DIPLOMADO
Ciencia de Datos

Módulo 7

Aprendizaje de máquina no supervisado

1. Clustering, modelos de mezclas gaussianas

Eduardo Espinosa Avila

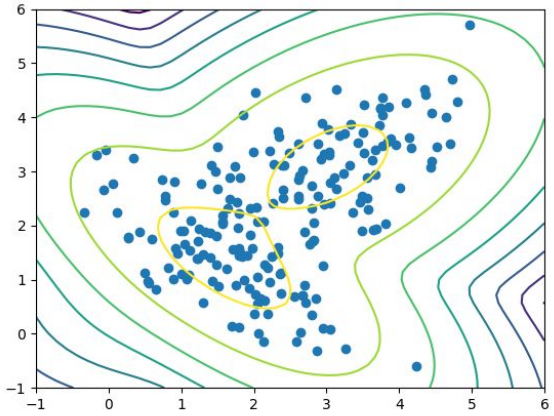
DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Agrupamiento (*clustering*)

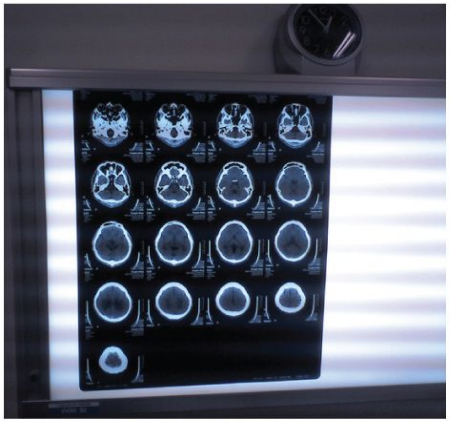
En ocasiones, es posible dividir una colección de observaciones en distintos subgrupos, basados únicamente en los atributos de las observaciones.

La intención es realizar división de datos en grupos (clusters) de observaciones, que son más similares dentro de un grupo que entre varios grupos.

Los grupos son formados, ya sea agregando observaciones o dividiendo un gran grupo de observaciones en una colección de grupos más pequeños.

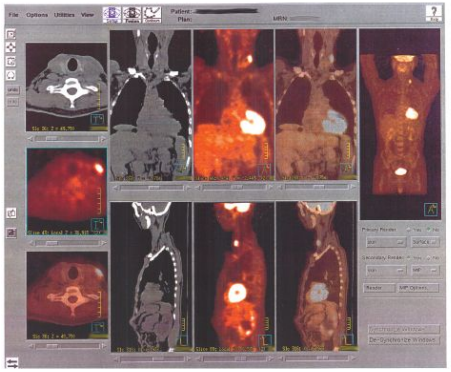


Modelos de mezclas gaussianas (GMM)

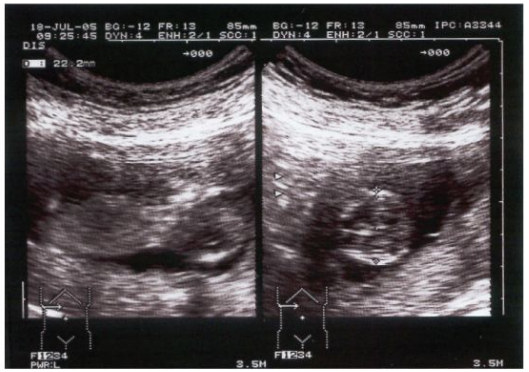


(a)

(b)



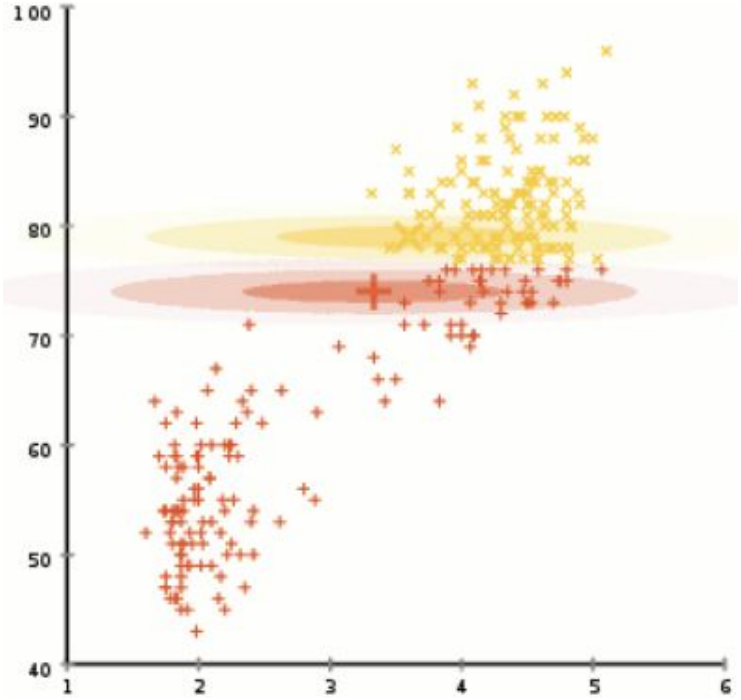
(c)



(d)

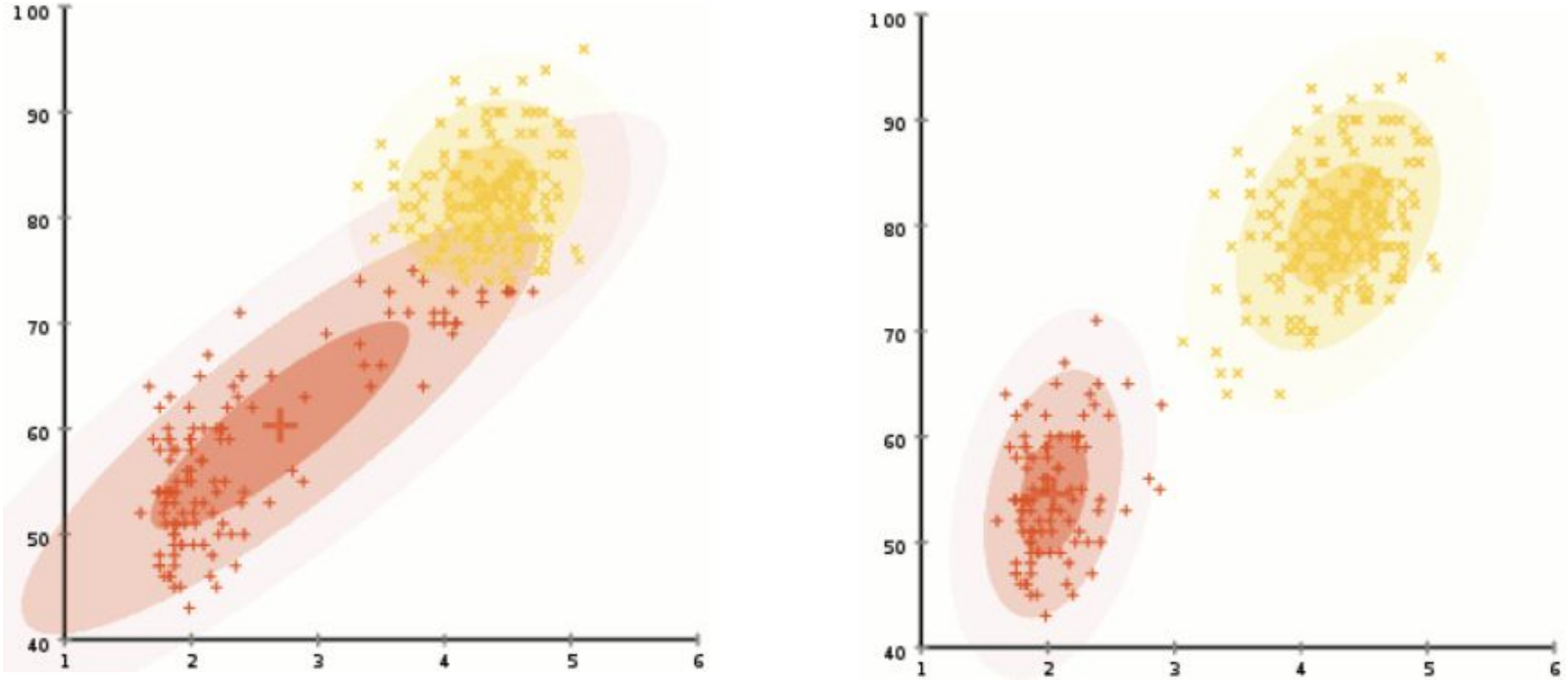
By OpenStax - https://cnx.org/contents/FPtK1zmh@8.25:fEI3C8Ot@10/Preface, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=30131135

Modelos de mezclas gaussianas, definición



Similar a *k-means*, en estos modelos se establece una cantidad *k* de grupos para iniciar.

Modelos de mezclas gaussianas, definición

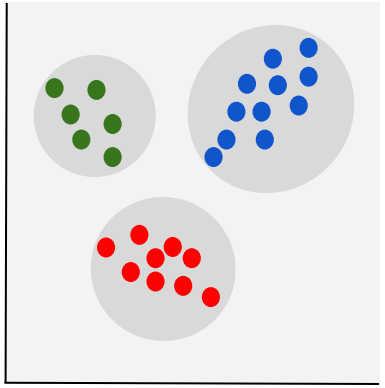


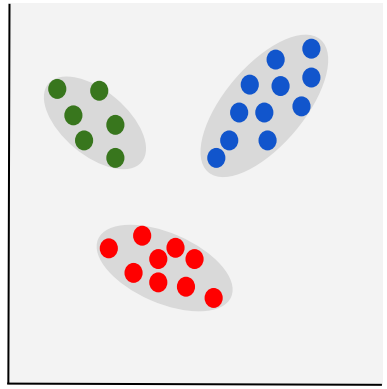
E iterativamente actualiza la media y la varianza de los grupos, así como la probabilidad de pertenencia de cada punto a cada grupo y el *peso* del mismo (cantidad de muestras de cada grupo).

Modelos de mezclas gaussianas vs *k-means*

Si bien el algoritmo inicia con un número predefinido de grupos al igual que *k-means*, existen diferencias fundamentales:

- *k-means* establece un círculo (hiperesfera) al centro del grupo, con radio definido por el punto más distante
- *GMM* funciona mejor cuando los datos no se ajustan a circunferencias (hiperesferas)
- *k-means* realiza clasificación *dura*: devuelve la clase a la que pertenece una muestra, mientras *GMM* hace clasificación *suave*: devuelve la probabilidad de pertenencia a cada clase.

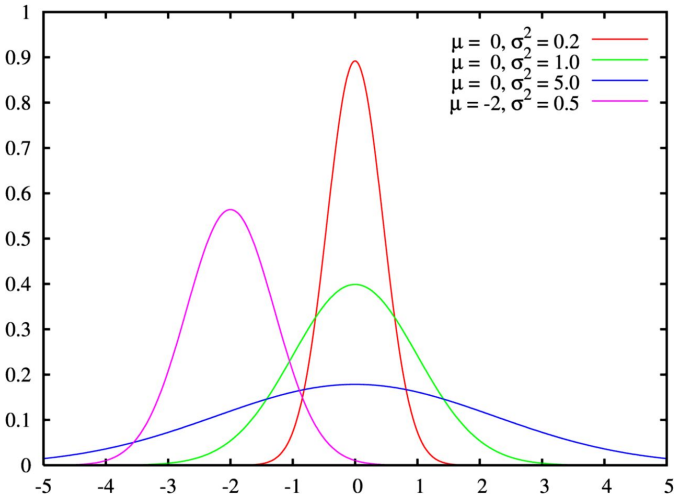




Expectation Maximization

Expectation Maximization (EM), es la técnica más popular para determinar los parámetros de un modelo de mezclas gaussianas. Se compone de dos pasos:

- Paso E: asignar de forma probabilística cada muestra a cada clase, basándose en la hipótesis actual de los parámetros.
- Paso M: actualizar las hipótesis de los parámetros, en función de las asignaciones actuales.



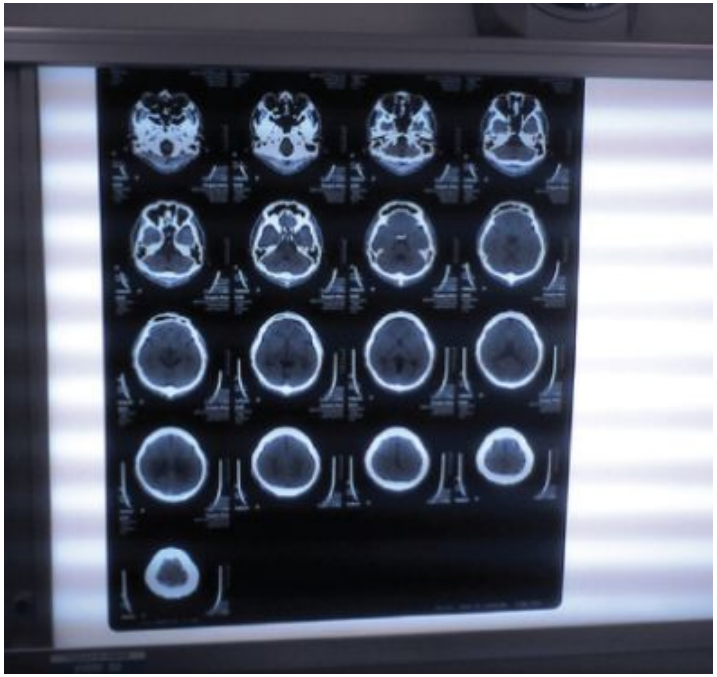
En el paso E se calcula el valor esperado de las asignaciones de grupos. En el paso M se obtiene la nueva máxima verosimilitud de las hipótesis

Expectation Maximization, aplicaciones

El algoritmo EM puede utilizarse en situaciones en las que se debe tomar en cuenta la distribución de las muestras. Por ejemplo:

- Reconstrucción de imágenes médicas

- Procesamiento de lenguaje natural:
  - *Baum–Welch algorithm*: para encontrar parámetros de un HMM; también útil en bioinformática
  - *Inside–outside algorithm*: para estimar probabilidades de producción en gramáticas libres de contexto probabilísticas



Referencias

- T.K. Moon
  The expectation-maximization algorithm
  http://home.ustc.edu.cn/~xiaosong/ppt/EM_tutorial.pdf

- Min Han and Yuyan Xu
  Application of Expectation Maximization Algorithm in Magnetic Induction Tomography
  https://link.springer.com/content/pdf/10.1007/s13534-015-0192-0.pdf

Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

https://twitter.com/laloea

https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/