Machine Learning

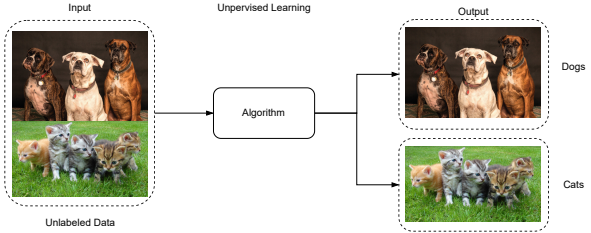
Agrupamiento

Jose Luis Paniagua Jaramillo
jlpaniagua@uao.edu.co

Agenda

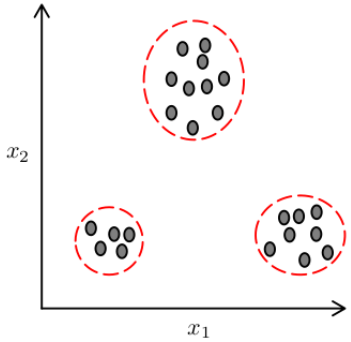
Agenda

Aprendizaje Supervisado vs Aprendizaje No Supervisado



Agenda

Agrupamiento

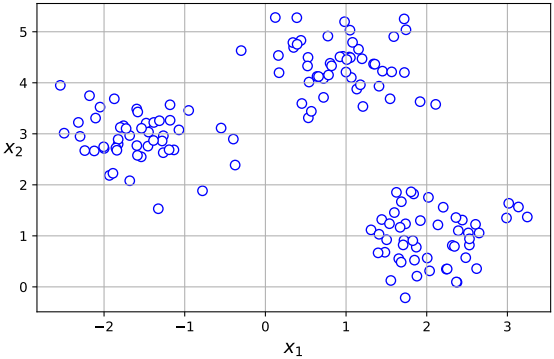


- Es una tecnica de **aprendizaje no supervisado**.
- El objetivo es agrupar datos con características similares en grupos (**clusters**).

Agenda

Agenda

k-means I



cuantos grupos hay?

k-means II

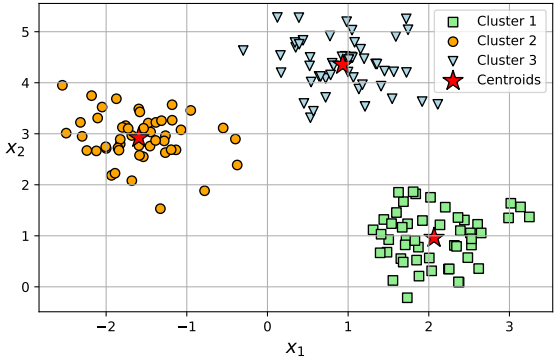
La función del algoritmo k-means es encontrar el centro (**centroide**) de cada grupo y asignar cada dato al grupo mas cercano.

k-means

El algoritmo k-means se puede resumir en cuatro pasos:

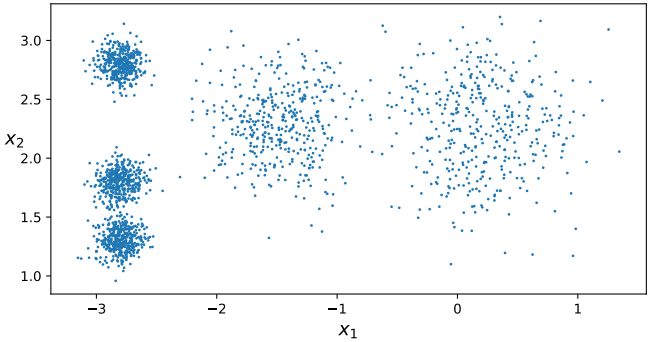
1. Elejir e inicializar aleatoriamente el numero de centroides k.
2. Asignar cada dato (**sample**) al grupo mas cercano.
3. Mover los centroides al centro de los datos que fueron asignados a este.
4. Repetir los pasos 2 y 3 hasta que:
   - Las asignaciones de los datos a los grupos no cambie.
   - Se cumpla una tolerancia definida por el usuario.
   - Se cumpla el numero total de iteraciones.

k-means III



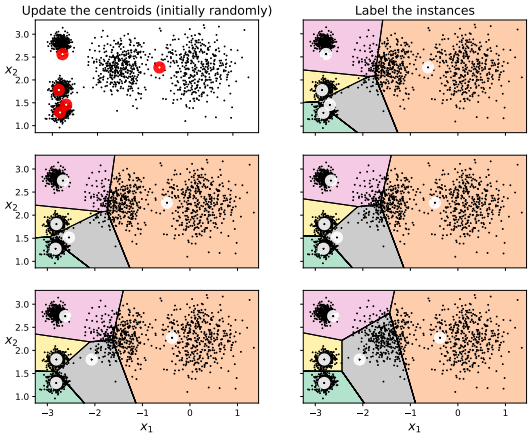
k-means IV

Cuantos grupos hay?



k-means V

Que se puede concluir acerca de los datos ubicados cerca de las fronteras de decisión?



k-means VI

Resumen

- Debido a la inicializacion aleatoria de los centroides, estos no siempre terminan en el centro del cluster.
- Para tratar de encontrar la solución optima, se utilizan estrategias como ejecutar el algoritmo varias veces y seleccionar el mejor resultado. Otra opción es inicializar de forma manual los centroides.
- Especificar el numero de clusters a priori es una de las limitantes de este algoritmo.
- No se obtienen buenos resultados cuando los datos estan dispersos. Por esta razon se deben normalizar antes de entrenar el algoritmo.

Metricas De Desempeño

Inercia del Modelo

Sum of Squared Distance

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w^{(i,j)} \left\| x^{(i)} - \mu^{(j)} \right\|_2^2$$

k-means++ I

Cual es el problema?

- El numero de clusters que se seleccionan no siempre es el correcto en problemas reales.
- Se asume que los clusters no se solapan.
- Se asume que al menos existe un dato en cada cluster.

k-means++ II

Solucion: k-means++

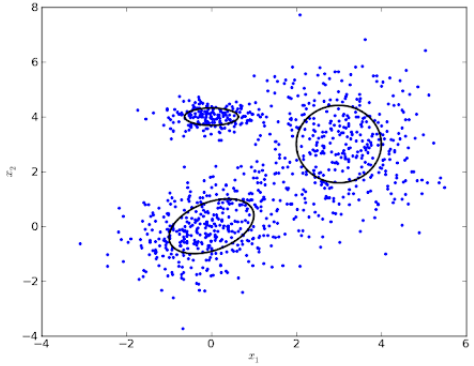
- Selecciona los centroides de tal manera que queden uno lejos del otro.

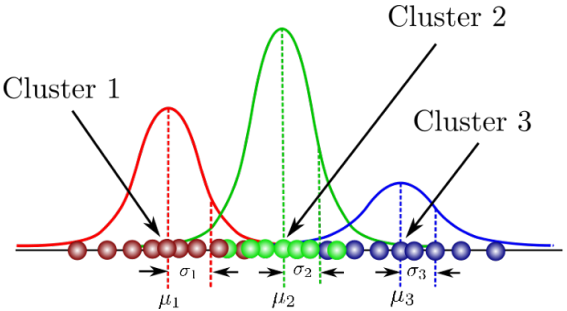
Pasos para la inicializacion en k-means++

1. Seleccionar un centroide $c^{(1)}$ de forma aleatoria.
2. Calcular la distancia mínima entre cada **sample** $x^{(i)}$ y el centroide $c^{(1)}$.
3. Seleccionar el sample $x^{(i)}$ mas alejado del centroide ya selecciondo como el nuevo centroide $c^{(i)}$
4. Repetir pasos 2 y 3 hasta que se hayan seleccionado todos los k centroides.
5. Continuar con el algoritmo k-means clásico.

Agenda

Gaussian Mixture Models I





Gaussian Mixture Models II

GMM

- Es un modelo probabilistico donde se asume que los datos fueron generados a partir de un mezcla de varias distribuciones gausianas cuyos parámetros no son conocidos.
- Todos los datos generados a partir de una distribución gausiana forman un **cluster** que normalmente tiene forma de **elipsoide**.
- Cada cluster puede tener diferente forma (**elipsoide**), tamaño, densidad y orientación.

Gaussian Mixture Models III

Como funciona el algoritmo?

1. Se deben establecer el numero k de distribuciones gausianas.
2. Para cada dato (**instancia**), un cluster es seleccionado de forma aleatoria de los k clusters definidos en el paso 1. La probabilidad de seleccionar uno de los clusters (j^{th}) en particular esta definida por el peso del cluster $\phi^{(j)}$.
3. Una vez la i^{th} instancia ha sido asignada al cluster j^{th} la localización de esta instancia es tomada aleatoriamente de una distribucion gausiana con media $\boldsymbol{\mu}^{(j)}$ y matriz de covarianza $\boldsymbol{\Sigma}^{(j)}$.

Gaussian Mixture Models IV

Que hace el algoritmo?

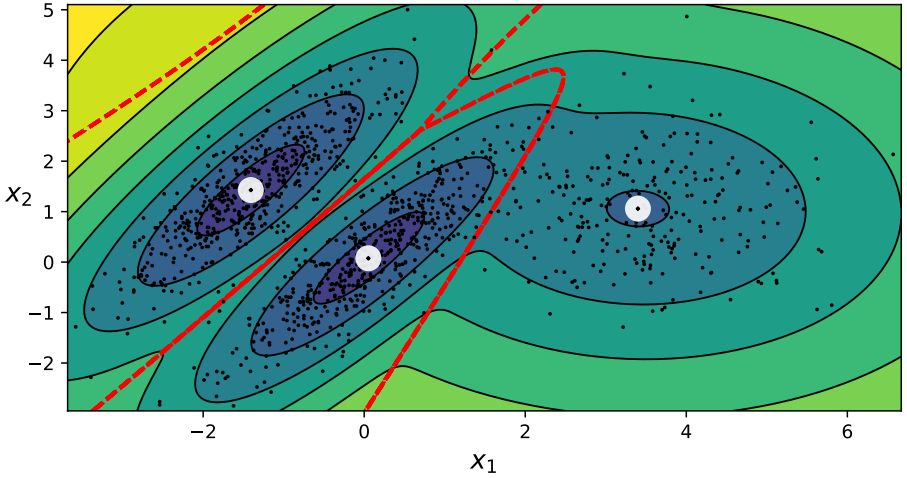
Dado un dataset X, el algoritmo estima los pesos ϕ y todos los parametos de las distribuciones de probabilidad $\mu^{(1)}$ a $\mu^{(k)}$ y $\Sigma^{(1)}$ a $\Sigma^{(k)}$

Expectation-Maximization (EM)

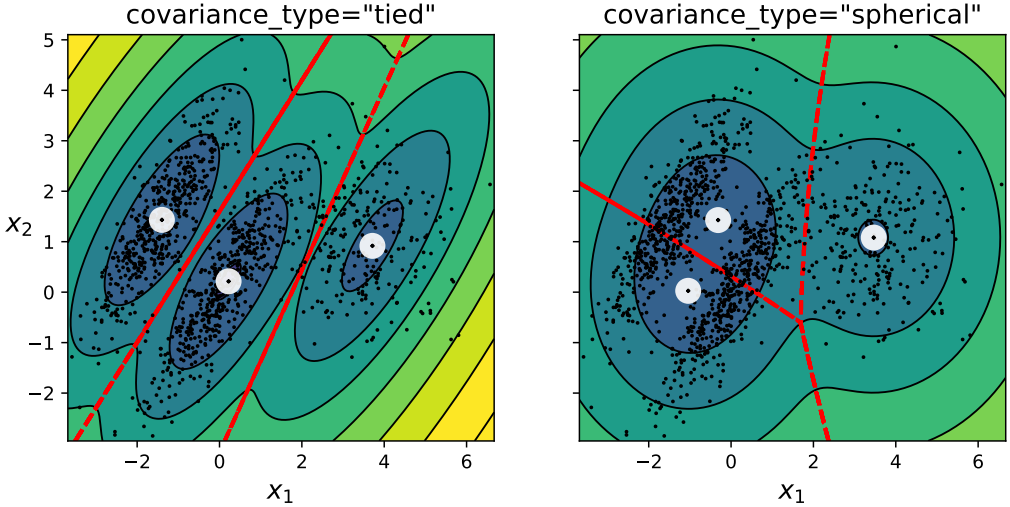
- el algoritmo **EM** es muy similar a **k-means**.
  1. inicializa los parametros de los clusters de forma aleatoria.
  2. Repite 2 pasos hasta que converja:
     1. primero asigna datos (**instancias**) a los clusters (**expectation step**).
     2. luego actualiza los clusters (**maximization step**)

EM es una generalización de **k-means**, ya que no solo encuentra los centroides ($\mu^{(1)}$ a $\mu^{(k)}$) si no que ademas encuentra el tamaño, la forma y orientación ($\Sigma^{(1)}$ a $\Sigma^{(k)}$) y su peso relativo ($\phi^{(1)}$ a $\phi^{(k)}$)

Gaussian Mixture Models V



Gaussian Mixture Models VI



Agenda

Referencias

Aurélien Géron.
Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.
O'Reilly Media, 2019.

`https://scikit-learn.org/stable/index.html`