一、阅读范围

《统计学习方法》第九章 EM 算法及其推广

二、负责讲解部分主要知识梳理

9.1.1 例子（三硬币模型）

第九章第一节主要介绍的就是 EM 算法，EM 算法是一种迭代算法，用于含有隐变量的概率模型参数的极大似然估计，或极大后验概率估计。每次迭代过程包括 E（expectation），求期望，和 M（maximization），求极大，两步组成。

当概率模型中变量全是观测变量，给定数据我们可以直接用极大似然估计法，或贝叶斯估计法估计模型参数。但是当模型含有隐变量时，就不能简单地直接使用这些估计方法。这就是为什么要提出 EM 算法的原因。

9.1.1 中例子的作用是介绍与引入 EM 算法。

例 9.1（三硬币模型） 假设有 3 枚硬币，分别记作 A，B，C。这些硬币正面出现的概率分别是π，p和q。进行如下掷硬币试验：先掷硬币 A，根据其结果选出硬币 B 或硬币 C，正面选硬币 B，反面选硬币 C；然后掷选出的硬币，掷硬币的结果，出现正面记作 1，出现反面记作 0；独立地重复n次试验（这里，$n=10$），观测结果如下：

$$1,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1,\ 1$$

假设只能观测到掷硬币的结果，不能观测掷硬币的过程。问如何估计三硬币正面出现的概率，即三硬币模型的参数。

三硬币模型可以写作

$$P(y|\theta)=\sum_z P(y,z|\theta)=\sum_z P(z|\theta)P(y|z,\theta)$$
$$=\pi P^y(1-P)^{1-y}+(1-\pi)q^y(1-q)^{1-y}$$

这里，随机变量y是观测变量，表示一次实验观测的结果是 1 或 0；随机变量z是隐变量，表示为观测到的掷硬币 A 的结果：$\theta=(\pi,p,q)$是模型参数。这一模型是以上数据的生成模型。也就是说，随机变量y的数据可以观测，随机变量z的数据不可观测。

进一步地，给出观测数据（$Y=(Y_1,Y_2,\cdots,Y_n)^T$）的似然函数：

$$P(Y|\theta)=\prod_{j=1}^{n}[\pi p^{y_j}(1-p)^{1-y_j}+(1-\pi)q^{y_j}(1-q)^{1-y_i}]$$

考虑求模型$\theta=(\pi,p,q)$的极大似然估计，即

$$\hat{\theta}=\arg\max_\theta \log_P(Y|\theta)$$

EM 算法首先选取参数的初值，记作$\theta^{(0)}=\left(\pi^{(0)},p^{(0)},q^{(0)}\right)$，然后通过 E，M 步迭代计算参数的估计值，直到收敛为止。第i次迭代参数的估计值为$\theta^{(i)}=\left(\pi^{(i)},p^{(i)},q^{(i)}\right)$。

EM 算法的第$i+1$次迭代如下。

E 步：计算在模型参数$\pi^{(i)},p^{(i)},q^{(i)}$下观测数据$y_j$来自掷硬币 B 的概率

$$\mu_j^{(i+1)}=\frac{\pi^{(i)}\left(p^{(i)}\right)^{y_j}\left(1-p^{(i)}\right)^{1-y_j}}{\pi^{(i)}(p^{(i)})^{y_j}(1-p^{(i)})^{1-y_j}+(1-\pi^{(i)})(q^{(i)})^{y_j}(1-q^{(i)})^{1-y_j}}$$

M 步：计算模型参数的新估计值：

$$\pi^{(i+1)} = \frac{1}{n}\sum_{j=1}^{n}\mu_j^{(i+1)}$$

$$p^{(i+1)} = \frac{\sum_{j=1}^{n}\mu_j^{(i+1)}y_j}{\sum_{j=1}^{n}\mu_j^{(i+1)}}$$

$$q^{(i+1)} = \frac{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})y_j}{\sum_{j=1}^{n}(1-\mu_j^{(i+1)})}$$

注：最终的结果与初值$\theta^{(0)} = \left(\pi^{(0)}, p^{(0)}, q^{(0)}\right)$的选取有关。

三、下周预计读书进度

第十章 隐马尔可夫模型