

Clustering

K-means

Luis Norberto Zúñiga Morales

15 de marzo de 2022

¿Qué es el aprendizaje no supervisado? Hagan un dibujo explicándolo.

Aprendizaje Supervisado vs Aprendizaje No Supervisado

Aprendizaje Supervisado

- $f : X \rightarrow Y$
- Conjunto de datos $X = \{(x_i, y_i)\}$
- Algoritmo A tal que $\hat{f} : X \rightarrow \hat{Y}$
- Funciones de error o pérdida $L(y, \hat{y})$

Aprendizaje Supervisado vs Aprendizaje No Supervisado

Si suponemos que (X, Y) son variables aleatorias representadas por una probabilidad conjunta $P(X, Y)$...

Aprendizaje Supervisado vs Aprendizaje No Supervisado

Si suponemos que (X, Y) son variables aleatorias representadas por una probabilidad conjunta $P(X, Y)$...

El Aprendizaje Supervisado se convierte como un problema de estimación:

$$P(X|Y)$$

Y esto nos permite encontrar los mejores parámetros μ que minimizan el error esperado de \mathbf{x}_i :

$$\mu(\mathbf{x}) = \operatorname{argmin}_{\theta} E_{Y|X} L(Y, \theta)$$

El clustering (agrupamiento) es un tipo de aprendizaje no supervisado:

- La tarea es agrupar objetos que compartan cierta similitud.

El clustering (agrupamiento) es un tipo de aprendizaje no supervisado:

- La tarea es agrupar objetos que compartan cierta similitud.
- Los algoritmos de clustering descubren la información natural de los datos.

El clustering (agrupamiento) es un tipo de aprendizaje no supervisado:

- La tarea es agrupar objetos que compartan cierta similitud.
- Los algoritmos de clustering descubren la información natural de los datos.
- Es útil cuando no se necesita o no se sabe la clase de los datos que se manejan.

El clustering (agrupamiento) es un tipo de aprendizaje no supervisado:

- La tarea es agrupar objetos que compartan cierta similitud.
- Los algoritmos de clustering descubren la información natural de los datos.
- Es útil cuando no se necesita o no se sabe la clase de los datos que se manejan.
- Es una herramienta útil para explorar los datos que se manejan.

Idea General:

- El algoritmo KMeans agrupa los datos tratando de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo.
- Este algoritmo requiere que se especifique el número de clústeres.
- K-means divide un conjunto de muestras N en grupos separados K , cada uno descrito por una media μ_j .

- El algoritmo K-means tiene como objetivo elegir centroides que minimicen la inercia, o el criterio de suma de cuadrados:

$$\sum_{i=0}^n \operatorname{argmin}_{\mu_j \in K} ||x_i - \mu_j||^2$$

K-means

Bibliografía