EM 算法

与 Fisher 打分法类似，期望最大化算法（Expectation-Maximization Algorithm，简称 EM 算法）也是一种常用的通用算法，用于确定最大似然估计量。该算法适用于目标数据只有部分观测的情况。

在许多实际应用中，这种数据缺失的情况经常出现，该算法在这种情况下可以通过将观测数据视为“完全观测”的一部分来应用。

像往常一样，我们用 X 表示观测数据，但我们这里假设的意思是仅观测到 X，而不是“完全数据” (X, Y)，

理论上 (X, Y) 也可以获取。如果 $(x, y) \mapsto \bar{p}_\theta(x, y)$ 是向量 (X, Y) 的概率密度，那么我们可以通过边缘化得到 X 的密度：

$$p_\theta(x) = \int \bar{p}_\theta(x, y) dy$$

（在离散分布的观测值的情况下，我们将积分换成求和就行）。

我们利用基于观测 X 的 θ 的最大似然估计量最大化似然函数 $\theta \mapsto p_\theta(X)$。如果显示的方程中的积分可以显式计算，则求解最大似然估计量是一个标准问题，可以通过解析法或迭代算法解决。如果积分不能解析求解，则每个 θ 的似然函数的计算需要对积分进行数值近似，而找到最大似然估计量可能需要进行多次这样的近似。EM 算法就是试图规避这些近似计算。

如果我们有“完全数据” (X, Y) 可供使用，我们可以利用 (X, Y) 来确定最大似然估计量。该估计量通常比仅基于 X 的最大似然估计量更好，它是使对数似然函数 $\theta \mapsto \log \bar{p}_\theta(X, Y)$ 最大的点，并且这一过程可能容易些。

当 Y 不可用时，自然的做法是将该对数似然函数替换为其条件期望，也就是说只知道 X 时候的情况：

$$\theta \mapsto \mathrm{E}_{\theta_0} \left(\log \bar{p}_\theta(X, Y) \mid X \right) \tag{1}$$

这是给定观测 X 的完整数据对数似然的条件期望。这个思路是将通常的对数似然函数替换为函数 (1)，并确定使后者最大化的点。

不幸的是，公式 (1) 中的期望通常依赖于真实参数 θ_0，这就是为什么我们在期望算子 E_{θ_0} 的下标中包含了它。由于 θ 的真实值未知，因此不能将该公式作为估计方法的基础。

所以EM 算法通过迭代解决了这个问题。给定一个适当选择的 θ 真实值的初始猜测 $\tilde{\theta}_0$，我们通过最大化公式 (1) 中的函数来确定估计值 $\tilde{\theta}_1$。然后，我们用 $\tilde{\theta}_1$ 代替 $\mathrm{E}_{\tilde{\theta}_0}$，最大化新的函数，依此类推。

1. 初始化 $\tilde{\theta}_0$。
2. E 步骤：给定 $\tilde{\theta}_i$，确定函数 $\theta \mapsto \mathrm{E}_{\tilde{\theta}_i} \left(\log \bar{p}_\theta(X, Y) \mid X = x \right)$。
3. M 步骤：定义 $\tilde{\theta}_{i+1}$ 为该函数达到最大值的点。

EM 算法给出了一系列值 $\tilde{\theta}_0, \tilde{\theta}_1, \ldots$，并且我们希望随着 i 的增加，$\tilde{\theta}_i$ 逐渐接近未知的最大似然估计量。

前面的说明很容易给人一种印象，EM 算法的结果是一种新的估计量。其实不然，因为如果由 EM 算法生成的序列 $\tilde{\theta}_0, \tilde{\theta}_1, \ldots$ 收敛到一个极限，那么该极限正是基于观测值 X 的最大似然估计量。

实际上，在某些规则条件下，对于每一个 i，我们有

$$p_{\tilde{\theta}_{i+1}}(X) \geq p_{\tilde{\theta}_i}(X) \tag{2}$$

因此，EM 算法的迭代给出了观察值 X 的似然函数的不断增大的值。如果算法“按预期”工作，$p_{\tilde{\theta}_i}(X)$ 的值将最终增加到似然函数的最大值，$\tilde{\theta}_i$ 将收敛到最大似然估计量。然而，通常情况下，并不一定收敛，需要逐个情况进行研究。例如，序列 $\tilde{\theta}_i$ 可能会收敛到局部极大值。此外，执行算法的两个步骤并不一定容易。

上面结论的证明

EM 算法生成的序列 $\tilde{\theta}_0, \tilde{\theta}_1, \tilde{\theta}_2, \ldots$ 给出了一个逐渐增加的似然值序列 $p_{\tilde{\theta}_0}(X), p_{\tilde{\theta}_1}(X), p_{\tilde{\theta}_2}(X), \ldots$

证明:
(X, Y) 的密度 $\bar{p}_\theta$ 可以分解为

$$\bar{p}_\theta(x, y) = p_\theta^{Y|X}(y \mid x) p_\theta(x)$$

对数将该乘积转换为求和，因此我们有

$$\mathrm{E}_{\tilde{\theta}_i}\left(\log \bar{p}_\theta(X, Y) \mid X\right) = \mathrm{E}_{\tilde{\theta}_i}\left(\log p_\theta^{Y|X}(Y \mid X) \mid X\right) + \log p_\theta(X)$$

由于 $\tilde{\theta}_{i+1}$ 是使该函数关于 θ 取得最大值的点，因此该表达式在 $\theta = \tilde{\theta}_{i+1}$ 时大于在 $\theta = \tilde{\theta}_i$ 时的值，

$$\mathrm{E}_{\tilde{\theta}_i}\left(\log \bar{p}_{\tilde{\theta}_{i+1}}(X, Y) \mid X\right) \geq \mathrm{E}_{\tilde{\theta}_i}\left(\log \bar{p}_{\tilde{\theta}_i}(X, Y) \mid X\right)$$

如果我们能证明 $\mathrm{E}_{\tilde{\theta}_i}\left(\log p_\theta^{Y|X}(Y \mid X) \mid X\right)$ 在 $\theta = \tilde{\theta}_{i+1}$ 时比在 $\theta = \tilde{\theta}_i$ 时小，那么对于 $\log p_\theta(X)$，相反的情况(更大)必定成立（并且差异必须由第二项补偿），由此推得公式 (2) 成立。因此，证明以下不等式就足够了：

$$\mathrm{E}_{\tilde{\theta}_i}\left(\log p_{\tilde{\theta}_{i+1}}^{Y|X}(Y \mid X) \mid X\right) \leq \mathrm{E}_{\tilde{\theta}_i}\left(\log p_{\tilde{\theta}_i}^{Y|X}(Y \mid X) \mid X\right)$$

该不等式形式为 $\int \log(q/p) dP \leq 0$，其中 p 和 q 分别为参数 $\tilde{\theta}_i$ 和 $\tilde{\theta}_{i+1}$ 时 Y 在给定 X 情况下的条件密度，P 为对应于密度 p 的概率测度。由于对所有 $x \geq 0$ 都有 $\log x \leq x - 1$，因此对于任何概率密度 p 和 q，都满足

$$\int \log(q/p) dP \leq \int (q/p - 1) dP = \int_{p(x)>0} q(x) dx - 1 \leq 0$$

这证明了前面的公式，完成了证明。

例 混合分布

假设一些物体或个体可以原则上分成几个大致均匀的聚类。不幸的是，我们无法观察到聚类标签，而是为每个物体测量一个向量 x_i。我们希望根据观测值 $x_1, \ldots, x_n$ 来确定这些物体的聚类。

我们可以假设每个观测值 x_i 是随机向量 X_i 的一个实现，如果该物体属于第 j 个聚类，则 X_i 的概率密度为 f_j。我们可以将上段中的“较为均匀”理解为，不同聚类的概率密度 $f_1, \ldots, f_k$ 之间重叠较少。我们将假设聚类的数量 k 是已知的，尽管我们也可以从数据中推断出这个数量。

确定聚类的一种方法是最大化如下的似然函数

$$\prod_{j=1}^{k} \prod_{i \in I_j} f_j\left(X_i\right)$$

在所有关于 $\{1, \ldots, n\}$ 被分成 k 个子集的划分 $\left(I_1, \ldots, I_k\right)$ 以及密度 f_j 中未知参数上进行最大化。此划分将给出聚类。

例如，选择期望向量为 μ_j 的正态密度作为 f_j (也就是说把这些概率密度都建立成正态的)，将得到 k -means 聚类：最佳分类由最小化以下表达式的划分给出

$$\min_{\left(\mu_1, \ldots, \mu_k\right) \in \mathbb{R}^k} \sum_{j=1}^{k} \sum_{i \in I_j}\left\|X_i-\mu_j\right\|^2$$

从计算的角度来看，这是一个不简单的问题，但可以使用迭代算法近似得到聚类。

另一种方法是假设每个物体随机分配到一个聚类。我们可以引入一个随机向量 $\left(C_1, \ldots, C_n\right)$ 来表示聚类标签（如果第 i 个物体属于第 j 个聚类，则 $C_i=j$），并将密度 f_j 视为给定 $C_i=j$ 时 X_i 的条件概率密度。类向量 $\left(C_1, \ldots, C_n\right)$ 是不可观测的。如果我们假设 $\left(C_1, X_1\right), \ldots,\left(C_n, X_n\right)$ 是独立同分布的向量，并且对于所有 i，$j=1, \ldots, k$，有 $\mathrm{P}\left(C_i=j\right)=p_j$，那么我们可以使用 EM 算法确定参数 $p=\left(p_1, \ldots, p_k\right)$ 和 $f=\left(f_1, \ldots, f_k\right)$ 中未知参数的最大似然估计。

完整数据由 $\left(C_1, X_1\right), \ldots,\left(C_n, X_n\right)$ 组成。相应的似然函数可以写为

$$(p, f) \mapsto \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f_j\left(X_i\right) 1_{\{C_i=j\}}=\prod_{i=1}^{n} \prod_{j=1}^{k}\left(p_j f_j\left(X_i\right)\right)^{1_{\{C_i=j\}}}$$

因此，EM 算法的 E 步是计算

$$\begin{aligned}
&\mathrm{E}_{\tilde{p}, \tilde{f}}\left(\log \prod_{i=1}^{n} \prod_{j=1}^{k}\left(p_j f_j\left(X_i\right)\right)^{1_{\{C_i=j\}}} \mid X_1, \ldots, X_n\right) \\
&\quad=\sum_{i=1}^{n} \sum_{j=1}^{k} \mathrm{E}_{\tilde{p}, \tilde{f}}\left(\left(\log p_j+\log f_j\left(X_i\right)\right) 1_{\{} C_i=j\right\} \mid X_i\Big)
\end{aligned}$$

(第二个等号是因为, 显然 C_i 只和 X_i 有关)
使用贝叶斯公式，我们可以得到条件概率密度 $\mathrm{P}\left(C_i=j \mid X_i=x\right)=p_j f_j(x) / \sum_c p_c f_c(x)$。因此，最后显示的公式等于

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \log p_j \frac{\tilde{p}_j \tilde{f}_j\left(X_i\right)}{\sum_c \tilde{p}_c \tilde{f}_c\left(X_i\right)}+\sum_{j=1}^{k} \sum_{i=1}^{n} \log f_j\left(X_i\right) \frac{\tilde{p}_j \tilde{f}_j\left(X_i\right)}{\sum_c \tilde{p}_c \tilde{f}_c\left(X_i\right)}$$

在 EM 算法的 M 步中，我们对 p 和 f 进行最大化。对于 p，只有第一项起作用。利用微积分可以证明，当达到最大值时

$$p_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}_j \tilde{f}_j(X_i)}{\sum_c \tilde{p}_c \tilde{f}_c(X_i)}$$

对于 f 的最大化，只有第二项起作用。此外，我们可以分别对每一个 j 项单独最大化，如果参数 $f_1, \ldots, f_k$ 彼此独立变化：在这种情况下，f_j 最大化

$$f_j \mapsto \sum_{i=1}^{n} \log f_j(X_i) \frac{\tilde{p}_j \tilde{f}_j(X_i)}{\sum_c \tilde{p}_c \tilde{f}_c(X_i)}$$

例如，如果我们选择期望向量为 μ_j 的正态密度作为 f_j，那么 $\log f_j(x)$ 等于 $-\frac{1}{2}\|x - \mu_j\|^2$（常数项略去），并对 μ_j 进行最大化，我们得到

$$\mu_j = \frac{\sum_{i=1}^{n} \alpha_{ij} X_i}{\sum_{i=1}^{n} \alpha_{ij}}, \quad \alpha_{ij} = \frac{\tilde{p}_j \tilde{f}_j(X_i)}{\sum_c \tilde{p}_c \tilde{f}_c(X_i)}$$

这是观测值 X_i 的加权平均，其中权重等于条件概率 $\alpha_{ij} = \mathrm{P}_{\tilde{p},\tilde{f}}(C_i = j \mid X_i)$，对于 $1 \le i \le n$，使用当前近似 $(\tilde{p}, \tilde{f})$ 计算参数。然后我们反复迭代这些更新公式，直到结果几乎不再变化。

从参数的最大似然估计中，我们还可以推导出概率 $\mathrm{P}_{p,f}(C_i = j \mid X_i)$ 的最大似然估计，即第 i 个物体属于第 j 个聚类的概率。我们可以将物体分配到该概率最大的聚类中。