Comenzado el	martes, 20 de febrero de 2024, 17:05
Estado	Finalizado
Finalizado en	martes, 20 de febrero de 2024, 17:30
Tiempo empleado	24 minutos 47 segundos
Calificación	**8,25** de 10,00 (**82,5**%)

Pregunta 1

Finalizado Puntúa 0,50 sobre 1,00

¿Cuál es el rango de valores posibles para el coeficiente de silueta (Silhoutte score)?¿Qué valor es el óptimo?¿Para qué tipo de algoritmo?

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

El silhouette score es una medida que se utiliza para determinar el número de clusters K apropiado para efectuar clustering mediante K-Means.

Comentario:

Pregunta 2

Finalizado Puntúa 1,00 sobre 1,00

Dada la siguiente línea de código:

```
Z = scipy.cluster.hierarchy.linkage(cityDist, 'complete')
```

¿Qué está almacenado en Z[i, 2]?

Z tiene almacenado un sistema de clustering, el cual usa la metodología de distancia completa.
Esto implica que a la hora de medir distancias entre dos clusters, se escogen los puntos más alejados entre estos dos.
En Z[i,2] esta almacenada esta distancia, para el cluster i.

Comentario:

Pregunta 3

Finalizado Puntúa 1,00 sobre 1,00

Menciona un método relevante en el procesamiento del lenguaje natural (NLP) para reducir la dimensionalidad que pueda considerarse como alternativa al PCA.

Para NLP, un método relevante es el Embedding, el cual convierte las palabras en vectores unitarios. Es capaz de reducir la dimensionalidad en base a la similitud de las palabras.

Comentario:

Pregunta 4

Finalizado Puntúa 0,75 sobre 1,00

En la siguiente línea de código, ¿qué métrica debería seleccionarse para este objeto?

```
AgglomerativeClustering(linkage='ward', metric=_____)
```

euclidean distance

Comentario:

Pregunta 5

Finalizado Puntúa 1,00 sobre 1,00

Si disponemos de 750 imágenes de 8x8 píxeles y deseamos aplicar PCA para reducir a 5 dimensiones, ¿que dimensión tendrá la matriz de loadings (W)?

X (N*D) * W (D*L) = Z (N*L)

Comentario:

Pregunta 6

Finalizado Puntúa 1,00 sobre 1,00

Identifica una variable fuertemente asociada a MntWines y otra que muestre escasa o ninguna asociación con MntWines.

feature_names	PC1	PC2
ID	-0.005422	-0.006155
Year_Birth	-0.090931	-0.625856
Income	0.410626	0.134237
Kidhome	-0.363594	-0.225476
Teenhome	-0.086597	0.661024
Recency	0.003690	0.032621
MntWines	0.405831	0.153844
MntFruits	0.394024	-0.166981
MntMeatProducts	0.447731	-0.146189
MntFishProducts	0.403682	-0.168723

Variable fuertemente asociada a MntWines: Income
Variable con escasa asociación a MntWines: Kidhome

Comentario:

Pregunta 7

Finalizado Puntúa 1,00 sobre 1,00

Dada la función de reconstrucción de PCA, ¿por qué queremos maximizar o elegir los autovalores más altos y qué hipótesis acerca de la media muestral hemos tenido que asumir?

PCA es un proceso estadistico el cual se basa en reducir el número de dimensiones de un dataset mientras procura conservar la maxima variancia del dataset (ya que nos otorga la mayor información). Ya que los datos pueden estar medidos a distinta escala, implicando una varianza distinta por columna, se deben de normalizar (y centrar) los datos para que cada variable aporte información sin sesgo de escala. Es por ello, que la media debe de ser 0.

Comentario:

Pregunta 8

Finalizado Puntúa 0,00 sobre 1,00

En el algoritmo de clustering aglomerativo (HAC), ¿qué objetivo principal tiene el enlace Ward (linkage='ward')?

El enlace Ward se utiliza ya que no es correcto asumir que todos los datos tienen una distribución Gaussiana, y forma esferica. Los enlaces Single y Complete en cambio, producen resultados muy malos cuando un dataset no sigue las pautas definidas previamente.

Comentario:

Pregunta 9

Finalizado Puntúa 1,00 sobre 1,00

Explica con tus palabras el significado del segundo argumento utilizado en el objeto de sklearn mostrado en el siguiente código de ejemplo:

```
KMeans(n_clusters=k, init='random')
```

Este código genera un módelo de KMeans, el cual por el segundo argumento, va a definir los centroides k en puntos aleatorios en el espacio, al principio del algoritmo.

Comentario:

Pregunta 10

Finalizado Puntúa 1,00 sobre 1,00

¿Cuál es el propósito de calcular un perfil de log-likelihood en PCA? ¿Nos interesa el mínimo o el máximo en la gráfica?"

El propósito de calcular un perfil de log-likelihood es indicar cuando un incremento de varianza ya no es significativo a la hora de añadir . Nos interesa obtener el máximo en la gráfica.

Comentario:

◀ Project I instructions

Ir a...