Expectation Maximization (EM)
期望最大

JhuoW

06-18 2022

1 最大似然估计 MLE

数据 $X=\{x_1,\cdots x_N\}$, 模型参数为 θ，Likelihood 定义为 $P(X|\theta)$：当参数为 θ 时，观测到给定数据 X 的概率。

$$P(X|\theta)=L(\theta|X)=P_\theta(X) \tag{1}$$

最大似然估计（Maximum Likelihood Estimation, MLE）：

$$\theta_{\mathrm{MLE}}=\arg\max_\theta P(X|\theta) \tag{2}$$

最大似然估计：给定一组样本 X，模型的参数 θ 是研究对象。若能找到参数 θ_{MLE}，使得样本发生的可能性最大，则此估计值 θ_{MLE} 为参数 θ 的最大似然估计。

举例来说，如果模型是单个 Gaussian Distribution 下，参数为 Gaussian Distribution 的参数（均值 μ, 标准差 Σ，$\theta=\{\mu,\Sigma\}$）. 给定一组数据 X, 要计算 X 来自什么样的 Gaussian，即：$P(\cdot|\theta)=f_\theta(\cdot)=\mathcal{N}(\cdot|\mu,\Sigma)$ 是一个 Gaussian Distribution 函数，目标为：

$$\theta_{\mathrm{MLE}}=\mu^\star,\Sigma^\star=\arg\max_{\mu,\Sigma}\sum_{i=1}^N\log\mathcal{N}(x_i|\mu,\Sigma) \tag{3}$$

即 MLE 的目标是找到最佳的高斯分布，是的从该分布中采样出数据 X 的概率最高。

如果只需要用一个 Gaussian 来拟合 X 的分布的话，这个 Gaussian 可以很容易用求导的方式获得 θ_{MLE} 的解析解：对 μ 求导：$\frac{\partial P(X|\mu,\Sigma)}{\partial\mu}$；对 Σ 求导：$\frac{\partial P(X|\mu,\Sigma)}{\partial\Sigma}$，令导数为 0，即可求得最佳的 μ，Σ，使得对应的高斯分布符合数据 $X=\{x_1,\cdots x_N\}$ 的分布。

但是，要用更复杂的模型（更多参数）来更准确的拟合 X 的分布，例如 Gaussian Mixture Model，即多个 Gaussian 的组合，其模型参数为：

$$\theta=\{\underbrace{\mu_1,\cdots,\mu_K}_{\text{每个 Gaussian 的 mean 参数}},\underbrace{\Sigma_1,\cdots,\Sigma_K}_{\text{每个 Gaussian 的 std 参数}},\underbrace{\alpha_1,\cdots,\alpha_{K-1}}_{\text{每个 Gaussian 的权重}}\} \tag{4}$$

假设是一个 K 个 Gaussian 的 Gaussian Mixture Model，那么 $\sum_{k=1}^{K} \alpha_k = 1$。

给定数据 $X = \{x_1, \cdots x_N\}$，若要用 K 维 Gaussian Mixture Model 来拟合该数据，就要优化所有 K 个 Gaussian 的均值参数，标准差参数，和权重参数，使得混合高斯分布采样出 X 的概率最大，即：

$$\begin{aligned} \theta_{\text{MLE}} &= \mu_1^\star, \cdots \mu_K^\star, \Sigma_1^\star, \cdots, \Sigma_K^\star, \alpha_1^\star, \cdots, \alpha_{K-1}^\star \\ &= \arg\max_\theta \sum_{i=1}^{N} \log \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \end{aligned} \tag{5}$$

如果要得到上式的解析解，要对 $\mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K, \alpha_1, \cdots, \alpha_{K-1}$ 求导，再令导数为 0 来求解，这非常困难，由此引出 EM 算法。

2 期望最大算法

求解 MLE 问题时，在最大化 log-likelihood:

$$\theta_{\text{MLE}} = \arg\max_\theta \log P(X|\theta) \tag{6}$$

难以直接对 θ 求导来得到解析解时（如高斯混合模型情况），可以使用 EM 算法来迭代求解：

$$\theta^{(t+1)} = \arg\max_\theta \int_z \log P(X, z|\theta) \cdot P(z|X, \theta^{(t)}) dz \tag{7}$$

X 为观测数据，z 为 latent variables（隐变量），隐变量必须不会影响 X 的边缘分布，即 $P(X) = \int_z P(X|z)P(z)dz$。

而公式7中

$$\begin{aligned} &\int_z \underbrace{\log P(X, z|\theta)}_{\text{每个 z 对应的值}} \cdot \underbrace{P(z|X, \theta^{(t)})}_{\text{z 的分布}} dz \\ =&\mathbb{E}_{P(z|X,\theta^{(t)})} [\log P(X, z|\theta)] \end{aligned} \tag{8}$$

所以，期望最大化算法求参数 θ 的迭代公式可改写为：

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_{P(z|X,\theta^{(t)})} [\log P(X, z|\theta)] \quad \text{期望最大化} \tag{9}$$

3 EM 算法收敛性证明

因为 EM 算法通过迭代的方式优化模型参数 θ,使得对数似然 $\log P(X|\theta)$ 最大。通过公式7,可以保证在 $\theta^{(t+1)}$ 参数下的模型比 $\theta^{(t)}$ 参数下的模型更拟合数据分布。通过公式7迭代更新参数 $\theta^{(t)} \to \theta^{(t+1)}$,可以使得 $\log P(X|\theta)$ 变大。收敛性即证明:

$$\log P(X|\theta^{(t)}) \leq \log P(X|\theta^{(t+1)}) \tag{10}$$

证明.

$$\begin{aligned}
&\because P(X, z) = P(z|X)P(X) \quad \text{always true, then} \quad P(X) = \frac{P(X, z)}{P(z|X)} \\
&\therefore P(X|\theta) = \frac{P(X, z|\theta)}{P(z|X, \theta)} \\
&\therefore \log P(X|\theta) = \log P(X, z|\theta) - \log P(z|X, \theta)
\end{aligned} \tag{11}$$

上式左右两边对分布 $P(z|X, \theta^{(t)})$ 求期望:

$$\mathbb{E}_{z\sim P(z|X,\theta^{(t)})} \underbrace{[\log P(X|\theta)]}_{\text{与 z 无关}} = \mathbb{E}_{z\sim P(z|X,\theta^{(t)})} [\log P(X, z|\theta) - \log P(z|X, \theta)] \tag{12}$$

上式左边 $= \log P(X|\theta)$,右边:

$$\begin{aligned}
&\mathbb{E}_{P(z|X,\theta^{(t)})} [\log P(X, z|\theta) - \log P(z|X, \theta)] \\
=& \underbrace{\int_z P(z|X, \theta^{(t)}) \log P(X, z|\theta) dz}_{Q(\theta,\theta^{(t)})} - \underbrace{\int_z P(z|X, \theta^{(t)}) \log P(z|X, \theta) dz}_{H(\theta,\theta^{(t)})}
\end{aligned} \tag{13}$$

注意到 $Q(\theta, \theta^{(t)}) = \int_z P(z|X, \theta^{(t)}) \log P(X, z|\theta) dz$ 就是 EM 算法的迭代更新函数,即 $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{(t)})$。结合公式12 和公式13:

$$\log P(X|\theta) = Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)}) \tag{14}$$

因此 log-likelihood under $\theta^{(t)}$ and $\theta^{(t+1)}$:

$$\begin{aligned}
\log P(X|\theta^{(t+1)}) &= Q(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t+1)}, \theta^{(t)}) \\
\log P(X|\theta^{(t)}) &= Q(\theta^{(t)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})
\end{aligned} \tag{15}$$

首先，根据 EM 的迭代求解公式，$\theta^{(t+1)}$ 由

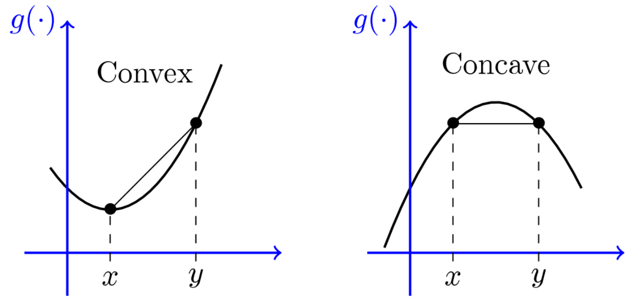
$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)}) \tag{16}$$

得到，所以 $Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta, \theta^{(t)})$ 一定成立。所以下式成立：

$$Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)}) \tag{17}$$

对于 $H(\theta, \theta^{(t+1)})$，首先介绍 Jensen Inequality:

Jensen Inequality:
If $g(x)$ is a convex function on R_X, and $\mathbb{E}[g(x)]$ and $g(\mathbb{E}[X])$ are finite, then $\mathbb{E}[g(x)] \geq g(\mathbb{E}[X])$。



显然 log 是 concave，所以 $\mathbb{E}[\log(\cdot)] \leq \log(\mathbb{E}[\cdot])$。同理 $-\log$ 是 convex，所以 $\mathbb{E}[-\log(\cdot)] \geq -\log(\mathbb{E}[\cdot])$。下面，计算 $H(\theta^{(t)}, \theta^{(t)}) - H(\theta, \theta^{(t)})$：

$$\begin{aligned}
&H(\theta, \theta^{(t)}) = \int_z P(z|X, \theta^{(t)}) \log P(z|X, \theta) dz \\
&H(\theta^{(t)}, \theta^{(t)}) - H(\theta, \theta^{(t)}) \\
=& \int_z P(z|X, \theta^{(t)}) \log P(z|X, \theta^{(t)}) dz - \int_z P(z|X, \theta^{(t)}) \log P(z|X, \theta) dz \\
=& \underbrace{\int_z P(z|X, \theta^{(t)}) \log \frac{P(z|X, \theta^{(t)})}{P(z|X, \theta)} dz}_{\mathrm{KL}(P(z|X,\theta^{(t)}) \| P(z|X,\theta))} \\
=& -\int_z P(z|X, \theta^{(t)}) \log \frac{P(z|X, \theta)}{P(z|X, \theta^{(t)})} dz \\
=& \mathbb{E}_{P(z|X,\theta^{(t)})} \left[-\log \frac{P(z|X, \theta)}{P(z|X, \theta^{(t)})} \right] \\
\geq& -\log \mathbb{E}_{P(z|X,\theta^{(t)})} \left[\frac{P(z|X, \theta)}{P(z|X, \theta^{(t)})} \right] = -\log \int_z \frac{P(z|X, \theta)}{P(z|X, \theta^{(t)})} \cdot P(z|X, \theta^{(t)}) dz \\
=& -\log \int_z P(z|X, \theta) dz = -\log 1 = 0
\end{aligned} \tag{18}$$

$$
\begin{aligned}
&\therefore H(\theta^{(t)},\theta^{(t)}) \geq H(\theta,\theta^{(t)}) \\
&\therefore H(\theta^{(t+1)},\theta^{(t)}) \leq H(\theta^{(t)},\theta^{(t)}) \\
&\because Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)}) \\
&\therefore Q(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \\
&\therefore \log P(X|\theta^{(t+1)}) \geq \log P(X|\theta^{(t)})
\end{aligned} \tag{19}
$$

所以通过 EM 算法的迭代得到新的 $\theta^{(t+1)}$ 增大 likelihood，使得模型更加拟合数据。 □

4 EM 算法公式推导

EM 算法 Maximize Likelihood Estimation 迭代公式：

$$
\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta} \int_z \log P(X,z|\theta) \cdot P(z|X,\theta^{(t)}) dz \\
&= \arg\max_{\theta} \mathbb{E}_{P(z|X,\theta^{(t)})}[\log P(X,z|\theta)]
\end{aligned} \tag{20}
$$

- E-Step：$\mathbb{E}_{P(z|X,\theta^{(t)})}[\log P(X,z|\theta)]$
- M-Step：$\arg\max_{\theta} \mathbb{E}_{P(z|X,\theta^{(t)})}[\log P(X,z|\theta)]$

其中，X 是观测数据，z 是隐变量，(X,z) 为完整数据，θ 为待优化模型参数，$P(\cdot|X)$ 为后验。

上一节通过收敛性证明，验证了上式每次迭代都朝着最大化 log-likelihood 的方向。本节推导 EM 的迭代公式。

公式11中得到：

$$
\log P(X|\theta) = \log P(X,z|\theta) - \log P(z|X,\theta) \tag{21}
$$

引入一个关于隐变量 z 的分布 $q(z)$，可以定义为任意关于 z 的非 0 分布。上式可以改写为：

$$
\log P(X|\theta) = \log \frac{P(X,z|\theta)}{q(z)} - \log \frac{P(z|X,\theta)}{q(z)} \tag{22}
$$

左右两边对 $q(z)$ 求期望。

$$
\text{左边} = \mathbb{E}_{q(z)} \log P(X|\theta) = \int_z q(z) \log P(X|\theta) dz = \log P(X|\theta) \underbrace{\int_z q(z) dz}_{=1} = \log P(X|\theta) \tag{23}
$$

$$
\begin{aligned}
\text{右边} &= \mathbb{E}_{q(z)}\left[\log\frac{P(X,z|\theta)}{q(z)} - \log\frac{P(z|X,\theta)}{q(z)}\right] \\
&= \underbrace{\int_z q(z)\log\frac{P(X,z|\theta)}{q(z)}dz}_{ELBO=\text{Evidence Lower Bound}} \underbrace{-\int_z q(z)\log\frac{P(z|X,\theta)}{q(z)}dz}_{\mathrm{KL}(q(z)||P(z|X,\theta))}
\end{aligned} \tag{24}
$$

所以

$$
\log P(X|\theta) = ELBO + \mathrm{KL}(q(z)||P(z|X,\theta)) \tag{25}
$$

其中 $P(z|X,\theta)$ 为后验（posterior）。而 $\mathrm{KL}(q(z)||P(z|X,\theta)) \geq 0$, 当分布 $q(z) = P(z|X,\theta)$ 时，等号成立。所以

$$
\log P(X|\theta) \geq ELBO = \int_z q(z)\log\frac{P(X,z|\theta)}{q(z)}dz \tag{26}
$$

因此，最大化 log-likelihood $\log P(X|\theta)$ 问题可以转化为最大化 $\log P(X|\theta)$ 的下界 ELBO，即：

$$
\hat{\theta} = \arg\max_\theta \log P(X|\theta) \Longleftrightarrow \hat{\theta} = \arg\max_\theta ELBO \tag{27}
$$

$$
\begin{aligned}
\hat{\theta} &= \arg\max_\theta ELBO \\
&= \arg\max_\theta \int_z q(z)\log\frac{P(X,z|\theta)}{q(z)}dz \quad \text{令关于}z\text{的分布}q(z) = P(z|X,\theta^{(t)}) \\
&= \arg\max_\theta \int_z P(z|X,\theta^{(t)})\left[\log P(X,z|\theta) - \underbrace{P(z|X,\theta^{(t)})}_{\text{与}\theta\text{无关，去掉不影响结果}}\right]dz \\
&= \arg\max_\theta \int_z P(z|X,\theta^{(t)})\log P(X,z|\theta)dz \\
&= \text{公式 (7)}
\end{aligned} \tag{28}
$$