

Modelo de Mistura de Gaussianas

Advanced Institute for Artificial Intelligence – AI2

<https://advancedinstitute.ai>

Modelo de Misturas de Gaussianas, do inglês *Gaussian Mixture Models* - GMMs, é uma técnica de **aprendizado não supervisionado** que pode ser entendida como uma generalização do k -Médias. Ao invés de estimarmos os centroides de cada agrupamento, tentamos estimar também a forma e proporção de cada Gaussiana que compõe a mistura.

Definição do problema: seja $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ um conjunto de dados tal que $\mathbf{x}_i \in \mathbb{R}^n$. Assumimos que as amostras são independentes e identicamente distribuídas (i.i.d.) a partir de uma função de densidade de probabilidade $p(\mathbf{x}_i)$, $\mathbf{x}_i \in \mathcal{X}$. Assumimos, também, que $p(\mathbf{x}_i)$ é uma **mistura finita** de K componentes, ou seja:

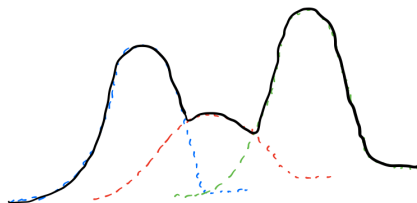
$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^K w_k p(\mathbf{x}_i|\theta_k), \quad (1)$$

em que $\theta_k = (w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ corresponde aos parâmetros relativos à Gaussiana k e $\sum_{k=1}^K w_k = 1$.

As componentes são densidades Gaussianas multivariadas dadas por:

$$p(\mathbf{x}_i|\theta_k) = \frac{1}{(2\pi)^{m'_k/2}|\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\}. \quad (2)$$

No caso, os pesos w_k representam a probabilidade que uma amostra $\mathbf{x}_i \in \mathcal{X}$ selecionada aleatoriamente tenha sido gerada pela componente k . Vejamos o exemplo de uma mistura 1D com $k = 3$.



Como não temos os rótulos, devemos utilizar alguma informação adicional para estimar $\theta = \{\theta_1, \theta_2, \dots, \theta_k\}$. Para tal tarefa, podemos empregar o conhecido algoritmo E-M (*Expectation-Maximization*), o qual segue a mesma ideia da abordagem de máxima verossimilhança.

A falta do rótulo dos dados caracteriza um **problema estatístico incompleto**, fazendo com que utilizemos as chamadas **variáveis latentes**, isto é, variáveis que **não são observadas**. Note que no caso do classificador Bayesiano temos o conjunto de rótulos, ou seja, observamos essas variáveis.

Seja $\mathcal{Z} = \{z_1, z_2, \dots, z_m\}$ um conjunto de variáveis latentes em que cada uma delas está relacionada com uma amostra do conjunto de treinamento \mathcal{X} . Dado o conjunto de parâmetros θ , temos que a função de verossimilhança é dada por:

$$L(\theta|\mathcal{X}, \mathcal{Z}) = p(\mathcal{X}, \mathcal{Z}|\theta), \quad (3)$$

em que $p(\mathcal{X}, \mathcal{Z}|\theta)$ corresponde à densidade conjunta de \mathcal{X} e \mathcal{Z} . Ademais, o estimador de máxima verossimilhança do vetor de parâmetros θ é dado, então, pela **maximização da verossimilhança marginal** (só temos acesso ao \mathcal{X}) dos dados observados, ou seja:

$$L(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \int p(\mathcal{X}, \mathcal{Z}|\theta) d\mathcal{Z}. \quad (4)$$

A formulação acima corresponde à calcular a integral da probabilidade conjunta em \mathcal{Z} .

No entanto, não temos como calcular essa integral manualmente, pois não temos conhecido de \mathcal{Z} . No entanto, o algoritmo E-M é um estimador iterativo para a verossimilhança marginal. Ele funciona por meio da aplicação sucessiva de dois passos:

- 1 *Expectation* (E): calcula o valor esperado do logaritmo da verossimilhança com relação à distribuição condicional de \mathcal{Z} dado \mathcal{X} utilizando a estimativa atual dos parâmetros θ :

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{\mathcal{Z}|\mathcal{X},\theta^{(t)}} [\log L(\theta|\mathcal{X}, \mathcal{Z})], \quad (5)$$

em que Q denota uma "quantidade".

- 2 *Maximization* (M): encontra os parâmetros que maximizam essa quantidade Q :

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)}). \quad (6)$$

De maneira geral, a ideia do algoritmo E-M é executar os passos 1 e 2 de maneira iterativa até que $|\theta^{(t+1)} - \theta^{(t)}| < \epsilon$. **A variável latente \mathcal{Z} nos diz, basicamente, a pertinência de uma dada amostra $x \in \mathcal{X}$ pertencer à um determinado agrupamento (classe).**

Basicamente, temos as seguintes informações acerca de nosso problema:

- Observações em \mathcal{X} podem ser discretas ou contínuas (ex: podem ser oriundas de distribuições Gaussianas).
- Variáveis latentes em \mathcal{Z} são discretas (pseudo-rótulos).
- Os parâmetros em θ são contínuos.

Basicamente, os passos do E-M são os seguintes:

- ❶ Inicializar o conjunto de parâmetros θ de maneira aleatória.
 - Passo E: para cada amostra $x_i \in \mathcal{X}$ associamos uma pontuação γ_{ki} que denota a probabilidade dessa amostra pertencer ao agrupamento k .
 - Passo M: dadas as pontuações para todas as amostras, ajusto θ_k para cada Gaussiana (agrupamento) k utilizando utilizando uma função de verossimilhança marginal (Equação 4).
- ❷ Avaliamos a função do logaritmos da verossimilhança. Caso ela tenha estabilizado entre as iterações, podemos assumir que o método convergiu e interrompemos o processo treinamento.

No entanto, uma maneira melhor de inicializar μ_k e Σ_k é via algoritmo do k -Médias, em que $k = 1, 2, \dots, K$. Já os pesos w_k podem ser inicializados como $w_k = \frac{C_k}{|\mathcal{X}|}$, em que C_k corresponde ao número de amostras do agrupamento k .

Vamos às equações de atualização:

- Passo E: temos que as pontuações na iteração t podem ser calculadas como segue:

$$\gamma_{ki}^{(t)} = \frac{w_k p(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K w_j p(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \quad (7)$$

A equação assim equivale ao preenchimento de uma matrix $\gamma \in \mathbb{R}^{K \times |\mathcal{X}|}$.

- Passo M: as seguintes equações são então utilizadas para atualizar os valores dos parâmetros:

$$C_k = \sum_{i=1}^{\mathcal{X}} T_{ki}^{(t)}, \quad (8)$$

em que $\mathbf{T}^{(t)} \in \mathbb{R}^{K \times |\mathcal{X}|}$ é uma matriz da iteração t e definida da seguinte forma:

$$\mathbf{T}_{ki}^{(t)} = \begin{cases} 1 & \text{a amostra } \mathbf{x}_i \text{ pertence ao agrupamento } k \\ 0 & \text{caso contrário.} \end{cases} \quad (9)$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{C_k} \sum_{i=1}^{|\mathcal{X}|} \gamma_{ki}^{(t)} \mathbf{x}_i, \quad (10)$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{C_k} \sum_{i=1}^{|\mathcal{X}|} T_{ki}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T (\mathbf{x}_i - \boldsymbol{\mu}_k), \quad (11)$$

e

$$w_k^{(t+1)} = \frac{C_k}{|\mathcal{X}|}. \quad (12)$$

O logaritmo da verossimilhança é utilizado como critério de parada do algoritmo. Essa verossimilhança é uma medida de **confiança** que temos que os dados do conjunto de dados \mathcal{X} são gerados pelos parâmetros estimados:

$$\log p(\mathcal{X}|\boldsymbol{\theta}) = \sum_{i=1}^{|\mathcal{X}|} \log \left[\sum_{k=1}^K w_k p(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right]. \quad (13)$$

A equação acima deve ser calculada ao final de cada iteração do algoritmo E-M.