Minería de Datos
IIC2433

Gaussian Mixture Models (GMM)

Vicente Domínguez

¿Qué veremos esta clase?

- Una nueva forma de hacer clústering

Algoritmos de clustering

- K-Means
- Mean Shift
- DBSCAN
- Clustering Jerárquico
- …

Son modelos "simples" o definidos en base a heurísticas

- Nos gustaría algo un poco más estadístico

Gaussian Mixture Models (GMM)
Definiciones

- Modelo no supervisado probabilístico
- Se basa en asumir que la distribución de los datos está compuesta por una mezcla de gaussianas
- Se puede ver como una generalización suave del modelo K-Means

Gaussian Mixture Models (GMM)

Propiedades

- Permite obtener la probabilidad de que un punto pertenezca a un cluster
- Un punto puede pertenecer a más de un cluster
- A diferencia de K-Means, no solo encuentra cluster de forma circular, sino que también de forma elíptica

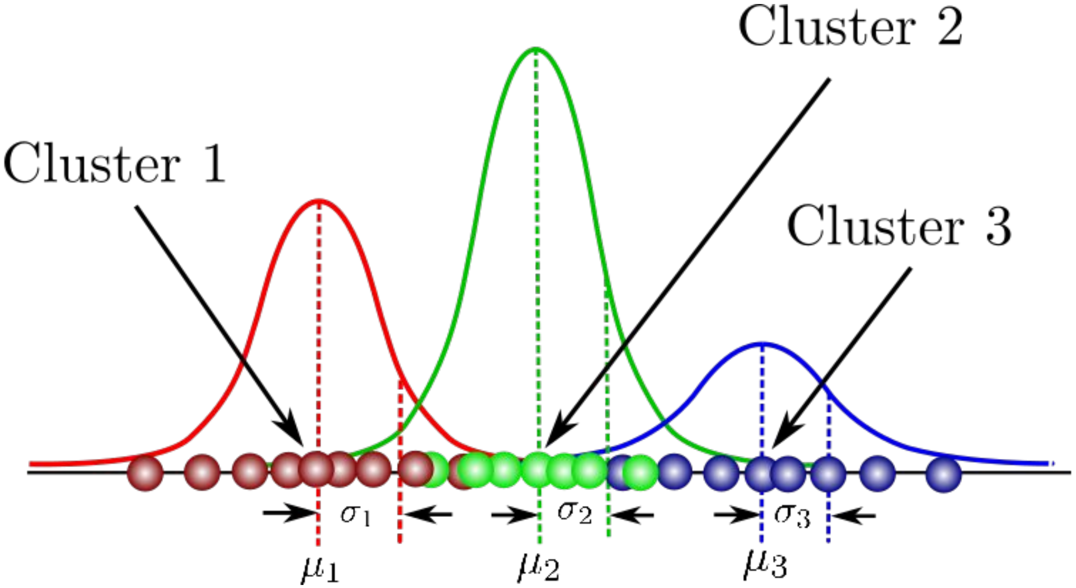
Gaussian Mixture Models (GMM)
Parámetros

- Al igual que K-Means, recibe un valor inicial que sería la cantidad de gaussianas a encontrar.

- Para cada guassiana, tratará de ajustar 3 parámetros.
    - Su media o su vector de medias
    - Su varianza o matriz de covarianza
    - Un parámetro escalador/ponderador

Gaussian Mixture Models (GMM)
Parámetros

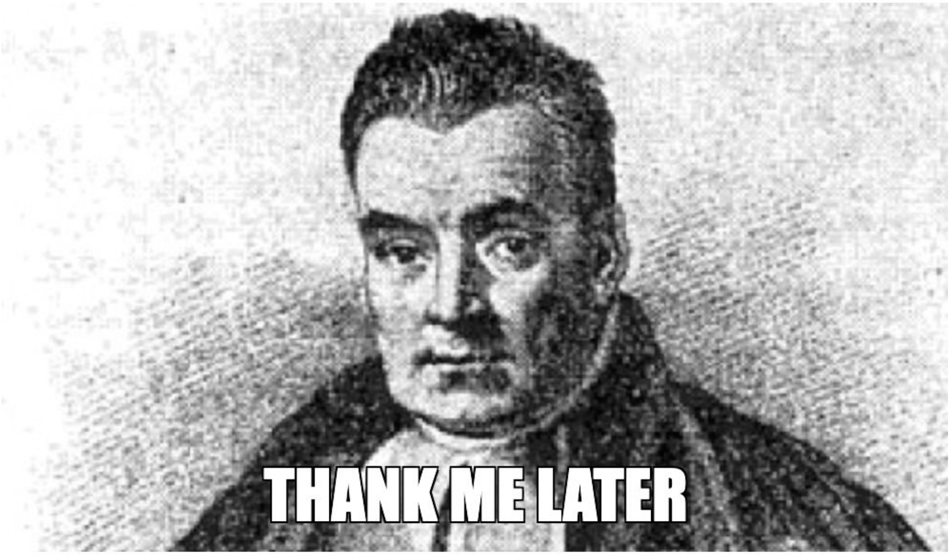
- Por ejemplo para K = 3 y gaussianas de 1 dimensión ajustará:
  - 3 medias
  - 3 varianzas
  - 3 valores escaladores



https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

Gaussian Mixture Models (GMM)
Parámetros

- ¿Cómo entrenamos estos parámetros?
  - Utilizando el algoritmo Expectation Maximization (EM)
  - Y el señor Bayes



Teorema de Bayes

- P(A = sí): Probabilidad del evento A sea “sí”
- P(A=sí|B=sí): Probabilidad de que el evento A sea “sí” DADO QUE el evento B fue “sí”
- Por simplicidad, usamos P(A) = P(A=“sí”)

$$P(A \mid B) = \frac{P(B|A) * P(A)}{P(B)}$$

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Teorema de Bayes

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Prior: Distribución de probabilidad a priori. El conocimiento de la probabilidad o incerteza de la clase antes de observar o condicionar los datos.

Likelihood: La probabilidad del evento bajo cierta clase o categoría, condicionada por los datos.

Evidence: Suma de las probabilidades del evento bajo todas las clases.

Posterior: Distribución de probabilidad condicional, que representa la probabilidad del evento condicionado luego de observar los datos.

EM

- Algoritmo compuesto por dos pasos, *Expectation* y *Maximization*
- Se utiliza para maximizar el valor de la *likelihood* de alguna distribución de probabilidad

- *Expectation:*
    - Se encarga de obtener el valor esperado actual, dado por los datos (*likelihood*)
- *Maximization:*
    - En base a los valores obtenidos en el paso E, se actualizan los parámetros maximizando el valor de la *likelihood*.

Gaussian Mixture Models (GMM)

Entrenamiento

- Para cada uno de los pasos del EM (E y M), se va calculando la probabilidad de cada dato de pertenecer a un cluster.
- Se actualizan los parámetros hasta que no varíen en un delta pequeño.

$$C_k \sim \mathcal{N}(\mu, \sigma^2)$$

Gaussian Mixture Models (GMM)

E step

- Se eligen parámetros iniciales para cada cluster, el parámetro ponderador parte inicialmente como 1/k para ponderar de forma uniforme inicialmente
- Para cada observación $\mathbf{x_i}$ se calcula su valor esperado con respecto a la gaussiana estimada.

$$C_k \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(\mathbf{x}|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(\mathbf{x} - \mu_k)^2}{2\sigma_k^2})$$

Gaussian Mixture Models (GMM)

M step

- Luego de obtener el valor esperado de pertenecer a la gaussiana, obtenemos su probabilidad a posteriori utilizando el teorema de bayes para cada cluster.
- Este es la *likelihood* o verosimilitud, de que un dato **x_i** haya sido generado por el cluster **K**

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

$$\mathbf{b}_k = \frac{f(\mathbf{x}|\mu_k, \sigma_k^2)\phi_k}{\sum_{k=1}^{K} f(\mathbf{x}|\mu_k, \sigma_k^2)\phi_k}$$

Gaussian Mixture Models (GMM)

M step

- Por último, re-calculamos los parámetros para cada cluster **k** y generamos una nueva iteración.
- El proceso se detiene cuando hay una variación menor a un delta en alguno de los parámetros definidos

$$\mu_k = \frac{\sum \mathbf{b}_k \mathbf{x}}{\sum \mathbf{b}_k} \qquad \sigma_k^2 = \frac{\sum \mathbf{b}_k (\mathbf{x} - \mu_k)^2}{\sum \mathbf{b}_k} \qquad \phi_k = \frac{1}{N} \sum \mathbf{b}_k$$

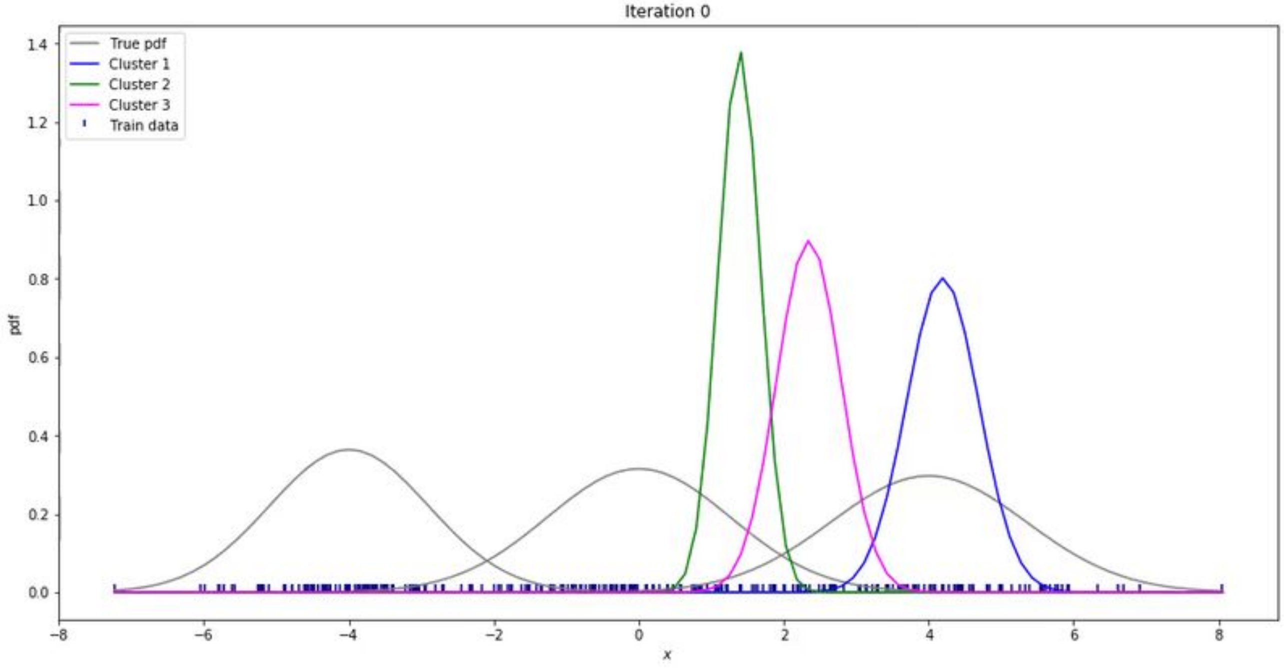
Gaussian Mixture Models (GMM)

M step

- Por último, re-calculamos los parámetros para cada cluster **k** y generamos una nueva iteración.
- El proceso se detiene cuando hay una variación menor a un delta en el valor obtenido en la likelihood

$$\mu_k = \frac{\sum \mathbf{b}_k \mathbf{x}}{\sum \mathbf{b}_k} \qquad \sigma_k^2 = \frac{\sum \mathbf{b}_k (\mathbf{x} - \mu_k)^2}{\sum \mathbf{b}_k} \qquad \phi_k = \frac{1}{N} \sum \mathbf{b}_k$$

Gaussian Mixture Models (GMM)



https://towardsdatascience.com/how-to-code-gaussian-mixture-models-from-scratch-in-python-9e7975df5252

Gaussian Mixture Models (GMM)

https://towardsdatascience.com/how-to-code-gaussian-mixture-models-from-scratch-in-python-9e7975df5252

Gaussian Mixture Models (GMM)

- Hay toda una demostración matemática de por qué esas son las ecuaciones a actualizar en el algoritmo. Si quieren saber más al respecto, acá pueden ver una explicación detallada.
- Al ser un enfoque con distribuciones de probabilidad, nos permite tener una base más sólida de por qué se formaron los clusters.
- Debemos recordar que es un algoritmo no supervisado. Por lo que independiente de los resultados, no tenemos certeza de que son unos buenos clusters.
- Los códigos utilizados para los plots pueden encontrarlos acá.
- GMM 1D, GMM 2D