ALGORITMO DE CLASIFICACIÓN CON MEZCLAS GAUSSIANAS

Juan Camilo Rojas Cortés

CONTENIDO

Objetivo del algoritmo

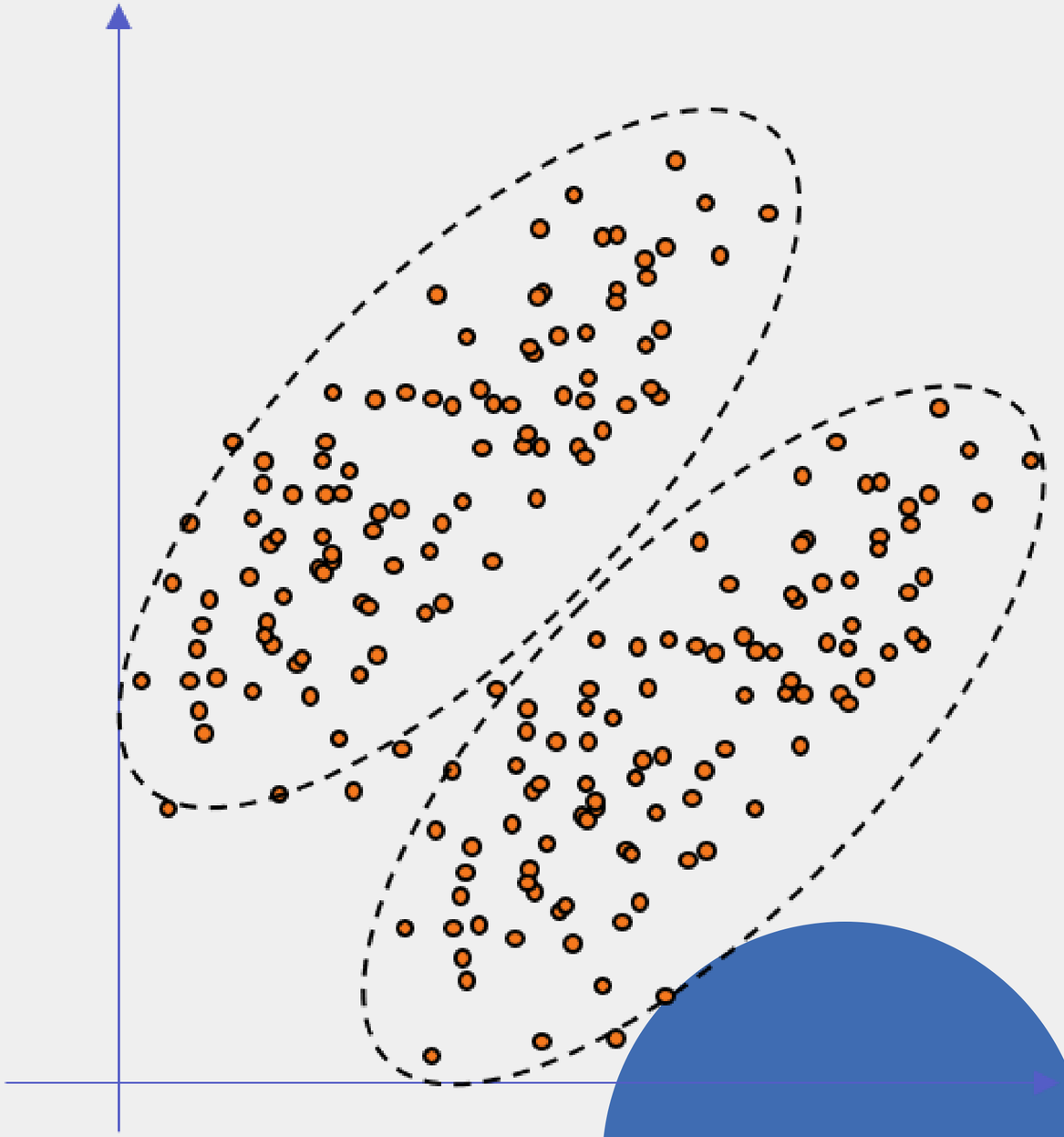
Conceptos estadísticos previos

Aproximación matemática al algoritmo

Ejemplo de implementación

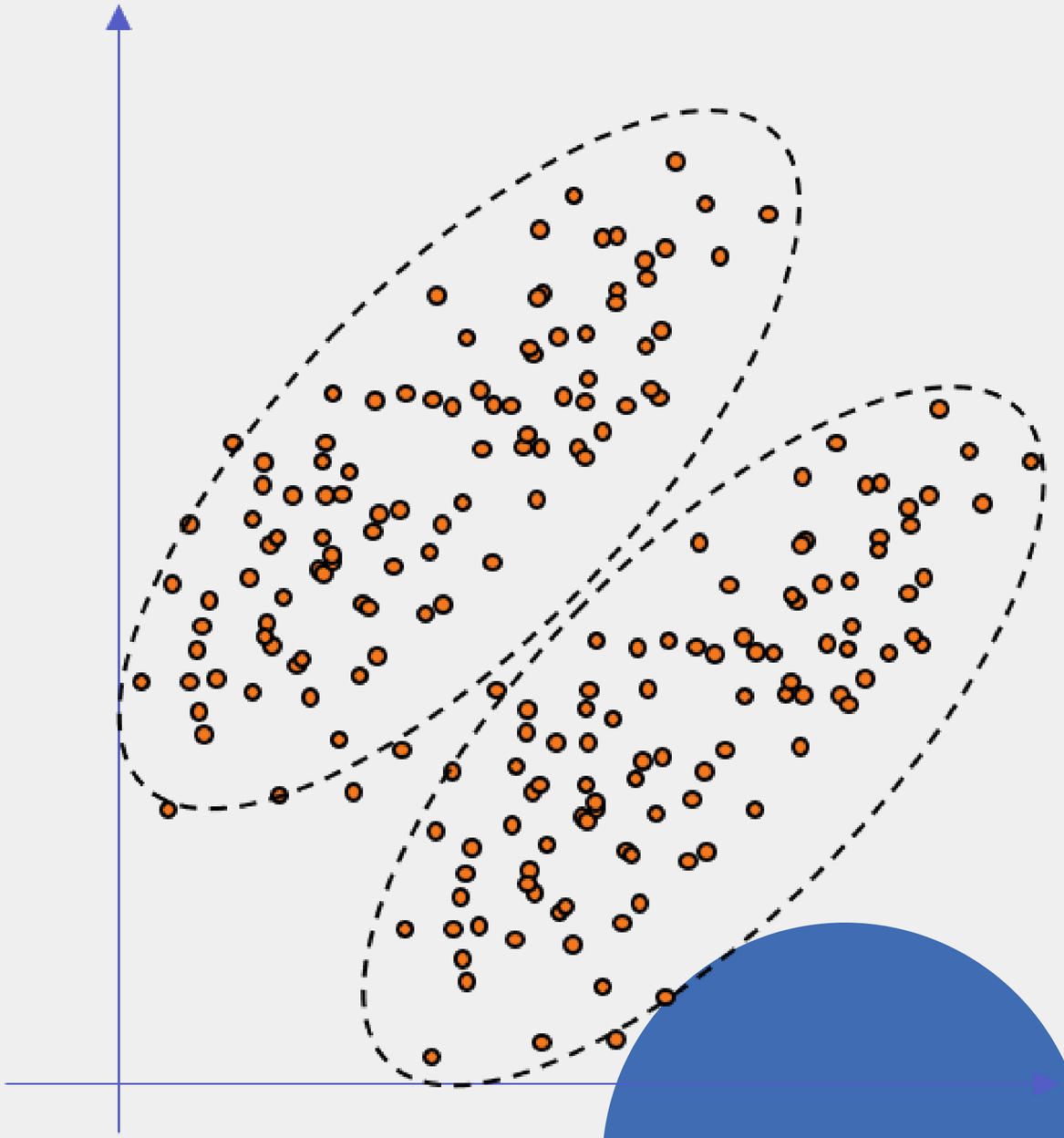
OBJETIVO DEL ALGORITMO

Procesar un grupo de datos de N dimensiones sin relación o clasificación aparente y agruparlos utilizando distribuciones normales (gaussianas) de probabilidad.



OBJETIVO DEL ALGORITMO

Procesar un grupo de datos de N dimensiones sin relación o clasificación aparente y agruparlos utilizando distribuciones normales (gaussianas) de probabilidad.

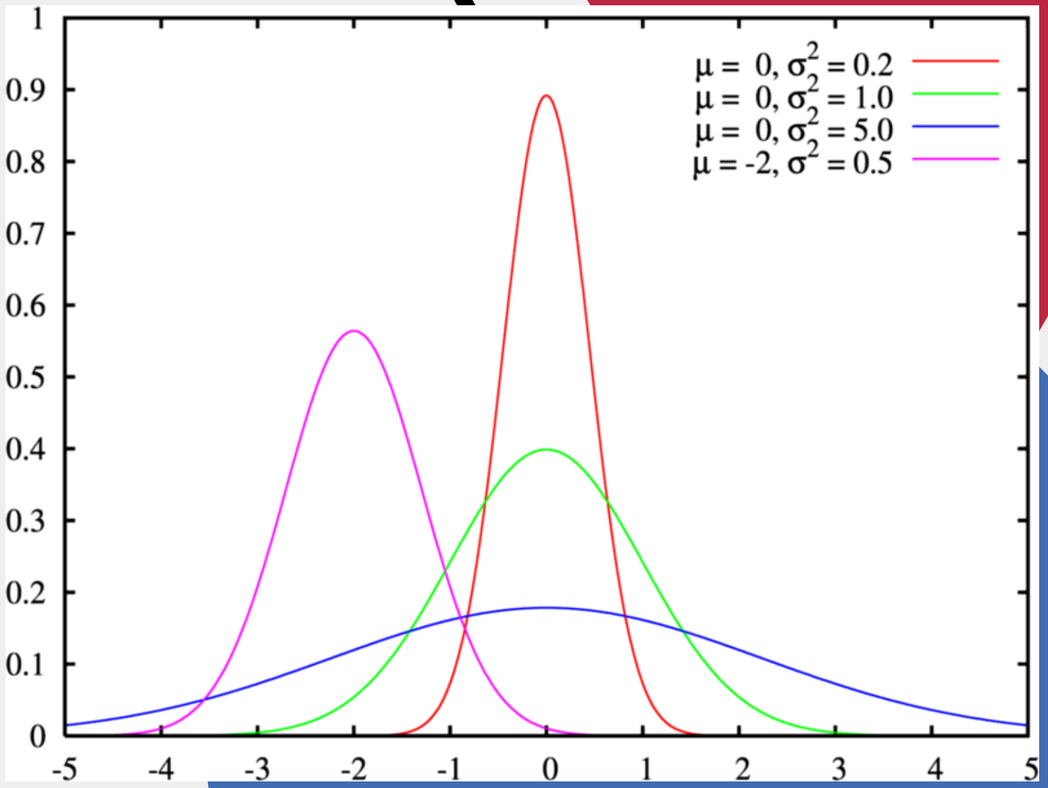


MEDIA

Parámetro que define el punto medio o promedio de los datos. Gráficamente, indica en qué punto está la mitad de la función de probabilidad.

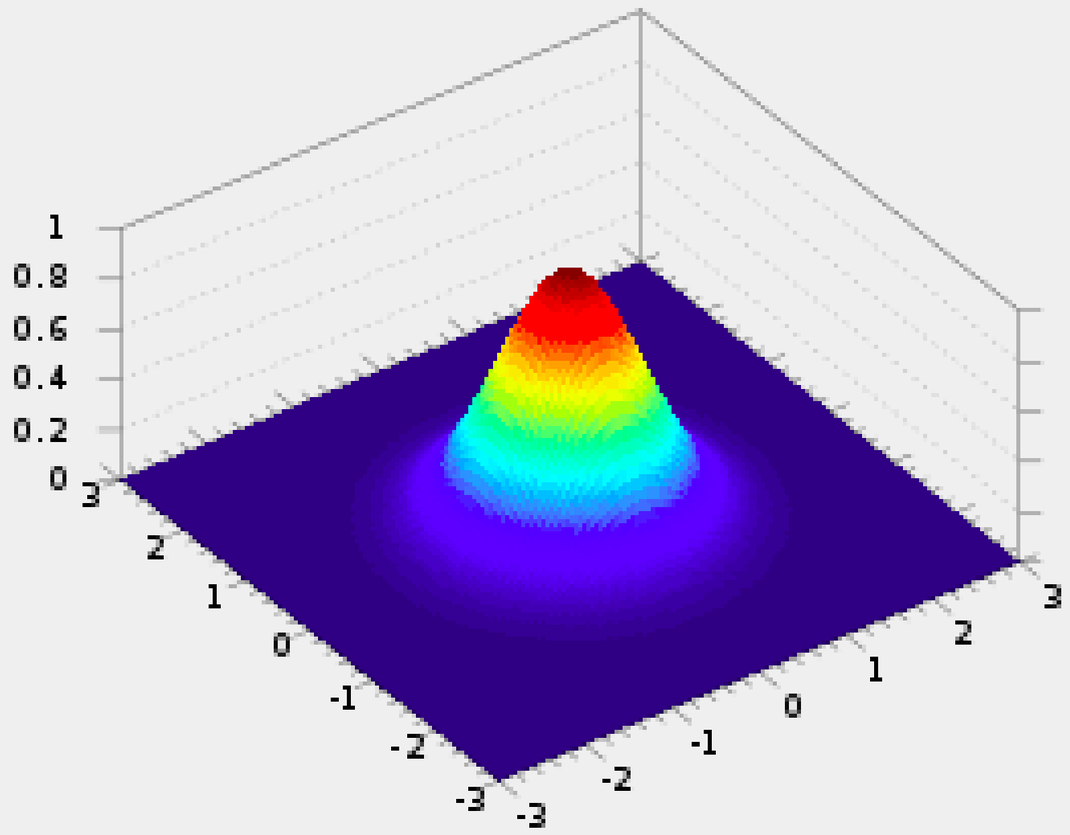
VARIANZA

Indica el nivel de dispersión de los datos. Gráficamente, define qué tan ancha o angosta es la función de probabilidad.



GAUSSIANA EN N DIMENSIONES

La función Gaussiana puede llevarse a tres o más dimensiones para efectos gráficos. Por ejemplo, para clasificación de datos bidimensionales, la función se ve de la siguiente forma:



MATRIZ DE COVARIANZAS

Para hacer distribuciones de datos con maś de una dimensión (gaussianas de 3 o más dimensiones) es necesario utilizar matrices de covarianzas, en lugar de una sola varianza.

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

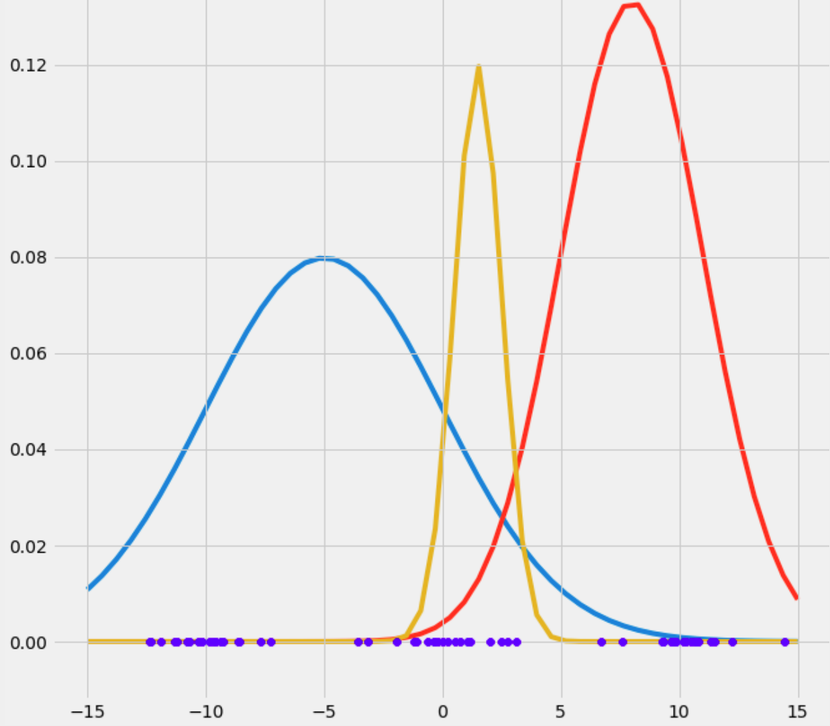
APROXIMACIÓN MATEMÁTICA AL ALGORITMO

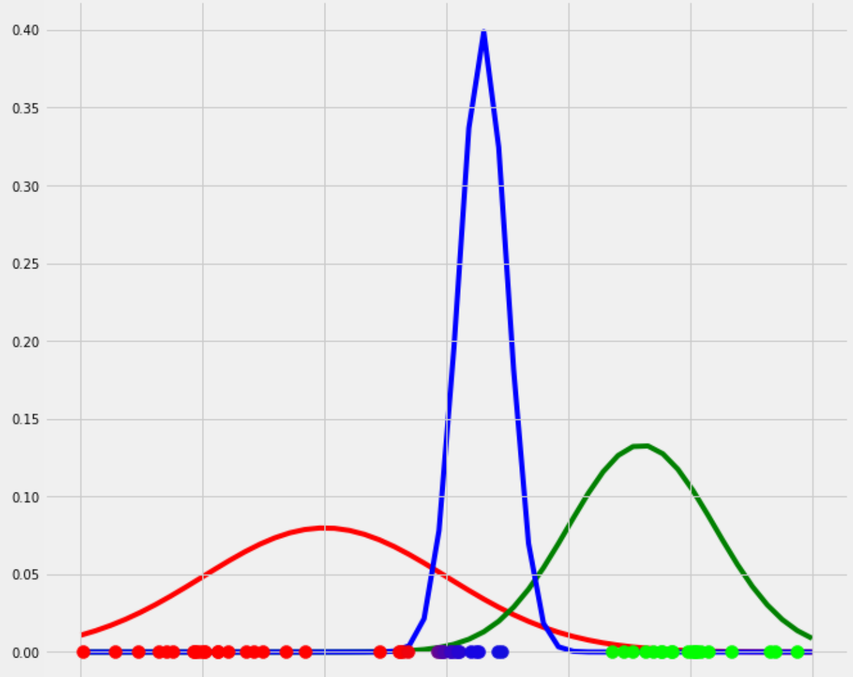
PASO E (EXPECTATION)

Con gaussianas existentes (inicialmente aleatorias) se calcula la probabilidad que tiene cada dato de pertenecer a cada una de estas

PASO M (MAXIMIZATION)

Se modifican las gaussianas con el objetivo de que, al final, todos los puntos tengan una probabilidad cercana a 1 de pertenecer a una de las gaussianas y cercana a 0 de pertenecer a todas las demás





PASO E

- Se calcula la probabilidad que tiene cada dato (r) de pertenecer a una gaussiana o clúster (c)
- La función N del término anterior se refiere a la función gaussiana multivariada

$$r_{ic} = \frac{\pi_c N(x_i \mid \mu_c, \Sigma_c)}{\sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \Sigma_k)}$$

$$N(x_i, \mu_c, \Sigma_c) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_c|^{\frac{1}{2}}} exp(-\frac{1}{2}(x_i - \mu_c)^T \Sigma_c^{-1}(x_i - \mu_c))$$

EJEMPLO DE IMPLEMENTACIÓN