斯坦福 ML 公开课笔记 12

本文对应斯坦福 ML 公开课的第 12 个视频，第 12 个视频与前面相关性并不大，开启了一个新的话题——无监督学习。主要内容包括无监督学习中的 K 均值聚类(K-means)算法，混合高斯分布模型(Mixture of Gaussians, MoG)，求解 MoG 模型的 EM 算法，以及 EM 的一般化形式，在 EM 的一般化形式之前，还有一个小知识点，即 Jensen 不等式(Jensen's inequality)。

K-Means 算法

在之前的算法和模型中，训练数据都是带有标记的，这样的算法是有监督学习。当训练数据没有标记时，成为无监督学习。聚类算法就是无监督学习最常见的一种，给定一组数据，需要聚类算法去发掘数据中的隐藏结构。

聚类算法应用很广。举例来说，对基因进行聚类，可以发掘不同物种中具有相同功能的基因片段；对顾客行为进行聚类可以把市场分为不同的几个部分，针对不同的顾客可以采用不同的促销策略；在 google 的新闻首页，对新闻进行聚类，使得描述同一事件的报道不全部展示；在图片分割中，可以利用图片不同部分的相似性来理解图片信息等。

下面对 K-Means 算法的流程进行介绍，给定输入数据为$S = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$，K-Means 算法如下：

1) 选择初始的 k 个聚类中心$\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$
2) 对每个样本数据来说，将其类别标号设为距离其最近的聚类中心的标号，即

$$\text{label}^{(i)} = arg\min_j ||x^{(i)} - \mu_j|| \tag{1}$$

3) 将每个聚类中心的值更新为与该类别中心相同类别的所有样本的平均值，即

$$\mu_j := \frac{\sum_{i=1}^{m} I\{label^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} I\{label^{(i)} = j\}} \tag{2}$$

4) 重复第 2 步和第 3 步，直到聚类中心的变化低于阈值为止

对于 K-Means 来说，它要优化的目标函数可以看成如下形式：

$$\text{J(label, μ)} = \sum_{i=1}^{m} ||x^{(i)} - \mu_{label^{(i)}}||^2 \tag{3}$$

可以将 K-Means 算法看做是目标函数 J 的坐标下降过程，在第 2 步，我们保持聚类中心不变，将样本类别设为距离最近的中心的类别，此时修改了类别的样本的的目标函数项会变小，即$\sum_{\text{修改类别值的样本}} ||x - \mu||^2$值变小，而没有修改类别的样本值不变，从而整体变小。在第 3 步中，更新了聚类中心点的值，这样使得对每个类别来说，其目标函数项会变小，即$\sum_{\text{属于某类的样本}} ||x - \mu||^2$变小，从而整体变小。

在 K-Means 算法中，如何选择初始的聚类中心数目 k 是一个普遍的问题。有很多自动选择聚类中心的算法，但不在本文的范围内。

由于公式 3 不是一个凸函数，因而 K-Means 算法能保证收敛到一个局部极值，不能保证收敛到全局极值最优值。一个较为简单的解决方法是随机初始化多次，以最优的聚类结果为最终结果。

在聚类结束后，如果一个中心没有得到任何样本，那么需要去除这个中心点，或者重新初始化。

聚类算法可用于离群点检测，离群点检测应用也很普遍，比如飞机零件的评测，信用卡

消费行为异常监控等。

混合高斯分布

混合高斯分布(MoG)也是一种无监督学习算法，常用于聚类。当聚类问题中各个类别的尺寸不同、聚类间有相关关系的时候，往往使用 MoG 更合适。对一个样本来说，MoG 得到的是其属于各个类的概率(通过计算后验概率得到)，而不是完全的属于某个类，这种聚类方法被成为软聚类。一般说来，任意形状的概率分布都可以用多个高斯分布函数去近似，因而，MoG 的应用也比较广泛。

先举一个直观上的例子来帮助大家理解。如下图所示：

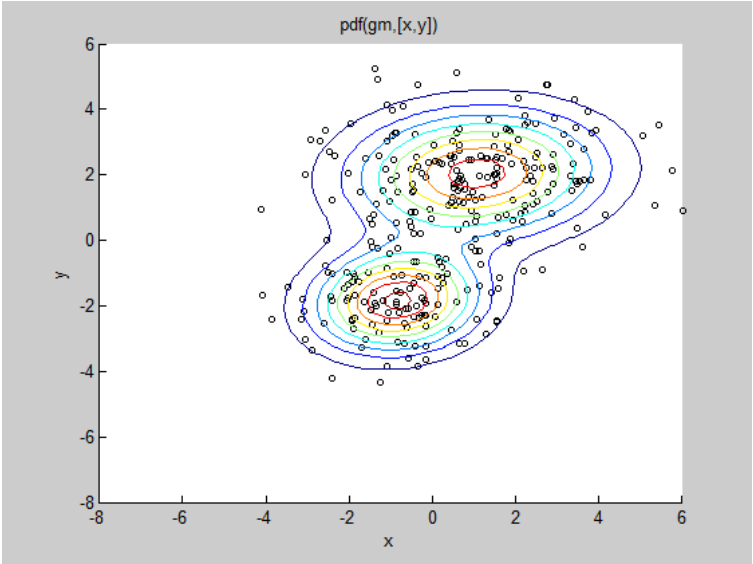

图 1 二维混合高斯分布

由图 1 可知，数据点由均值为(-1,-2)和(1,2)的两个高斯分布生成。根据数据点属于两个高斯分布的后验概率大小对数据点进行分类，可得图 2 所示的聚类结果。

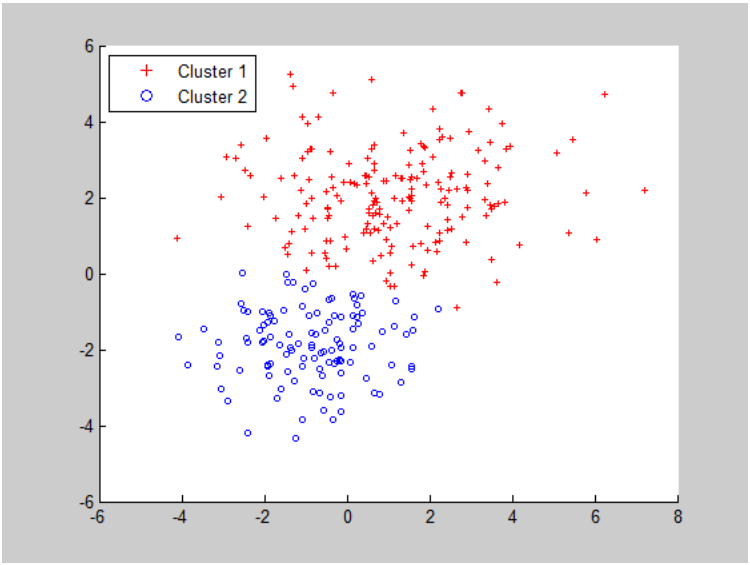

图 2 二维混合高斯分布聚类结果

下面，我们形式化的对 MoG 进行解释。首先，对问题进行形式化。

在 MoG 问题中，数据属于哪个分布可以看成是一个隐含变量 z。则 MoG 模型存在两个假设

假设一：z 服从多项式分布，即：

$$z^{(i)} \sim Multinomial(\phi) \tag{4}$$

对于多项式分布的参数，需要服从$\sum_j \phi_j = 1$。特殊的，当只有两个分布时，z 服从伯努利分布(两点分布)。

假设二：已知 z 时，x 服从正态分布，即条件概率 p(x|z)服从正态分布，即：

$$\mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right) \sim \mathrm{N}(\mu_j, \Sigma_j) \tag{5}$$

则 x 与 z 的联合分布概率函数为

$$\mathrm{p}\left(x^{(i)}, z^{(i)}\right) = \mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right) * p(z^{(i)}) \tag{6}$$

在 MoG 模型中，若$z^{(i)}$已知，那么就与 GDA(高斯判别分析，见笔记 5)一样了。这时，可以写出其似然函数：

$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} logp\left(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma\right)$$

$$= \sum_{i=1}^{m} [log\mathrm{p}\left(x^{(i)}\middle|z^{(i)}; \mu, \Sigma\right) + logp(z^{(i)}; \phi)] \tag{7}$$

因而，极大似然估计的结果为：

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} I\{z^{(i)} = j\} \tag{8}$$

$$\mu_j = \frac{\sum_{i=1}^{m} I\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} I\{z^{(i)} = j\}} \tag{9}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} I\{z^{(i)} = j\}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} I\{z^{(i)} = j\}} \tag{10}$$

公式 8、9、10 与笔记 5 中 GDA 的结论一致。

EM 算法求解 MoG

但是，我们现在并不知道$z^{(i)}$的值，因而，需要使用 EM 算法进行迭代估计出$z^{(i)}$从而得到参数。EM 算法的基本思想如下：

1) 设置初始参数θ；例如 MoG 中的ϕ, μ, Σ；
2) E-step：根据当前参数与观测数据 X，估计隐含变量 z 的分布；
3) M-step：根据 z 的分布，对参数进行重新估计；
4) 第 2 步和第 3 步反复进行，直到参数变化小于阈值或者目标函数的变化小于阈值为止。

具体说来，在 E-step 中，z 的概率分布的更新公式如下：

$$w_j^{(i)} = p\left(z^{(i)} = j\middle|x^{(i)}, \phi, \mu, \Sigma\right)$$

$$= \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_k p(x^{(i)}|z^{(i)} = k; \mu, \Sigma)p(z^{(i)} = k; \phi)} \tag{11}$$

其中，如假设中所言，$p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)$是正态分布，$p(z^{(i)} = j; \phi)$是多项式分布，将密度函数代入，即可得到给定观察值 x，z 的条件概率。

在 M-step 中，根据 E-step 中得到的 z 的分布，重新对参数进行估计。有

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)} \tag{12}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \tag{13}$$

$$\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}} \tag{14}$$

分别对比公式 8、9、10 与 12、13、14，发现 12、13、14 中只是将 8、9、10 中的指示函数替换为了 E-step 中的概率值。这说明在 MoG 的 EM 训练过程中，并不硬性的规定一个样本只能属于一个类，而是以概率的形式表达样本与类别的关系。当然，在训练中止的时候，大部分 z 的概率值都会接近于 0 或者 1，只有少部分会比较中和，这样使得 MoG 模型对不确定的样本处理的更好。

与 GDA 不同的是，在 MoG 中，不同的高斯分布所采用的协方差矩阵可以是不一样的。而 GDA 中则是一样的。

Jensen 不等式

上述解决 MoG 参数估计的 EM 算法只是 EM 算法的特例。EM 还有它的一般化形式。

在叙述 EM 的一般化形式之前，首先要先叙述一个会用到的定理，即 Jensen 不等式。

定理：

若 f 为凸函数，即$f''(x) \geq 0$，注意，并不要求 f 一定可导，但若存在二阶导数，则必须恒大于等于 0。再令 x 为随机变量，则存在不等式

$$f(E[X]) \leq E[f(x)] \tag{15}$$

进一步，若二阶导数恒大于 0，则不等式等号成立当且仅当 x=E[x]，即 x 是固定值。

若二阶导数的不等号方向逆转，则不等式的不等号方向逆转。

为使便于理解，看图 3.

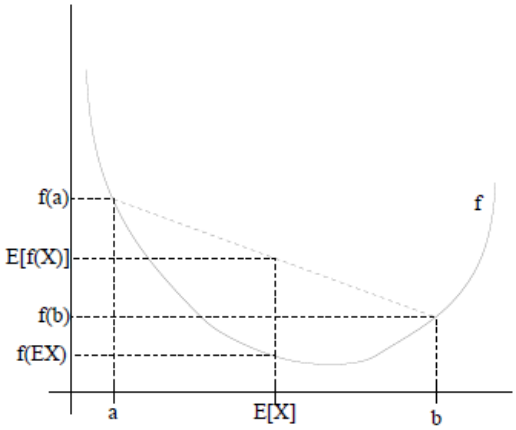

图 3 Jensen 不等式举例说明

EM 算法的一般化形式

在有隐含变量的模型中，它的似然函数为

$$\ell(\theta) = \sum_{i=1}^{m} log p(x^{(i)}; \theta) = \sum_{i=1}^{m} log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \tag{16}$$

注意，这里我们开始讨论 EM 的一般化形式，所以与在 MoG 中讨论 EM 不同，我们不

知道$p(x^{(i)}|z^{(i)})$与$p(z^{(i)})$所服从的具体分布，我们只要知道它们都是一种概率分布就足够了。

如果直接在公式 16 中应用极大似然估计，因为在对数函数中有连加，因而求偏导时会异常麻烦，导致参数估计十分困难。

我们可以对公式 16 进行处理。处理过程如下：

$$\max_{\theta}\sum_{i} logp(x^{(i)};\theta)=\max_{\theta}\sum_{i=1}^{m} log\sum_{z^{(i)}} p(x^{(i)},z^{(i)};\theta)$$

$$=\max_{\theta}\sum_{i=1}^{m} log\sum_{z^{(i)}} Q_i(z^{(i)})\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{17}$$

这里，Q 是一种概率分布，即$Q_i(z^{(i)})\geq 0$，$\sum_{z^{(i)}} Q_i(z^{(i)})=1$，稍后会叙述如何选择 Q 的概率分布。

继续推导有：

$$\sum_{i=1}^{m} log\sum_{z^{(i)}} Q_i(z^{(i)})\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}=\sum_{i=1}^{m} logE\left[\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\right]$$

$$\geq\sum_{i=1}^{m} E\left[log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\right]=\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{18}$$

在公式 18 中，用到了 Jensen 不等式，不过 log 函数是凹函数，所以不等号逆转了。另外还用到了期望的定义：若 x~p(x),则 E[g(x)]=Σp(x)g(x)。

由公式 16、17、18 可知，

$$\ell(\theta)\geq\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}=\text{lowbound}(\theta) \tag{19}$$

这时我们就找出了似然函数的一个下界，可以看到，该下界已经将取对数放到求和里面了，因而对其求偏导较为简单。假设当前的参数为$\theta^{(t)}$，在下界上进行极大似然估计后得到新参数$\theta^{(t+1)}$，如果能保证$\ell(\theta^{(t+1)})\geq\ell(\theta^{(t)})$，那么我们就可以在下界函数上进行极大似然估计就可以了。

如何能保证这一点呢？只要我们能在当前参数$\theta^{(t)}$处，使公式 17 的等号成立就行了。证明如下：

$$\ell(\theta^{(t+1)})\geq\text{lowbound}(\theta^{(t+1)})\geq\text{lowbound}(\theta^{(t)})=\ell(\theta^{(t)}) \tag{20}$$

第一个不等号意为下界函数，第二个不等号意为在下界函数上做极大似然估计，第三个等号是我们的假设。

如何使公式 17 中等号成立呢？回顾 Jensen 不等式中令等号成立的条件，只要使 x=E[x]即可，在公式 17 中即意味着使

$$\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}=\text{constant} \tag{21}$$

如此，则加上$\sum_{z^{(i)}} Q_i(z^{(i)})=1$的条件，我们就可以这样选择 Q：

$$Q_i(z^{(i)})=\frac{p(x^{(i)},z^{(i)};\theta)}{\sum_{z^{(i)}} p(x^{(i)},z^{(i)};\theta)}=p(z^{(i)}|x^{(i)};\theta) \tag{22}$$

有没有觉得公式 22 很眼熟呢？回顾公式 11，我们发现 Q 的设置与 MoG 的 E-step 的公式很相似。

公式 22 中的 Q 即为对 z 的概率估计。

由以上分析，我们就得到了 EM 算法的一般化形式。一般化形式的思想是，在 E-step，找到对于当前参数θ，使公式 19 等号成立的 Q 分布；在 M-step，对似然函数下界进行极大似然估计，得到新的参数。形式化表述为：

E-step:

$$Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\theta) \tag{23}$$

M-step:

$$\theta := \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{24}$$

为了便于理解，这里以一幅图来对 EM 算法进行总结。

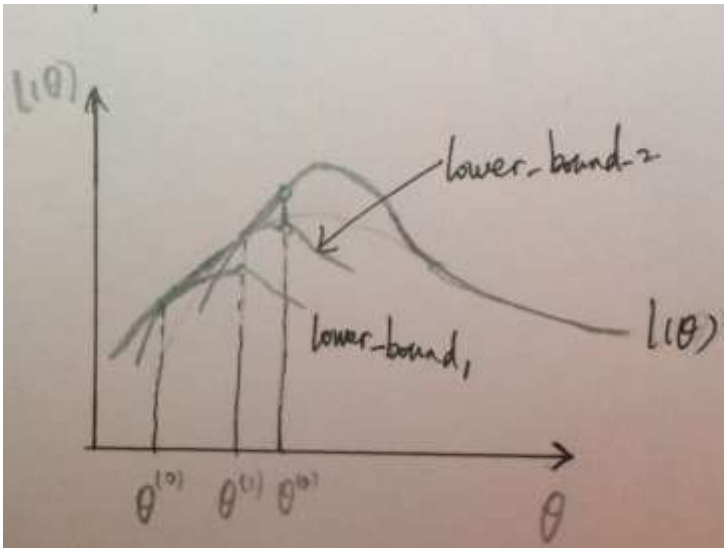

图 4 EM 基本思想

图 4 中所展现的内容就是我们刚才所述主要思想，存在一个我们不能直接进行求导的似然函数，给定初始参数，我们找到在初始参数下紧挨着似然函数的下界函数，在下界上求极值来更新参数。然后以更新后的参数为初始值再次进行如上操作，这就是 EM 进行参数估计的方法。

当然似然函数不一定是如图 4 中那样只有一个极值点，因而 EM 算法也有可能只求出局部极值。当然，可以如 K-Means 那样多次选择初始参数进行求，然后取最优的参数。

其实，在 EM 的一般化形式中，可以将目标函数看做是

$$J(Q,\theta) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \tag{25}$$

这样，EM 算法就可以看做是对目标函数的坐标上升过程，在 E-step 中，θ不变，调整 Q 使函数变大；在 M-step 中，Q 不变，调整θ使目标函数变大。