Pregunta /Posible respuesta

¿Cuál de las siguientes es una forma razonable de seleccionar el número de componentes (k) principales de un conjunto de datos con n dimensiones?

- (●) Se selecciona el menor valor de k de modo que las componentes principales contengan por lo menos el 99% de la varianza de los datos
- ( ) Se selecciona k = n/2
- ( ) Se utiliza el WoE
- ( ) Se seleccionan los tres primeros

En base a:

La varianza es la medida de la dispersión de un conjunto de una variable:

Pregunta /Posible respuesta

¿Qué es un auto-codificador?

- (●) Una red neuronal
- ( ) Un algoritmo de clustering
- ( ) Un algoritmo de clasificación
- ( ) Un sistema de recomendación

En base a:

Los auto-codificadores ("autoencoders") son redes neuronales implementadas con tres capas (sólo una capa oculta) que aprende a producir a la salida exactamente la misma información que recibe a la entrada.

Pregunta /Posible respuesta

Al implementar un modelo de clasificación para la detección de fraude se ha observado que se ha identificado correctamente 85 casos de fraude y 890 no fraudulentos. Por otro lado, el modelo marca como fraude 10 (falsos positivos) casos que no lo son y deja escapar 15 (falsos negativos) que si lo son. ¿Cuál es la exactitud de este modelo?

85/ (85+10) = 0,89

En base a:

- Exactitud (Precision): $$P = \frac{TP}{TP + FP}$$

Pregunta /Posible respuesta

¿En cuál de las siguientes aplicaciones se podría utilizar una algoritmo de clustering, por ejemplo k-means?

- [x] A partir de los patrones de uso en un sito web identificar los diferentes grupos de usuarios que existen.
- [ ] A partir de las ventas de un almacén en un periodo de rebajas estimar las ventas en el siguiente.
- [ ] A partir del histórico de navegación en una tienda electrónica estimar la probabilidad de que un nuevo visitante realice una compra.
- [x] Teniendo en cuenta los datos de ventas de un almacén averiguar que productos forman grupos (por ejemplo se compran con frecuencia juntos) y por lo tanto se deben poner juntos.

En base a:

En el análisis de clustering se busca **agrupar** las observaciones de un conjunto de datos de tal manera que **los miembros de un mismo cluster son más similares entre sí de lo que son los miembros de los otros grupos**.

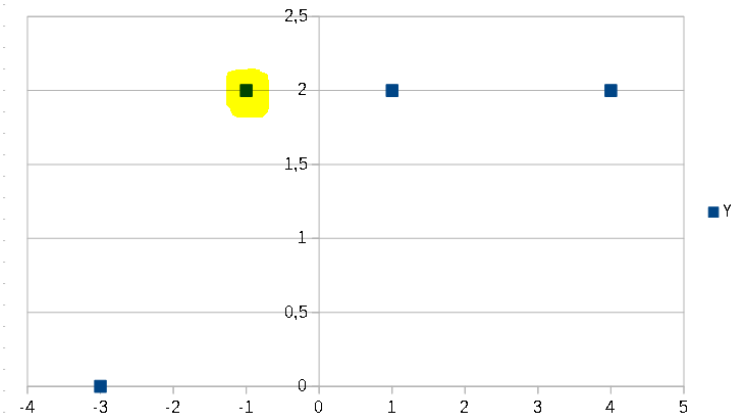
La similitud entre dos registros de datos se calcula utilizando una métrica.

Pregunta /Posible respuesta

Se han estimado los siguientes centroides con el método de k-means: c1 = [1; 2], c2 = [-3; 0] y c3 = [4; 2]. Suponiendo que se utiliza la distancia Euclídea, ¿A que clúster pertenece el punto [-1; 2]?

- (●) C1
- ( ) C2
- ( ) C3
- ( ) Ninguno

En base a:



Pregunta /Posible respuesta

¿Qué se mide con el VIF?

- (●) Multicolinealidad
- ( ) Colinealidad
- ( ) La capacidad de predecir de un modelo
- ( ) La velocidad de convergencia de un modelo

En base a:

El factor de inflación de la varianza (variance inflation factor, VIF) cuantifica la multicolinealidad en un análisis de regresión.

Pregunta /Posible respuesta

¿Qué se puede medir con el WoE (Weight of Evidence)?

- (●) La capacidad de predecir de una agrupación o nivel de una variable
- ( ) La capacidad de predecir de un modelo
- ( ) El peso de una variable utilizada en un modelo
- ( ) La existencia de sobreajuste en un modelo

En base a:

El peso de la evidencia (WoE, Weight of Evidence) es un valor que indica la capacidad predictiva de cada una de los niveles de una variable:

$$WoE_i = \ln \left| \frac{R_i(T)}{R_i(F)} \right|$$

Pregunta /Posible respuesta

¿Cuáles de los siguientes métodos de aprendizaje automático son supervisados?

- [ ] K-means
- [ ] Gaussian mixtures
- [x] Arboles de decisión
- [x] Regresión logística

En base a:

Las redes neuronales es un paradigma de aprendizaje automático supervisado inspirado en el funcionamiento del sistema nervioso.

Pregunta /Posible respuesta

En un gran superficie se han observado que 100 clientes han comprado leche, 80 han comprado pan y 30 han comprado mantequilla de un total de 200 clientes. De los 80 clientes que han comprado pan 40 han comprado también leche y 10 han comprado leche, pan y mantequilla. ¿Cuál es el soporte de conjunto Leche y Pan?

40/200= 0,2 --> 20%

En base a:

- Soporte (Support): es el porcentaje de transacciones en las que aparece X. El soporte de una regla de asociación es el porcentaje de transacciones que contiene X e Y

$$supp(X \Rightarrow Y) = supp(X \cup Y)$$

Pregunta /Posible respuesta

En un gran superficie se han observado que 100 clientes han comprado leche, 80 han comprado pan y 30 han comprado mantequilla de un total de 200 clientes. De los 80 clientes que han comprado pan 40 han comprado también leche y 10 han comprado leche, pan y mantequilla. ¿Cuál es la confianza de la regla {Leche, Pan} -> {Mantequilla}?

(10/200) / (40/ 200) = 0,25

En base a:

- Confianza (Confidence): es la fracción de las transacciones en las que aparece X y también aparece Y

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Pregunta /Posible respuesta

Al realizar un modelo de clasificación con regresión logística ¿Cuál de las siguientes afirmaciones son ciertas?

- [ ] Al añadir nuevas variables al conjunto de datos siempre se consigue un modelo mejor en el conjunto de entrenamiento y validación.
- [x] Al añadir muchas variables nuevas es más probable que aparezca sobreajuste (overfit) en el conjunto de entrenamiento.
- [ ] Se selecciona el menor valor de k de modo que las componentes principales contengan por lo menos el 99% de la varianza de los datos

En base a:

El **sobreaprendizaje** o **sobreajuste** (overfitting) aparece cuando el algoritmo de aprendizaje memoriza el ruido existente en los datos con los que se esta creando el modelo.

Pregunta /Posible respuesta

Al implementar un modelo de clasificación para la detección de fraude se ha observado que se ha identificado correctamente 85 casos de fraude y 890 no fraudulentos. Por otro lado, el modelo marca como fraude 10 (falsos positivos) casos que no lo son y deja escapar 15 (falsos negativos) que si lo son. ¿Cuál es la precisión de este modelo?

(85+890) / (85+10+890+15) = 0,975

En base a:

- Precisión (Accuracy): $$A = \frac{TP + TN}{TP + FP + TN + FN}$$