Standford CS229 2022Fall，第12.1讲：高斯混合模型与 EM 算法

高斯混合模型与 EM 算法

在本组笔记中，我们讨论用于密度估计的 EM（期望最大化）算法。

假设我们像往常一样获得了一个训练集 $\{x^{(1)}, \ldots, x^{(m)}\}$。由于我们处于无监督学习的设定下，这些点没有附带任何标签。我们希望通过指定一个联合分布 $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$ 来对数据进行建模。这里，$z^{(i)} \sim \text{Multinomial}(\phi)$（其中 $\phi_j \geq 0$，$\sum_{j=1}^{k} \phi_j = 1$，参数 ϕ_j 给出了 $p(z^{(i)} = j)$），并且 $x^{(i)}|z^{(i)} = j \sim N(\mu_j, \Sigma_j)$。我们用 k 表示 $z^{(i)}$ 可以取值的个数。因此，我们的模型假设每个 $x^{(i)}$ 是通过从 $\{1, \ldots, k\}$ 中随机选择 $z^{(i)}$，然后根据 $z^{(i)}$ 从 k 个高斯分布中的一个抽取得到的。这被称为高斯混合模型。同时请注意，$z^{(i)}$ 是潜在的随机变量，意味着它们是隐藏/未观测的。这将使我们的估计问题变得困难。

我们模型的参数是 ϕ、μ 和 Σ。为了估计它们，我们可以写出数据的似然函数：

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}; \phi, \mu, \Sigma) = \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma)p(z^{(i)}; \phi).$$

然而，如果我们令该公式的导数等于零并尝试求解，我们会发现无法以闭合形式找到参数的最大似然估计。（您可以在家自己尝试一下。）

随机变量 $z^{(i)}$ 表明每个 $x^{(i)}$ 来自于哪个高斯分布。请注意，如果我们知道 $z^{(i)}$ 的值，那么最大似然问题就会变得简单。具体来说，我们可以将似然函数写成：

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi).$$

对 ϕ、μ 和 Σ 最大化此式，可得参数：

$$\phi_j = \frac{1}{m} \sum_{i=1}^{m} 1\{z^{(i)} = j\},$$

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}},$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}.$$

事实上，我们看到，如果 $z^{(i)}$ 的值已知，那么最大似然估计几乎与我们在估计高斯判别分析模型参数时所得到的结果相同，只是这里的 $z^{(i)}$ 扮演了类别标签的角色。[1]

然而，在我们的密度估计问题中，$z^{(i)}$ 的值是未知的。我们该怎么办？

EM 算法是一种迭代算法，包含两个主要步骤。应用于我们的问题时，在 E 步骤中，它试图“猜测” $z^{(i)}$ 的值；在 M 步骤中，它根据我们的猜测更新模型的参数。由于在 M 步骤中我们假装第一步的猜测是正确的，因此最大化过程变得简单。以下是该算法：

重复直到收敛：
{
(E 步骤) 对于每个 i, j，设置

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

(M 步骤) 更新参数：

$$\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)},$$

$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

}

在 E 步骤中，我们计算给定 $x^{(i)}$ 和当前参数设置下，$z^{(i)}$ 的后验概率。即，使用贝叶斯规则，我们得到：

$$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

这里，$p(x^{(i)} | z^{(i)} = j; \mu, \Sigma)$ 是通过在 $x^{(i)}$ 处计算均值为 μ_j、协方差为 Σ_j 的高斯分布密度得到的；$p(z^{(i)} = j; \phi)$ 由 ϕ_j 给出，等等。在 E 步骤中计算出的值 $w_j^{(i)}$ 代表了我们对 $z^{(i)}$ 值的“软”猜测[2]。

此外，您应该将 M 步骤中的更新与当 $z^{(i)}$ 完全已知时我们得到的公式进行对比。它们是相同的，只是现在我们用 $w_j^{(i)}$ 代替了指示函数 “$1\{z^{(i)} = j\}$”，后者表示每个数据点来自哪个高斯分布。

EM 算法也让人联想到 K-means 聚类算法，不同之处在于，我们不再使用“硬”聚类分配 $c^{(i)}$，而是使用“软”分配 $w_j^{(i)}$。与 K-means 类似，它也容易陷入局部最优，因此多次使用不同的初始参数重新初始化可能是个好主意。

很明显，EM 算法有一个非常自然的解释，即反复尝试猜测未知的 $z^{(i)}$；但它是如何产生的？我们能否对其做出任何保证，例如关于其收敛性？

在下一组笔记中，我们将描述 EM 的一个更通用的视角，这将使我们能够轻松地将其应用于其他也存在潜在变量的估计问题，并允许我们给出收敛性保证。

注释：

[1] 这里与我们在 PS1 中用高斯判别分析得到的公式有其他一些细微差别，首先是因为我们将 $z^{(i)}$ 推广为多项分布而非伯努利分布，其次是因为这里我们为每个高斯分布使用了不同的 Σ_j。

[2] “软”一词指的是我们的猜测是概率值，取值范围在 $[0, 1]$ 内；相比之下，“硬”猜测是指单一的最佳猜测（例如取值在 $\{0, 1\}$ 或 $\{1, \ldots, k\}$ 中）。