en que la Propiedad se terminó (se estrenó), a través de la diferencia de la fecha con la antigüedad.

- Crearía también una Feature que represente una relación entre los m^2 totales y los m^2 cubiertos. De la forma:

$$\frac{m^2 \text{ cubiertos}}{m^2 \text{ totales}}$$

Notar que sería un número menor o igual a 1. Y no preocupa que m^2 totales sea cero, porque sino la Propiedad no tendría sentido.

- Haría la concatenación de las Features categóricas: Provincia y Ciudad, ya que están altamente relacionadas.

Las Features que no tendría en cuenta son:

- Población_ciudad. (3)
- Cantidad_escuelas_cercanas.
- m^2_totales
- m^2_cubiertos
- Servicios.

(Es decir, las descartaría después de la creación de la Feature ya mencionada.) (4)

- Fecha (2)

También, es muy importante la etapa de encoding de las Features que son categóricas.

Para nueva Feature Provincia_ciudad utilizaría Binary Encoding, ya que es muy probable que haya muchos valores diferentes. (5)

Para las Features Tipo_Propiedad y estado_Propiedad aplicaría One Hot Encoding. (6) (7)

Ejemplo:

Año_estreno: 2010
Relación_mz_cub_total: 0.5.
Habitaciones: 8
Dormitorios: 4
Baños: 2
Cocheras: 1
Provincia_ciudad_1: 0
Provincia_ciudad_2: 0
Provincia_ciudad_3: 1
⋮
Provincia_ciudad_Log2(cantidad_valores): 0
Estado_malo: 0
Estado_regular: 0
Estado_bueno: 1
Antiguedad: 11

b) Como en modelos basados en árboles las Features numéricas no afectan, pero en las redes neuronales sí afectan drásticamente, realizó una normalización de las features numéricas. Más que nada a las features Año_estreno y antiguedad, que son las que más diferencia entre valores puede haber.

5 Feature Engineering 9 / 20

Punto a

✓ **- 2 pts Elimina datos importantes, como la fecha de publicación o la ciudad.**

✓ **- 3 pts Usa los features que vienen pero no considera agregar nuevos features.**

✓ **- 2 pts No procesa el campo Servicios.**

✓ **- 1 pts No aprovecha que el estado de la propiedad tiene un orden implicito para codificarlo con label encoding.**

✓ **- 2 pts Utiliza binary encoding para la ciudad sin evaluar codificarlo con mean encoding.**

Punto b

✓ **- 1 pts Falta detalle de qué campos deberia normalizar y cómo.**

1. Es la fecha de publicación
2. Es la fecha de publicación. Ud cree que el precio es el mismo si la propiedad se vendió el año pasado o hace 20 años?
3. Por?
4. Por?
5. Por qué binary encoding y no mean encoding?
6. Por qué?
7. El estado quizas sea mejor encodearlo con label encoding ya que hay un orden implicito en los datos.