Teoria e Aplicação de Misturas Gaussianas (GMM)

Introdução

O **Modelo de Mistura Gaussiana (GMM)** é uma técnica estatística amplamente utilizada para modelar dados que podem ser representados como uma combinação de várias distribuições normais (*Gaussianas*). Ele é usado principalmente para tarefas como **agrupamento** (*clustering*) e **modelagem de densidade**.

Neste relatório, explicaremos a teoria por trás do GMM, destacando os principais conceitos matemáticos e como eles são aplicados para separar um conjunto de dados em grupos.

—

A Intuição por Trás da Mistura Gaussiana

Imagine que temos um conjunto de dados, como alturas de pessoas, e queremos dividi-los em dois grupos: "adolescentes" e "adultos". Cada grupo pode ser representado por uma distribuição normal (ou Gaussiana), caracterizada por:

- Uma **média** (μ), que define o centro da distribuição.
- Uma **variância** (σ^2), que define a dispersão dos dados ao redor da média.
- Um **peso** (w), que indica a proporção de dados pertencentes àquele grupo.

O modelo GMM assume que os dados são gerados a partir de uma combinação dessas distribuições normais. A fórmula geral para a densidade de probabilidade do GMM é:

$$p(x) = \sum_{k=1}^{K} w_k \cdot N(x|\mu_k, \sigma_k^2),$$

onde:

- K é o número de grupos (ou componentes).
- w_k é o peso da k-ésima componente, com $\sum_{k=1}^{K} w_k = 1$.
- $N(x|\mu_k, \sigma_k^2)$ é a função densidade da distribuição normal:

$$N(x|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}.$$

—

O Algoritmo Expectation-Maximization (EM)

O GMM utiliza o algoritmo EM (*Expectation-Maximization*) para ajustar os parâmetros (w_k, μ_k, σ_k^2) às características dos dados. O EM alterna entre duas etapas principais:

1. Etapa Expectation (E-Step)

Nesta etapa, calculamos as responsabilidades $\gamma(i, k)$, que representam a probabilidade de cada ponto x_i pertencer ao grupo k. A fórmula é:

$$\gamma(i, k) = \frac{w_k \cdot N(x_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} w_j \cdot N(x_i|\mu_j, \sigma_j^2)}.$$

Aqui:

- O numerador calcula a probabilidade do ponto x_i pertencer ao grupo k, levando em conta o peso w_k e a densidade normal $N(x_i|\mu_k, \sigma_k^2)$.
- O denominador normaliza as probabilidades para garantir que $\sum_{k=1}^{K} \gamma(i, k) = 1$.

A responsabilidade $\gamma(i, k)$ pode ser interpretada como uma "confiança" do modelo sobre qual grupo o dado pertence.

—

2. Etapa Maximization (M-Step)

Nesta etapa, usamos as responsabilidades calculadas na E-Step para atualizar os parâmetros do modelo (w_k, μ_k, σ_k^2):

- Atualização das médias:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma(i, k) x_i}{\sum_{i=1}^{N} \gamma(i, k)}.$$

  A nova média μ_k é uma média ponderada dos dados atribuídos ao grupo k, onde os pesos são as responsabilidades $\gamma(i, k)$.

- Atualização das variâncias:

$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma(i, k)(x_i - \mu_k)^2}{\sum_{i=1}^{N} \gamma(i, k)}.$$

  A nova variância σ_k^2 mede a dispersão dos dados ao redor da nova média μ_k, ponderada pelas responsabilidades.

- Atualização dos pesos:

$$w_k = \frac{\sum_{i=1}^{N} \gamma(i, k)}{N}.$$

  O novo peso w_k representa a proporção de pontos atribuídos ao grupo k.

—

Interpretação dos Parâmetros

Após várias iterações do algoritmo EM, os parâmetros convergem para valores que melhor descrevem os dados. Cada componente Gaussiana terá:

- Uma média (μ_k) que representa o centro do grupo.
- Uma variância (σ_k^2) que descreve a dispersão dos dados no grupo.
- Um peso (w_k) que indica o tamanho relativo do grupo.

Os valores finais das responsabilidades $\gamma(i, k)$ podem ser usados para classificar os dados em grupos ou para realizar agrupamento "suave", onde cada ponto pertence parcialmente a vários grupos.