Probabilistic Methods
Expectation-Maximization

Belén C. Saldías F.
https://belencarolina.com
Departamento de Ciencia de la Computación

Un poco sobre mi

Mi investigación

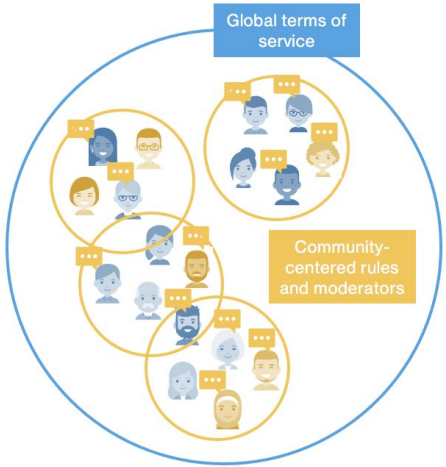
MIT center for constructive communication

Designing tools, methods, and systems to understand and address societal fragmentation.

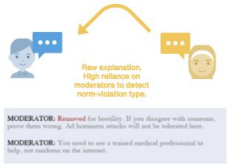
https://www.media.mit.edu/projects/explaining-machine-supported-community-content-moderation/overview/

Explaining machine-supported community content moderation

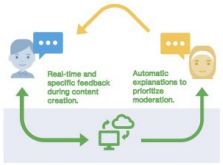
Belén Saldías & Deb Roy



How can we provide **better explanations** behind **norm violations** to help **users engage in diverse communities** and scale the **community-guided content moderation**?

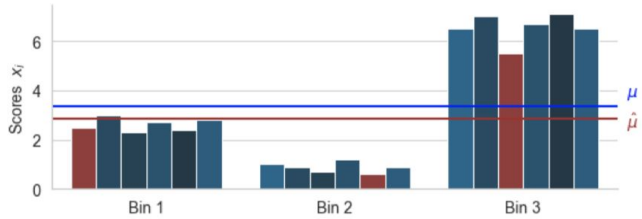


Current scenario

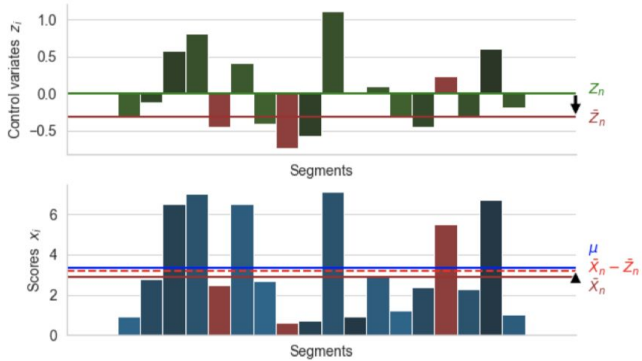


Rationale-focused scenario

Mi investigación



(a) Stratified sampling forces sampled segments (shown in red) to be evenly distributed across bins, resulting in better estimates when the score variance within bins is lower than the variance across bins.



(b) Control variates allow for reversing the shift of the sample mean $\bar{X}_n$ depending on the strength of the correlation between X and Z. In this illustration, where X and Z are highly correlated ($\sim$0.9), $\bar{Z}_n < 0$ reflects the negative shift in $\bar{X}_n$.

Figure 1: Complementary strategies for reducing the variance of the estimated average score.

Saldías, B., Foster, G., Freitag, M., & Tan, Q. (2022). ***Toward More Effective Human Evaluation for Machine Translation***. In HumEval @ ALC 2022.

Objetivos de la clase

- Formalizar la noción de *likelihood*, y entender cómo esto ayuda a resolver problemas de estimación.
- Entender las ventajas y aplicaciones de hacer *clustering* usando **Mezcla de Gaussianas** (*Gaussian Mixture*).
- Comprender cómo funciona y aplicar el método de estimación de parámetros **Esperanza-Maximización** (*Expectation-Maximization*) *(EM)*.

Soft Clustering

Regresión logística - Maximizar likelihood

- Encontrar **W** tal que

$$\max P(Y|\Theta)$$

$$\Theta_i = P(\hat{y}_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{D} \beta_j * x_{ij})}}$$

Formalizar la noción de *likelihood*.

Función de densidad de probabilidad

- Datos o instancias → eventos, observaciones, o realizaciones de variables aleatorias subyacentes (latentes).

- Variable discreta A:
  - P(A) codifica la probabilidad para cada categoría, clase o estado en el que A puede estar.
  - P(A = a) = P(a) es la probabilidad de observar el evento específico de que A tome valor a.

- Variable continua X:
  - P(X) asigna una probabilidad de densidad a todos los posibles valores de X.
  - P(x1) corresponde al valor escalar obtenido de evaluar P(X = x1).

Reglas bases

- *Product rule*: regla fundamental de las probabilidades.

$$P(A, B) = P(A|B)P(B)$$

- *Sum rule*: dada una distribución conjunta, permite obtener una **marginal**.

$$P(x_1) = \sum_{x_2} \dots \sum_{x_N} P(x_1, x_2, \dots, x_N)$$

Reglas bases

- *Bayes' rule*: vincula probabilidades condicionales. Válido en todas las aplicaciones de teoría de probabilidades.

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

Probabilidad vs. *Likelihood*

- **Probabilidad**: se refiere a la **posibilidad de que ocurra un resultado particular en función de un modelo** (y sus parámetros).

- ***Likelihood***: se refiere a **qué tan bien una muestra proporciona explicación para un modelo** (y sus parámetros).

$$P(\Theta|X) = \frac{P(X|\Theta)P(\Theta)}{P(X)} \qquad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Probabilidad vs. *Likelihood*

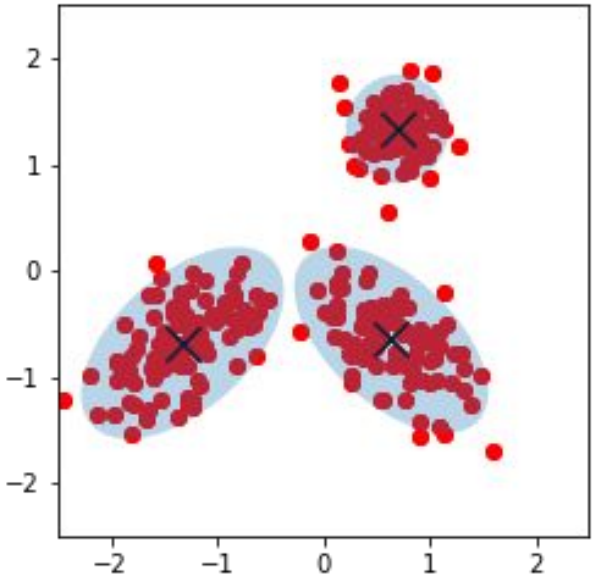
$$P(\Theta|X) = \frac{P(X|\Theta)P(\Theta)}{P(X)} \qquad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

- Lanzar una moneda
- Vender un computador al cliente que acaba de entrar

Mezcla de Gaussianas (*Gaussian Mixture*)

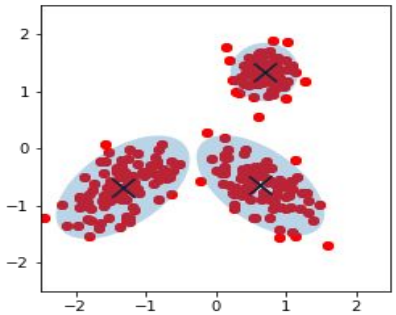
Mezcla de Gaussianas (*Gaussian Mixture*)

- Encontrar el **conjunto de *clusters*** más probables dado un set de datos.



Mezcla de Distribuciones

- Mezcla finita:
  - Conjunto de K distribuciones de probabilidad.
  - Cada distribución representa un *cluster*.
  - Cada distribución da la probabilidad de que una instancia haya sido generada por ella.
  - Cada distribución no es igual de probable que el resto.
  - Existe una distribución de probabilidad que gobierna los tamaños relativos de los *clusters*.



¿Qué parámetros se deben estimar en este caso?

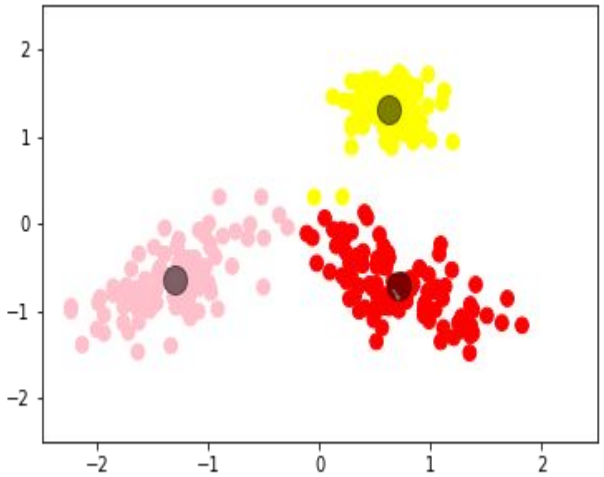
K-Means vs. EM en Mezcla de Gaussianas

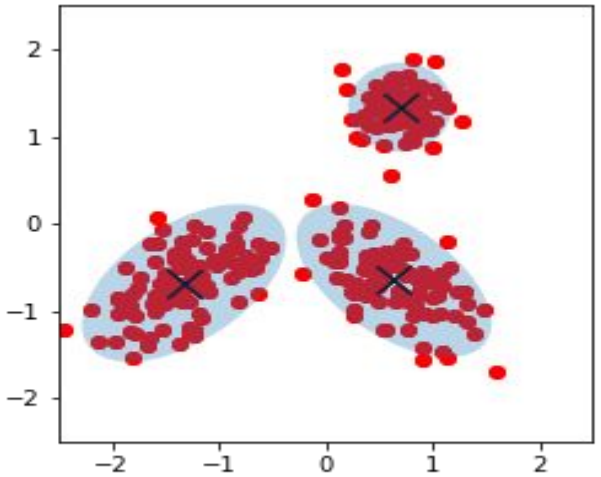
- K-Means
  - Cada instancia pertenece a un solo segmento → asignación *hard*.
  - El segmento de cada instancia es asignado según distancia Euclidiana.
  - En 2-D se produce un círculo, en R-D una hiperesfera.
  - No toma en cuenta covarianza de los datos.

- EM Mezcla de Gaussianas
  - Asignación *soft* de pertenencia a un grupo.
  - Entrega probabilidad de pertenencia.
  - No depende de la distancia.
  - Depende de la probabilidad de que una instancia haya sido generada por una distribución.
  - Toma en cuenta la matriz de covarianza de los datos, para determinar la probabilidad.

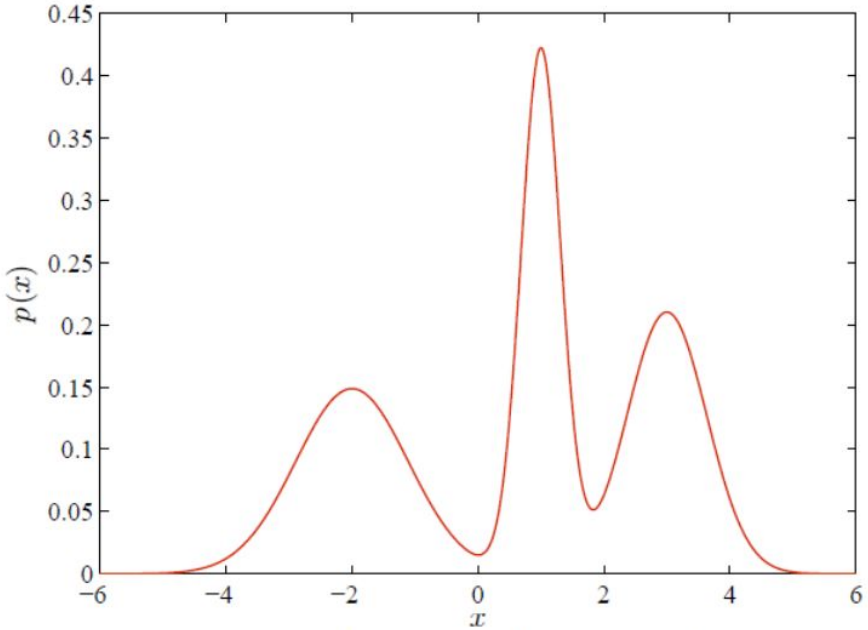
- *K-Means se ve afectada por la norma L2, la mezcla de Gaussianas no.*

K-Means vs. EM en Mezcla de Gaussianas



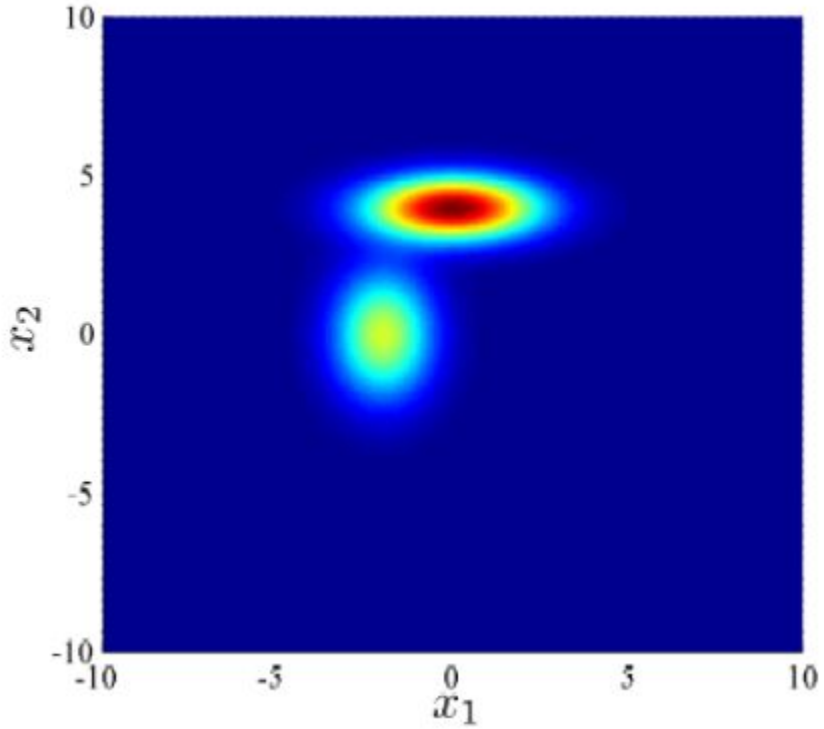


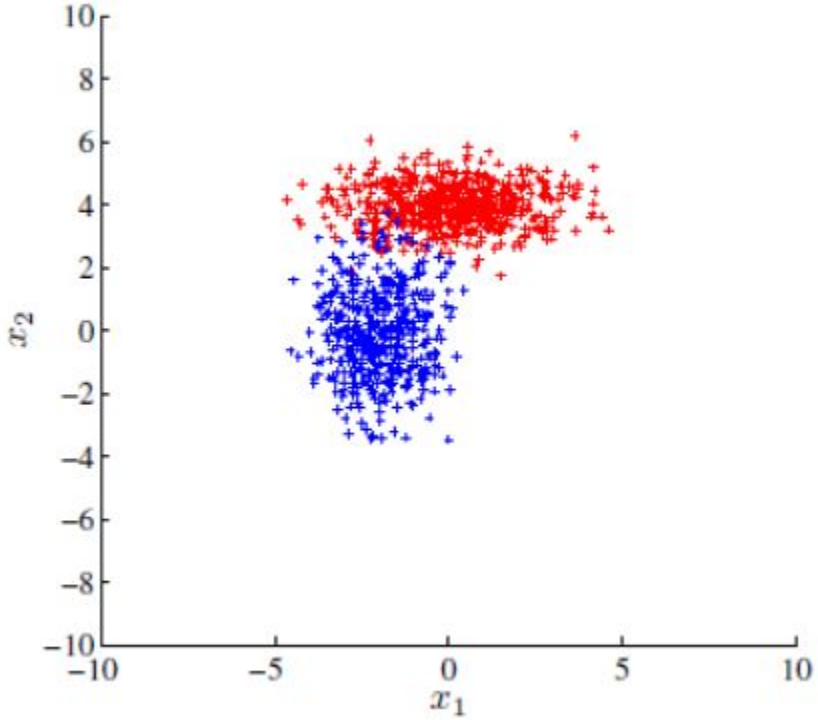
Mezcla de 3 Gaussianas 1-D



$$\mu_1 = -2,\ \mu_2 = 1,\ \mu_3 = 3,$$
$$\sigma_1^2 = 0.8,\ \sigma_2^2 = 0.1,\ \sigma_3^2 = 0.4,\ w_1 = w_2 = w_3 = 1/3$$

Mezcla de 2 Gaussianas 2-D





Obtención de parámetros de la mezcla

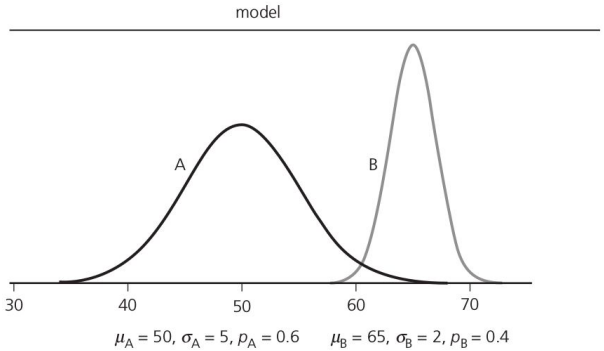
- Si se conocieran las clases de las instancias, se pueden obtener fácilmente los parámetros de las Gaussianas.

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N - 1}$$

Probabilidad de pertenencia a *cluster*

- Al conocer los 5 parámetros del modelo, encontrar la probabilidad de pertenencia a cada distribución es fácil.



$$P(A|x_i) = \frac{P(x_i|A)P(A)}{P(x_i)}$$

Probabilidad de pertenencia a *cluster*

- Al conocer los 5 parámetros del modelo, encontrar la probabilidad de pertenencia a cada distribución es fácil.
- **El problema es que una quiere hacer clustering porque no conoce ninguno de los parámetros que definen las distribuciones, ni a qué conjunto pertenecen los datos.**

$$P(A|x_i) = \frac{\mathcal{N}(x_i \mid \mu_a, \sigma_a) P_A}{P(x_i)}$$

EM - Algoritmo para Mezcla de Gaussianas

- Problema:
  - No se conocen las distribuciones latentes en los datos
  - No se conocen los parámetros de estas distribuciones
  - No se sabe a qué conjunto pertenecen los datos

- Solución:
  - Se adopta la idea de K-means y se itera probabilísticamente
  - Iniciar con medias iniciales μ y estimaciones para Σ según resultado de K-means
  - **Usar esos parámetros para calcular la probabilidad de pertenencia esperada a cada cluster *(E)***
  - **Usar esas probabilidades para a través de maximizar likelihood re-estimar los parámetros *(M)***

- Esta es una instancia del algoritmo *EM → Expectation - Maximization*

¿Hasta cuándo iterar?

- K-Means termina cuando las instancias no se cambian más de *cluster* de una iteración a otra → se alcanza un punto fijo.

- EM
  - Converge a un punto fijo, pero nunca llega ahí.
  - Converge cuando la log-likelihood prácticamente ya no cambia.
  - Diferencia de log-like es menor a 10^-10 durante 10 iteraciones seguidas.

- Marginal log-likelihood

$$\prod_{i=1}^{N} P(x_i) = \prod_{i=1}^{N} \sum_{j=1}^{K} P(x_i \mid c_j) P(c_j)$$

Esperanza-Maximización (*Expectation-Maximization*)
(EM)

MLE: Maximum Likelihood Estimation

- Considerar el problema de estimar el set de parámetros Θ de un modelo probabilístico, dado un dataset X.

- MLE asume:
  - Los datos no dependen unos de otros (la ocurrencia de uno no afecta la de otros).
  - Todas las instancias pueden ser modeladas de la misma manera.
  - ➢ i.i.d. Estructuras dependientes pueden ser capturadas por modelos más sofisticados.

$$\Theta_{ML} = \arg\max_{\Theta} \sum_{i=1}^{N} logP(x_i|\Theta)$$

EM - Algoritmo para Mezcla de Gaussianas

- Expectation E-Step:
  - Se asume la existencia de variables latentes
  - Se calculan las probabilidades de pertenencia a las Gaussianas
  - Se obtiene el valor esperado de las pertenencias

- Maximization M-Step:
  - Cálculo de parámetros de máxima verosimilitud
  - Cálculo de parámetros de las variables latentes
  - Se maximiza la likelihood de la distribución dados los datos

- Los parámetros encontrados en M se usan para recalcular E

EM - Obtención de parámetros de la mezcla

- Se trabaja con probabilidad de pertenencia, no con clases conocidas.
- Las probabilidades actúan como pesos.

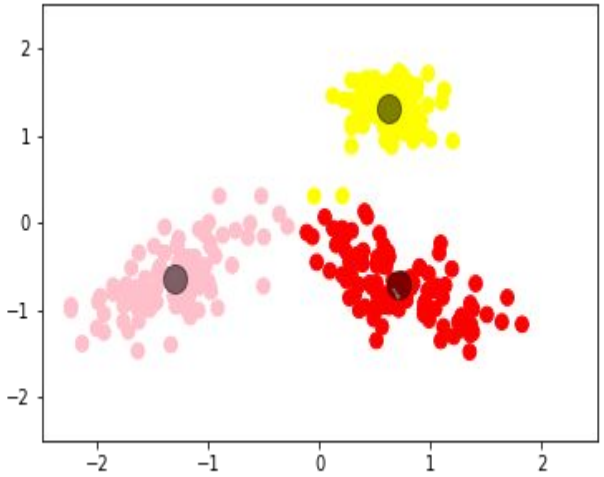
$$w_{Ai} = P(A|x_i)$$

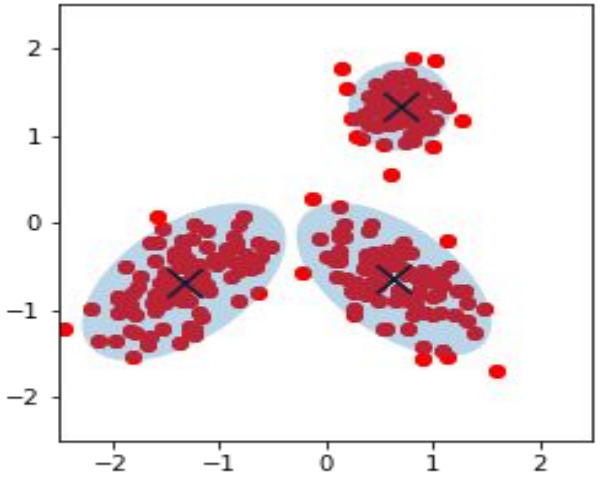
$$\mu_A = \frac{\sum_{i=1}^{N} w_{Ai}\, x_i}{\sum_{i=1}^{N} w_{Ai}}$$

$$\sigma_A^2 = \frac{\sum_{i=1}^{N} w_{Ai}(x_i - \mu)^2}{\sum_{i=1}^{N} w_{Ai}}$$

Objetivo

- Aplicar EM para encontrar *clusters* multidimensionales





Probabilistic Methods
Expectation-Maximization

Belén C. Saldías F.
https://belencarolina.com
Departamento de Ciencia de la Computación