计算机应用数学-2025 秋

Lecture 1.9: 高斯混合模型与 EM 算法

2025.12.12

Lecturer: 宋骐 *Scribe:*

1 Gaussian Mixture Model

1.1 Mixture Model

MM is a probabilistic model that contains k sub-distribution among the overall distribution.

The k sub-model is the Hidden Variable of the Mixture model.

A MM can use any distribution.

GMM: a MM that uses Gaussian distribution as the sub-model.

Example:

X_i - height of all students.

$z_i \in \{1, 2\}$ 1: Boys, 2: Girls.

Distribution of X_i is not a simple normal (Gaussian) distribution, but a "Mixture" of two Gaussian distributions.

$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

$X \sim (1 - \Delta) \cdot X_1 + \Delta \cdot X_2$

$\Delta \in \{0, 1\}, Pr(\Delta = 1) = \alpha$

Define $\phi_{\theta_i}(X)$ as the normal density (正态密度)

the probabilistic distribution of GMM is

$P(X|\theta) = (1 - \alpha)\phi_{\theta_1}(X) + \alpha\phi_{\theta_2}(X)$

$\theta = (\alpha, \theta_1, \theta_2) = (\alpha, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

Similarly, for k-GMM

$P(X|\theta) = \sum_i^k \alpha_i \phi(X|\theta_i)$

$\theta = (\alpha, \mu, \sigma^2) = (\alpha_1, \cdots, \alpha_k, \mu_1, \cdots, \mu_k, \sigma_1^2, \cdots, \sigma_k^2)$

2 Estimating GMM parameters

2.1 Maximum Likelihood Estimation for Gaussian Model

Let's start with estimating the parameters of a single GM.

极大似然估计:

- 求似然函数，取对数.
- 对对数求导并使之为 0，解对数似然方程.

即为未知参数的最大似然估计值.

For a Gaussian distribution $\mathcal{X} \sim N(\mu, \sigma^2)$

Given n observable samples.

$$
\begin{aligned}
L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (\frac{1}{\sqrt{2\pi}\sigma})^n exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
ln\ L(\mu, \sigma^2) &= ln\left[(2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \\
&= ln\left[(2\pi\sigma^2)^{-\frac{n}{2}}\right] + ln\left[exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \\
&= -\frac{n}{2}ln(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

$$
\begin{cases}
\dfrac{\partial\, lnL}{\partial\, \mu} &= \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}(x_i-\mu) = 0 \\
\dfrac{\partial\, lnL}{\partial\, \sigma^2} &= -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(x_i-\mu)^2 = 0
\end{cases}
$$

$$\Downarrow$$

$$\begin{cases} \sum_{i=1}^{n} (x_i - \mu) = 0 \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \end{cases}$$

$$\therefore \ \mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

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

2.2 MLE for GMM

For GMM, we don't know a certain data sample belongs to which sub-model. We need to estimate α, μ, σ^2.

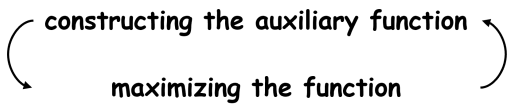
$$\begin{aligned} log\ L(\theta) &= \sum_{j=1}^{N} log[P(X_j|\theta)] \\ &= \sum_{j=1}^{N} log[\sum_{i}^{k} \alpha_i \phi(X|\theta_i)] \end{aligned}$$

$\log \sum$ is not fun, there is no way to reduce it !

Expectation-Maximization (EM) procedure can be used to handle $\log \sum$.

① Summarization of EM: It uses Jensen's inequality to create a lower bound (called an auxiliary function) for the likelihood that uses $\sum \log$ instead.

Repeat:



leading to a local maximum of the log likelihood for GMM.

Let's start with the overall process: assign each data point to a cluster, for each cluster, consider it as a Gaussian Model, calculate the θ.

- First, generate the cluster assignment.

$$z_i|w \sim Categorical(w)$$

w_k is the probability that i's cluster is k.

$$P(z_i = k|w) = w_k$$

We call w_k the mixture weights

$$\sum_k w_k = 1, 0 \leq w_k \leq 1$$

- Then, generate x_i from the cluster's distribution:

$$x_i|z_i = k \sim N(\mu_k, \sum_k)$$

$$\begin{aligned} \log likelihood(\theta) &= \log \prod_i \sum_k P(X_i = x_i|z_i = k, \theta)P(z_i = k|\theta) \\ &= \sum_i \log \sum_k P(X_i = x_i|z_i = k, \theta)P(z_i = k|\theta) \end{aligned}$$

z_i is the hidden variable, taking values $i = 1, \ldots, k$.

Similarly:

$$\begin{aligned} \log likelihood(\theta) &= \sum_i \log \sum_k P(X_i = x_i, z_i = k|\theta) \\ &= \sum_i \log \sum_k P(z_i = k|\theta)P(X_i = x_i|z_i = k, \theta) \\ &= \sum_i \log \sum_k P(z_i = k|x_i, \theta_t)\frac{P(X_i=x_i,z_i=k|\theta)}{P(z_i=k|x_i,\theta_t)} \\ &= \sum_i \log E_z \frac{P(X_i=x_i,z_i=k|\theta)}{P(z_i=k|x_i,\theta_t)} \end{aligned}$$

Lemma(Jensen's Inequality). If f is convex, then $f(EX) \leq E(f(X))$. If f is convex, $-f$ is concave, thus $-f(EX) \geq -E(f(X)) = E(-f(X))$.

Here, $-f(X) = \log(X)$ which is concave, thus $\log(EX) \geq E \log X$.

$$\begin{aligned} \log likelihood(\theta) &= \sum_i \log E_z \frac{P(X_i=x_i,z_i=k|\theta)}{P(z_i=k|x_i,\theta_t)} \\ &\geq \sum_i E_z \log \frac{P(X_i=x_i,z_i=k|\theta)}{P(z_i=k|x_i,\theta_t)} (Jensen'sInequality) \\ &= \sum_t \sum_k P(z_i = k|x_i, \theta_t) \log \frac{P(X_i=x_i,z_i=k|\theta)}{P(z_i=k|x_i,\theta_t)} \\ &=: A(\theta, \theta_t) \end{aligned}$$

- E-step: compute $P(z_i = k|x_i, \theta_t) =: \gamma_{ik}$ for each i, k
- M-step: $\max_\theta A(\theta, \theta_t) = \sum_i \sum_j \gamma_{ik} \log \frac{P(X_i=x_i, z_i=k|\theta)}{\gamma_{ik}}$, γ_{ik} doesn't depend on θ, thus, $\max_\theta \sum_i \sum_j \gamma_{ik} \log P(X_i = x_i, z_i = k|\theta)$.

② Back to GMM

- E-step: Using Bayes Rule

$$\begin{aligned} P(z_i = k|x_i, \theta_t) &= \frac{P(X_i=x_i|z_i=k,\theta_t)P(z_i=k|\theta_k)}{P(X_i=x_i|\theta_t)} \\ &= \frac{N(x_i;\mu_{kt},\sum_{kt})w_{kt}}{\sum_{k'} N(x_i;\mu_{k't},\sum_{k't})w_{k't}} =: \gamma_{ik} \end{aligned}$$

- M-step: $\max_\theta A(\theta, \theta_t) = \sum_i \sum_j \gamma_{ik} \log P(X_i = x_i, z_i = k|\theta)$

  Update θ, which is the collection $w, \mu, \sum$, by setting derivative of A to 0 with constraint $\sum_k w_k = 1$

  setting the derivatives to 0, we get the result for the cluster means:

$$\mu_{k,t+1} = \frac{\sum_i x_i \gamma_{ik}}{\sum_i \gamma_{ik}}$$

  Similarly, $\sum_{k,t+1} = \frac{\sum_i \gamma_{ik}(x_i-\mu_{k,t+1})(x_i-\mu_{k,t+1})^T}{\sum_i \gamma_{ik}}$

  Update for w is tricker beacuse of the constraint. We need to use Lagrangian to solve this constrained optimization problem.

  The Lagrangian:

$$L(\theta, \theta_t) = A(\theta, \theta_t) + \lambda(1 - \sum_k w_k)$$

Taking the derivative

$$\begin{aligned} \frac{\partial L(\theta,\theta_t)}{\partial w_{k'}} &= \frac{\partial A(\theta,\theta_t)}{\partial w_{k'}} - \lambda \\ &= \frac{\partial}{\partial w_{k'}}(\sum_i \sum_k \gamma_{ik} log P(X_i = X_i, Z_i = k|\theta)) - \lambda \end{aligned} \tag{1}$$

By the probabilistic model for generating data according to GMM:

$$\begin{aligned} P(X_i = X_i, Z_i = k|\theta) &= P(Z_i = k|w) \cdot P(X_i = x|Z_i = k, \mu_{k,t+1}, \sum_{k,t+1}) - \lambda \\ &= w_k \cdot N(X, \mu_{k,t+1}, \sum_{k,t+1}) \end{aligned}$$

plugging back to (1)

$$\begin{aligned}\frac{\partial L(\theta,\theta_t)}{\partial w_{k'}} &= \sum_i \frac{\partial}{\partial w_{k'}}[\gamma_{ik'} log[w_{k'} N(X;\mu_{k',t+1},\sum_{k',t+1})]] - \lambda \\ &= \sum_i \frac{\partial}{\partial w_{k'}}[\gamma_{ik'} log(w_{k',t+1})] + \frac{\partial}{\partial w_{k'}}[N(X,\mu_{k',t+1},\sum_{k',t+1})]\end{aligned}$$

Here,$N(X,\mu_{k',t+1},\sum_{k',t+1})$ dose not depend on $w_{k'}$

$$\begin{aligned}\frac{\partial L(\theta,\theta_t)}{\partial w_{k'}} &= \sum_i \frac{\partial}{\partial w_{k'}}[\gamma_{ik'} log(w_{k'})] - \lambda \\ &= \sum_i \gamma_{ik'} \frac{1}{w_{k'}} - \lambda \\ &= \frac{1}{w_{k'}} \sum_i \gamma_{ik'} - \lambda\end{aligned}$$

Setting the derivative to 0,

$$w_{k',t+1} = \frac{\sum_i \gamma_{ik'}}{\lambda}$$

We know $\sum_{k'} w_{k',t+1} = 1$, so λ is the normalization factor

$$\begin{aligned}\lambda = \sum_k \sum_i \gamma_{ik} &= \sum_i \underbrace{(\sum_k P(Z_i = k|x_i;\theta))}_{\text{=1 as it is the sum over the whole probability distribution.}} \\ &= \sum_i 1 \\ &= n\end{aligned}$$

Finally $w_{k',t+1} = \frac{\sum_i \gamma_{ik'}}{n}$

To sum:

- E-step : What is the current estimate of the probability

  that x_i comes from duster k ?

  $\rightarrow \gamma_{ik}$

- M-step : update $\mu, \sum, w$.

3 Multinomial Mixture Models and its estimation(多项式混合模型)

EM can be used for GMM, similar idea can be used for k-means.

Task: Document review for a law firm, we need to scan tens of thousands of emails to determine their relevance to a legal matter.

It could be helpful to pre-group them according to their topics.

Multinomial Mixture Model(MMM) would mdoel each topic as a distribution over English words.

I. MMM

Let $x_1, x_2, ..., x_N$ be per-document word count vectors, where n is the total number of documents.

$X_{iv} = m$ if v appears m times in document i.

We assume each document contains M words in total, $i.e.$ $\sum_v X_{iv} = M$.

The MMM models all documents as independent draws from the probability mass function

$$P(X) = \sum_{j=1}^{k} \pi_j f(x_j; \theta_j)$$

Each component

$$f(x_j; \theta_j) \propto \prod_{v=1}^{p} \theta_{jv}^{x_v}$$

is the pmf of a multinomial distribution, $Mult(\theta_j, M)$, where $\theta_j \in \mathbb{R}^p$

p is the total number of possible words,

θ_{jv} is the unknown probability of selecting word v for topic j.

II. EM for MMMs.

$$\begin{aligned}
\log p(X_{1:n}, Z_{1:n}; \pi, \theta_{1:k}) &= \sum_{i=1}^{n} \log p(X_i, Z_i; \pi, \theta_{1:k}) \\
&= \sum_{i=1}^{n} \log p(Z_i; \pi) + \sum_{i=1}^{n} \log p(X_i | Z_i; \theta_{i:k}) \\
&= \sum_{i=1}^{n} \log \pi_{Z_i} + \sum_{i=1}^{n} \sum_{v=1}^{p} x_{iv} \log \theta_{Z_i v} + \text{parameter-free terms} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} \prod_{Z_{i:j}} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{v=1}^{p} x_{iv} \prod_{Z_{i:j}} \log \theta_{jv} + \cdots
\end{aligned}$$

π_j is the unknown probability of selecting topic j, $\sum_{j=1}^{k} \pi_j = 1$

· E-step

$$\mathbb{E}_{p(\cdot|X_{1:n},\pi,\theta)} \left[\log p(X_{1:n}, Z_{1:n}; \pi_j, \theta_{1:k})\right] = \sum_{i=1}^{n} \sum_{j=1}^{k} \tau_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{v=1}^{p} x_{iv} \tau_{ij} \log \theta_{jv}$$

where $\tau_{ij} = p(Z_i = j \mid X_i; \pi, \theta_{1:k})$

By Bayes's Rule:

$p(Z_i = j \mid X_i; \pi, \theta_{1:k}) \propto p(Z_i = j; \pi) \cdot p(X_i \mid Z_i, \theta_{1:k})$

$\therefore\ p(Z_i = j \mid X_i, \pi, \theta_{1:k}) = \frac{\pi_j f(X_i;\theta_j)}{\sum_{l=1}^{k} \pi_l f(X_i;\theta_j)}$

· The M-step:

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \tau_{ij}$$

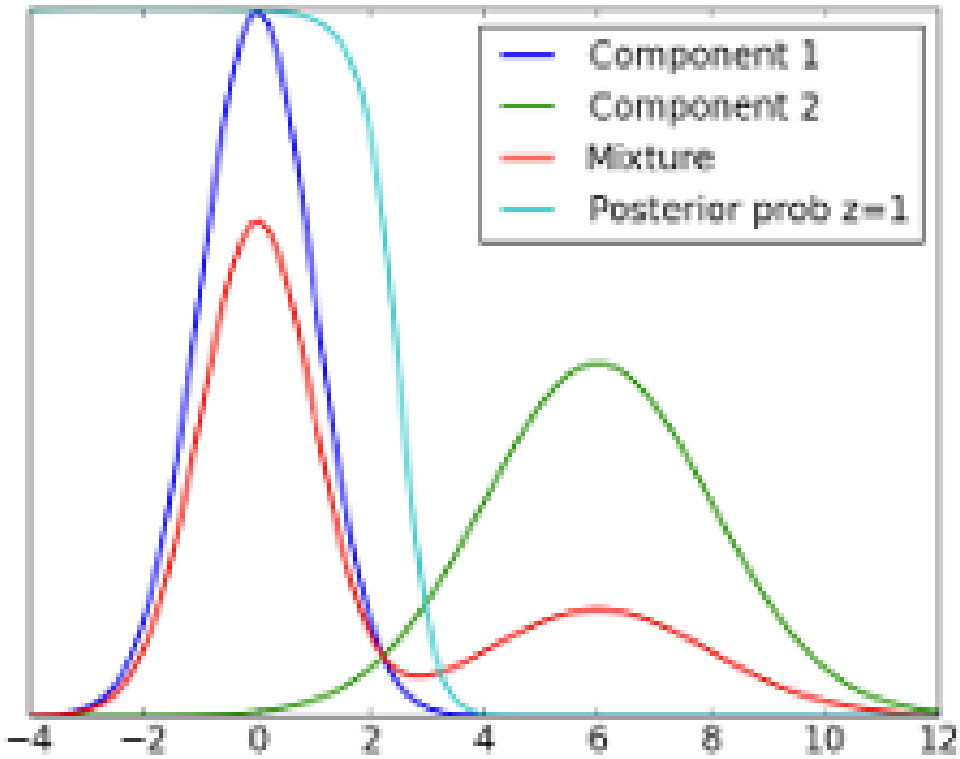
follow the similar steps of EM for GMM.

4 Model Inference * 后验概率推理

Assume we have the parameters for each sub-model. Given a data point, determine which sub-model it belongs to.

$p(z|x) \propto p(z)p(x|z)$

Example 1:



$x = 2$, with two GMs,
$Pr(z = 1) = 0.7$
$Pr(z = 2) = 0.3$
$p(x|z = 1) = Guassian(2; 0.1) \approx 0.054$
$p(x|z = 2) = Guassian(2; 6.2) \approx 0.027$
$p(z = 1|x) = \frac{Pr(z=1)p(x|z=1)}{Pr(z=1)p(x|z=1)+Pr(z=2)|p(x|z=2)} = 0.824$

5 Conclusion

Practical considerations for EM:

- GMMs are not appropriate for all data types (e.g., discrete data).

  We can use other models as the mixture model.

- The solutions of the EM algorithm are usually suboptimal.

  Initialization strategies can help.

- No single satisfying approach to choosing k.

  Methods we developed for k-means also apply to the GMM.

- We assume all data points are independent drawn from a common mixture.

  What if not? $\Rightarrow$ Hidden Markov models.