混合模型和 EM

陈雷慧
leihuichen@gmail.com
2019 年 4 月 18 日

目录

前言

混合模型

混合模型/隐变量模型

混合模型是一种常用的数据建模方法。它以简单的模型为基础模块，通过合理的引入隐藏变量，将基础模块组合在一起形成具有丰富表达能力的“组合式”模型。

- 连续隐变量模型：引入的隐变量是连续的
- 离散隐变量模型：引入的隐变量是离散的
- 常见的混合模型：隐马尔科夫模型、K 均值聚类算法、混合高斯分布、线性动态系统

讨论隐变量模型的思想和本质，以及离散隐变量模型的最大似然的求解方法——EM 算法的算法原理

K 均值聚类算法

K 均值聚类算法

问题描述

给定包含 N 个数据点的集合 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$，它是由 D 维欧式空间中的随机变量 $\boldsymbol{x}$ 的 N 次观测组成。

- 目标：将这些点分成 K 组，组内点之间的“距离”更近，组间点之间的“距离”更远
- 注意：K 值是事先指定的

形式化定义

引入一组 D 维向量 $\boldsymbol{\mu}_k$，其中 $k = 1, \cdots, K$。且 $\boldsymbol{\mu}_k$ 是与第 k 个聚类关联的一个代表。例如，将 $\boldsymbol{\mu}_k$ 视为聚类的中心。

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2.$$

其中，r_{nk} 指示了第 n 个数据点是否被分配到了第 k 类。

K 均值聚类算法

二阶段优化

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2.$$

首先，为 $\boldsymbol{\mu}_k$ 选择一些初始值。

- 第一阶段：关于 r_{nk} 最小化 J，保持 $\boldsymbol{\mu}_k$ 固定。J 是 r_{nk} 的一个线性函数，则 $r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2 \\ 0, & \text{otherwise} \end{cases}$
- 第二阶段：关于 $\boldsymbol{\mu}_k$ 最小化 J，保持 r_{nk} 固定。J 是 $\boldsymbol{\mu}_k$ 的二次函数，对其求导，并令导数为 0:

$$2 \sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0$$

可得 $\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \boldsymbol{x}_n}{\sum_n r_{nk}}$

K 均值聚类算法

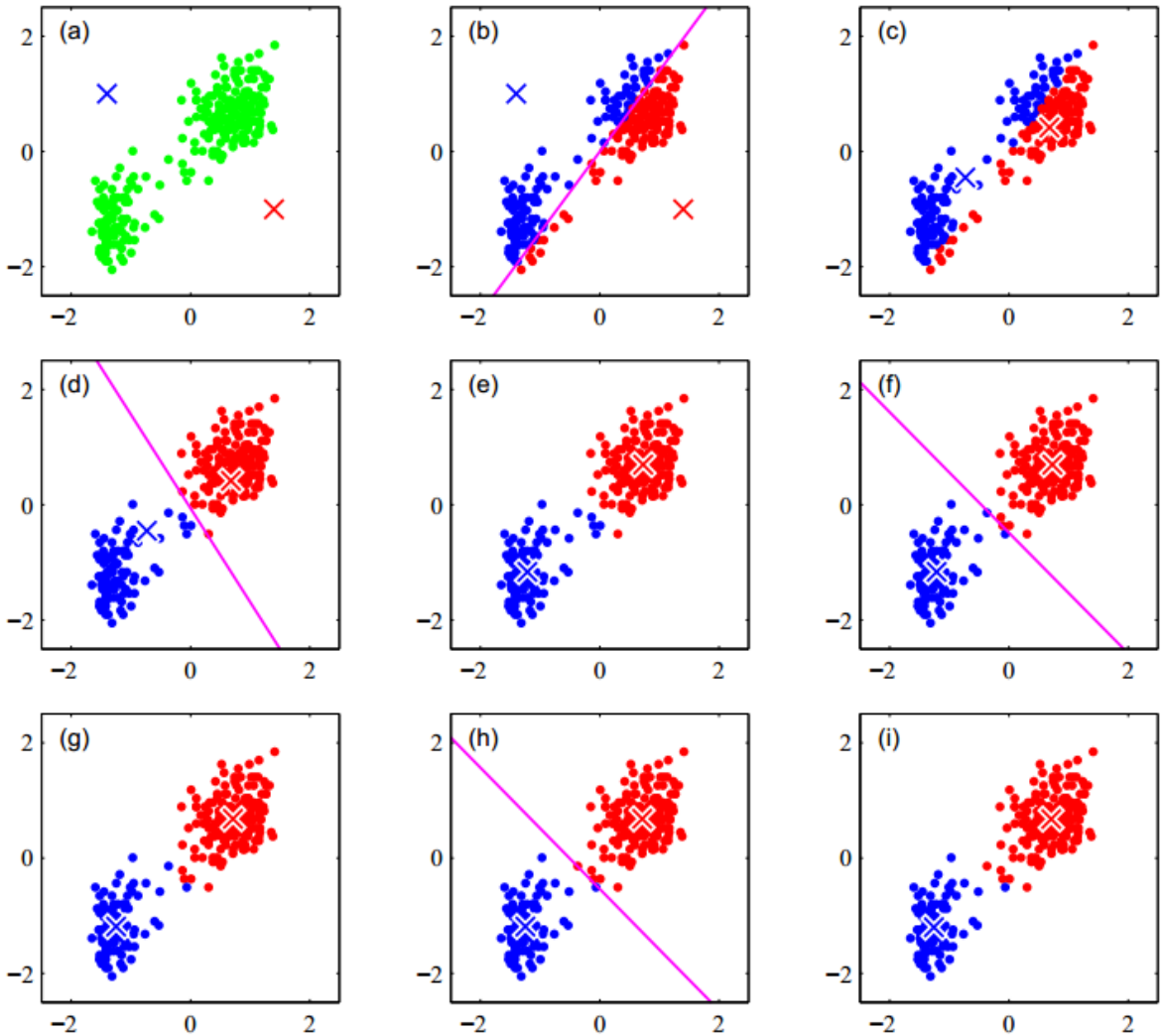


图 1: K 均值聚类实例

K 均值聚类算法

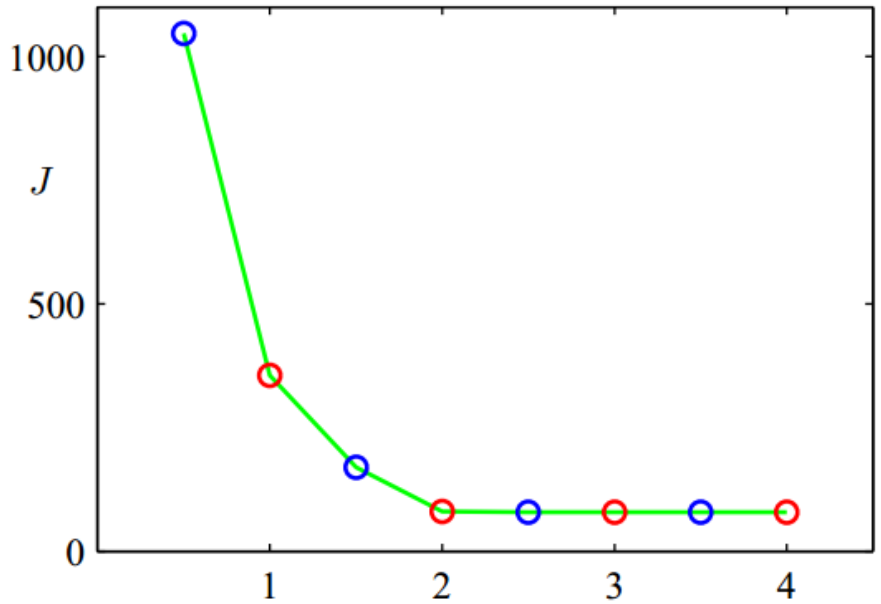


图 2: K 均值聚类实例

由于每个阶段都减小了目标函数 J 的值，因此算法的收敛性得到了保证。然而，算法可能收敛到 J 的一个局部最小值而不是全局最小值。

K 均值聚类算法

推广

K 均值算法的基础是将欧氏距离的评分作为数据点与代表向量之间不相似程度。这不仅限制了能够处理的数据变量的类型，而且使得聚类中心的确定对于异常点不具有鲁棒性。

- 引入两个向量 $\boldsymbol{x}$ 与 $\boldsymbol{x}'$ 之间更加一般的不相似程度的度量 $V(\boldsymbol{x}, \boldsymbol{x}')$:

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} V(\boldsymbol{x}_n, \boldsymbol{\mu}_k)$$

- 通常会将聚类原型限制为等于某个分配到那个聚类的数据向量。需要 $O(N_k^2)$ 次对 $V(.,.)$ 的计算

注意：K 均值算法在每一次迭代中，每个数据点被分配到一个唯一的与之最近的聚类中——硬聚类。位于两个聚类中心的大概中间的位置，应该怎么分配?

应用——图像分割

图 3: 图像分割

K 均值聚类算法

应用——数据压缩(有损)

- 无损数据压缩：能从压缩表示中精确地重建原始数据
- 有损数据压缩：重建过程中会出现一些错误

对于 N 个数据点中的每一个，只存储它被分配的聚类种类 k。以及 K 个聚类中心 μ_k 的值，其中 $K \ll N$。这样，每个数据点都根据它最近的中心 μ_k 确定。新的数据点可以类似的压缩。

对于图像，可以考虑相邻像素组成的小块。

混合高斯

混合高斯

表示

混合高斯分布一般用来对数据进行概率建模。

单独的高斯分布由于太简单、表达能力有限，面对复杂的实际问题往往无能为力。如果取多个高斯分布，并以合适的系数加权平均，就能得到表达能力更丰富的一类模型——高斯混合模型：

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

其中，K 为高斯分布的个数，$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$ 为第 k 个高斯分布的均值和方差，π_k 是第 k 个高斯分布的加权系数，也叫混合系数。满足 $0 \leq \pi_k \leq 1$ 和 $\sum_{k=1}^{K} \pi_k = 1$。

混合高斯

表示

引入隐变量，构造混合高斯的另一种表示。想象混合高斯分布的数据点生成过程：首先以一定的概率从 K 个高斯分布中选择一个，其中第 k 个高斯分布被选中的概率为 π_k，然后依据被选中的高斯分布生成一个数据点。重复上述过程，生成多个数据点，这些数据点将符合混合高斯分布。

为了表示选中哪一个高斯分布，引入一个二值指示变量 $\boldsymbol{z}$。该变量是一个 K 维的，每个元素的取值只能是 0 或者 1，且同时只能有一个元素取 1。由于第 k 个高斯分布被选中的概率为 π_k，即

$$p(z_k = 1) = \pi_k.$$

$$\sum_{k=1}^{K} \pi_k = 1.$$

混合高斯

表示

随机变量 $\boldsymbol{z}$ 的分布可表示为

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

在给定 $\boldsymbol{z}$ 的条件下，$\boldsymbol{x}$ 的条件概率分布是一个高斯分布

$$p(\boldsymbol{x}|z_k = 1) = N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

或者

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} N(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

混合高斯

表示

从而 $\boldsymbol{x}$ 的边缘概率分布可以通过将联合概率分布对所有可能的 $\boldsymbol{z}$ 求和的方式得到，即

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) = \sum_{\boldsymbol{z}} \prod_{k=1}^{K} [\pi_k N(x|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_k}$$
$$= \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

可见，用隐藏变量重新表达的高斯混合模型获得的 $\boldsymbol{x}$ 的边缘分布与之前的一致。

表示

将隐变量显式的写出，可以极大的简化后面最大似然的求解。通过引入期望最大化 (EM) 算法，可以看到这一点。

首先来看看在不引入隐变量的情况下，如何做最大似然的求解。

先给出 $\gamma(z_k) = p(z_k = 1|\boldsymbol{x})$ 这一后验概率。

$$\begin{aligned}\gamma(z_k) = p(z_k = 1|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|z_k = 1)p(z_k = 1)}{\sum_{j=1}^{K} p(\boldsymbol{x}|z_j = 1)p(z_j = 1)} \\ &= \frac{\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}\end{aligned}$$

这个后验，可以看作分量 k 对于 “解释” 观测值 $\boldsymbol{x}$ 的 “责任” (responsibility)。

混合高斯

示例

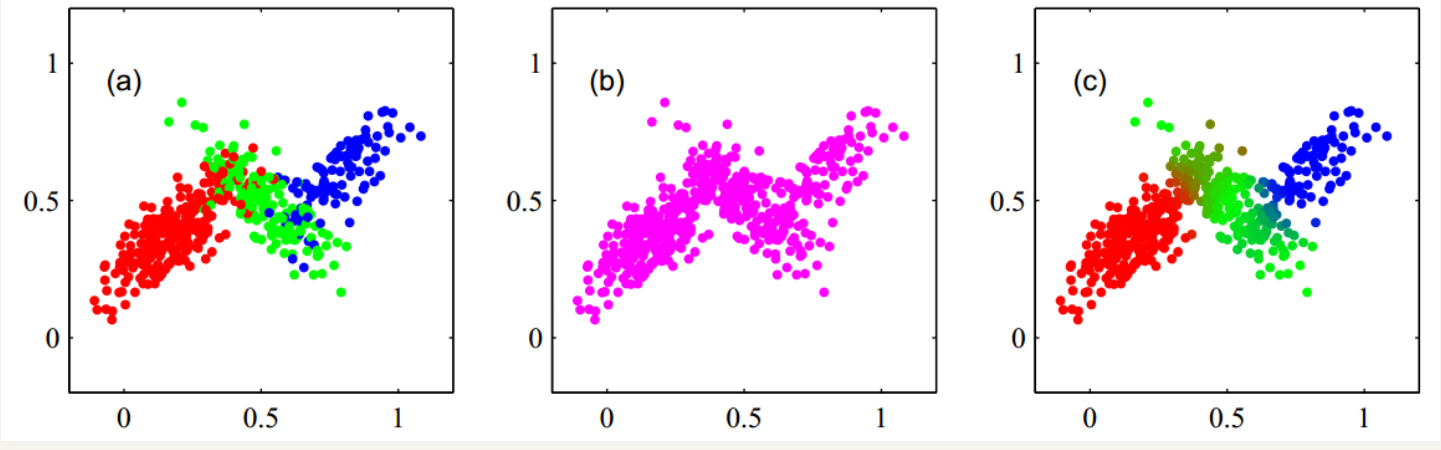
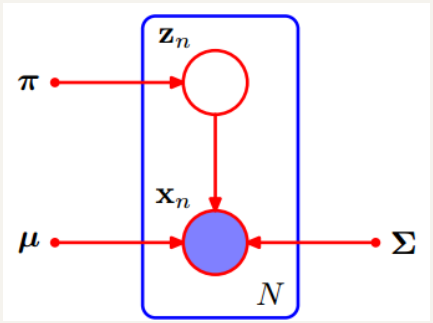


图 4: 3 个高斯分布组成的混合高斯分布

- (a) 从联合分布 $p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$ 中抽取的样本
- (b) 从边缘概率分布 $p(\boldsymbol{x})$ 中抽取样本
- (c) 同样的样本，颜色表示与数据点 $\boldsymbol{x}$ 关联的责任 $\gamma(z_{nk})$ 关联

最大似然表示

假设有一个观测的数据集 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$。使用高斯混合模型对数据进行建模。将数据点表示成一个 $N \times D$ 的矩阵 $\boldsymbol{X}$，其中行为 $\boldsymbol{x}_n^T$；隐变量表示成 $N \times K$ 的矩阵 $\boldsymbol{Z}$，其中行为 $\boldsymbol{z}_n^T$。假定数据点相互独立，那么依据图模型



可得对数似然函数为

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \}$$

最大似然求解

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \}$$

其中，参数有：$\boldsymbol{\pi}$、$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$

对 $\boldsymbol{\mu}_k$ 求偏导，并令导数为 0:

$$\begin{aligned} \frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}_k} &= \sum_{n=1}^{N} \frac{\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) \\ &= \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \end{aligned}$$

可得 $\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$，其中 $N_k = \sum_{i=1}^{N} \gamma(z_{nk})$。

最大似然求解

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \}$$

其中，参数有：$\boldsymbol{\pi}$、$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$

对 $\boldsymbol{\Sigma}_k$ 求偏导，并令导数为 0:

$$\begin{aligned}\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\Sigma}_k} &= -\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})[|\boldsymbol{\Sigma}|^{-1}\frac{\partial|\boldsymbol{\Sigma}_k|}{\partial\boldsymbol{\Sigma}_k} + \frac{\partial(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)}{\partial\boldsymbol{\Sigma}_k}] \\ &= -\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}[\boldsymbol{I}-(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}] = 0\end{aligned}$$

可得 $\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^T$，其中 $N_k = \sum_{i=1}^{N}\gamma(z_{nk})$。

最大似然求解

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \}$$

其中，参数有：$\boldsymbol{\pi}$、$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$

关于 π_k 有 $\sum_{k=1}^{K} = 1$ 的归一化约束，因此在原目标函数的基础上加上一个拉格朗日乘子。

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

对 π_k 求偏导，并令导数为 0:

$$\frac{\partial \ln p(\boldsymbol{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)}{\partial \pi_k} \cdot \pi_k = \sum_{n=1}^{N} \gamma(z_{nk}) + \pi_k \lambda = 0$$

两边同时对 k 求和，可得 $\lambda = -N$，进一步的，$\pi_k = \frac{N_k}{N}$。

最大似然求解

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

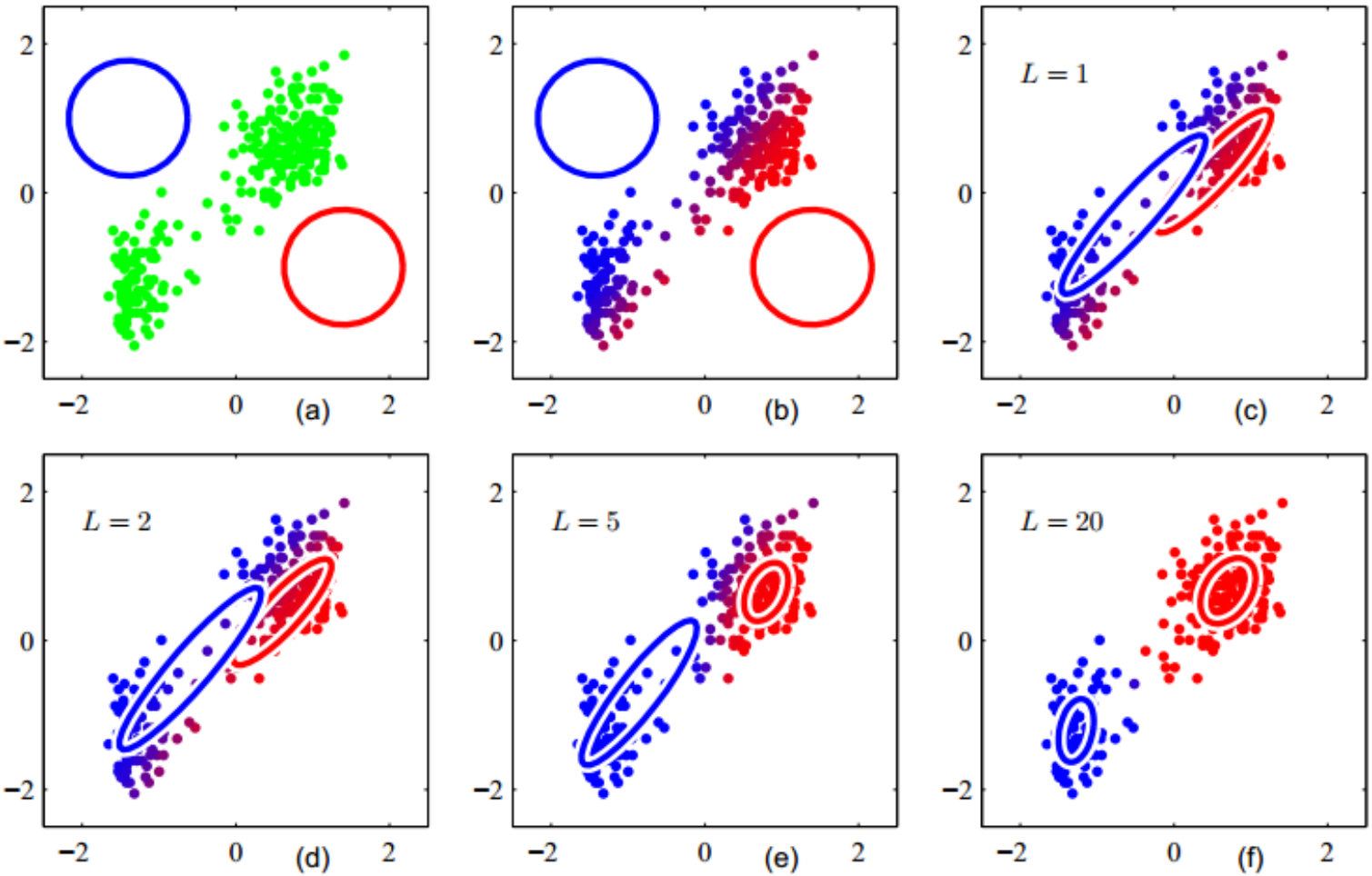
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

其中，$N_k = \sum_{i=1}^{N} \gamma(z_{nk})$

这三个式子，并不构成最大似然的解析解，因为 $\gamma(z_{nk})$ 以一种复杂的方式依赖于这些参数。采用交替进行两个更新的方法，分别对应 E 步和 M 步。首先，给三个参数赋上初始值；E 步计算 “责任” $\gamma(z_{nk})$；M 步计算 $\boldsymbol{\pi}$、$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$；不断迭代，直至收敛。

混合高斯



混合高斯

应用最大似然方法求解高斯混合模型参数时，将要面临两个问题：奇异性和唯一性。

奇异性

考虑一个高斯混合模型，它的分量的协方差矩阵为 $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$，其中 $\boldsymbol{I}$ 为单位矩阵。需要说明的是，不仅仅是对角阵，奇异性问题对任意形式的协方差矩阵都是存在的。假设这个分量的均值正好等于某个数据点，即 $\boldsymbol{\mu}_k = \boldsymbol{x}_n$，则该数据点在这个高斯分布下的似然值为

$$N(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\sigma^D}$$

当 $\sigma \to 0$ 时，上式趋于无穷大，因此，似然函数也趋于无穷大。

启发式方法：如果检测到高斯分量收缩到一个数据点，那么就将它的均值重新设定为一个随机选择的值，并且重新将它的方差设置为某个较大值，然后继续最优化。

混合高斯

应用最大似然方法求解高斯混合模型参数时，将要面临两个问题：奇异性和唯一性。

唯一性

对于任意给定的最大似然解，一个由 K 个分量混合而成的概率分布总共会有 $K!$ 个等价解，对应于 $K!$ 种将 K 个参数集合分配到 K 个分量上的方式。

以 $K=3$ 的一维模型为例，假设得到的一组解：$\boldsymbol{\pi}=[0.5,0.3,0.2]^T$，$\mu_1=1$，$\mu_2=2$，$\mu_3=3$，$\Sigma_1=0.1^2$，$\Sigma_2=0.2^2$，$\Sigma_3=0.3^2$。如果交换两组参数的顺序，可以得到另一组等价的解：$\boldsymbol{\pi}=[0.3,0.5,0.2]^T$，$\mu_1=2$，$\mu_2=1$，$\mu_3=3$，$\Sigma_1=0.2^2$，$\Sigma_2=0.1^2$，$\Sigma_3=0.3^2$。这些不同组合的解共有 $3!=6$ 中。

但是，这个问题与找到一个好的概率模型无关，因为任意等价的解互相之间一样好。

EM 算法

一般化的 EM 算法

符号说明：假设有一个观测的数据集 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$。使用高斯混合模型对数据进行建模。将数据点表示成一个 $N \times D$ 的矩阵 $\boldsymbol{X}$，其中行为 $\boldsymbol{x}_n^T$；隐变量表示成 $N \times K$ 的矩阵 $\boldsymbol{Z}$，其中行为 $\boldsymbol{z}_n^T$；用 $\boldsymbol{\theta}$ 表示所有模型的参数。因此，对数似然为：

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \sum_{\boldsymbol{z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

假定给定数据集 $\boldsymbol{X}$，对应的 $\boldsymbol{Z}$ 是知道的，那么 $\boldsymbol{X}$ 和 $\boldsymbol{Z}$ 的联合分布的对数似然为 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$。

将 $\{\boldsymbol{X}, \boldsymbol{Z}\}$ 称为完全数据集，把单独的 $\{\boldsymbol{X}\}$ 称为不完全数据。

一般化的 EM 算法

在实际中，我们只知道 $\boldsymbol{X}$，$\boldsymbol{Z}$ 是不知道的，所有关于 $\boldsymbol{Z}$ 的信息只能从其后验分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ 中得知，因此，无法求得完全数据的似然。

但是可以考虑完全数据的似然在后验分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ 下的期望。

- E 步：用参数的当前值 $\boldsymbol{\theta}^{old}$ 求得隐藏变量的后验分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$；
- M 步：用此后验分布计算完全数据似然的期望，并将此期望记作 $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$:

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$$

然后求得此期望最大值时的参数：

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old})$$

一般化的 EM 算法

EM 算法：给定完全数据的似然 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ 的表达式，目的是求得参数 $\boldsymbol{\theta}$ 使得观测数据的似然 $p(\boldsymbol{X}|\boldsymbol{\theta})$ 最大。

- 选择参数的初始值，记为 $\boldsymbol{\theta}^{old}$；
- E 步：计算隐藏变量的后验分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$；
- M 步：最大化完全数据的对数似然在布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ 下的期望，从而得到 $\boldsymbol{\theta}^{new}$

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}),$$

其中

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

- 检查迭代是否收敛，如果收敛则停止，否则令 $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$。

一般化的 EM 算法证明

为什么要最大化完全数据的对数似然在后验分布下的期望？为什么这样迭代后就会收敛到观测数据似然 $p(\boldsymbol{X}|\boldsymbol{\theta})$ 的极大值?

$$\begin{aligned}\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} \\ &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} \\ &= L(q,\boldsymbol{\theta}) + KL(q||p)\end{aligned}$$

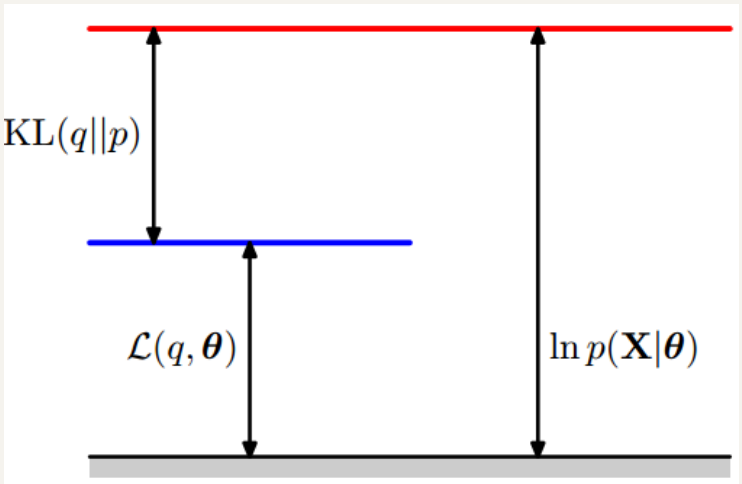
其中,

$$L(q,\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})}, \quad KL(q||p) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

一般化的 EM 算法证明

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = L(q, \boldsymbol{\theta}) + KL(q||p)$$

将目标函数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ 分解成了 $L(q, \boldsymbol{\theta})$ 和 $KL(q||p)$ 两项。$L(q, \boldsymbol{\theta})$ 是 $\boldsymbol{\theta}$ 的函数，是 $q(\boldsymbol{Z})$ 的泛函。$KL(q||p)$ 是分布 p 和 q 之间的 KL 距离，也是一个泛函，具有非负的特性，即 $KL(q||p) \geq 0$，当且仅当 $q = p$ 时，$KL(q||p) = 0$。分解的示意图为



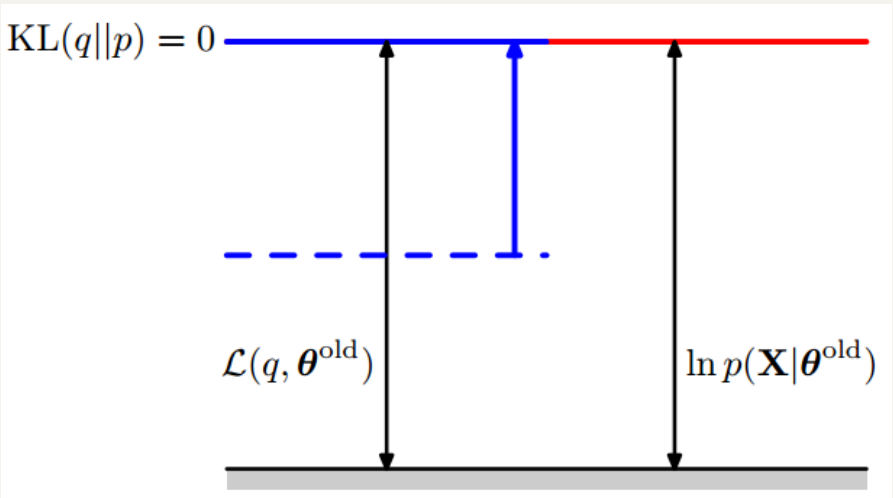
由于 $KL(q||p) \geq 0$，因此 $L(q, \boldsymbol{\theta})$ 是 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ 的下界，通过最大化下界，以最大化 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$。

一般化的 EM 算法证明

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = L(q,\boldsymbol{\theta}) + KL(q||p)$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

E 步：假设参数向量的当前值为 $\boldsymbol{\theta}^{old}$，那么 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ 是固定的。而等式右边的取值仅仅与 $q(\boldsymbol{Z})$ 相关。并且下界的最大值出现在 $KL(q||p) = 0$，即 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$ 时。

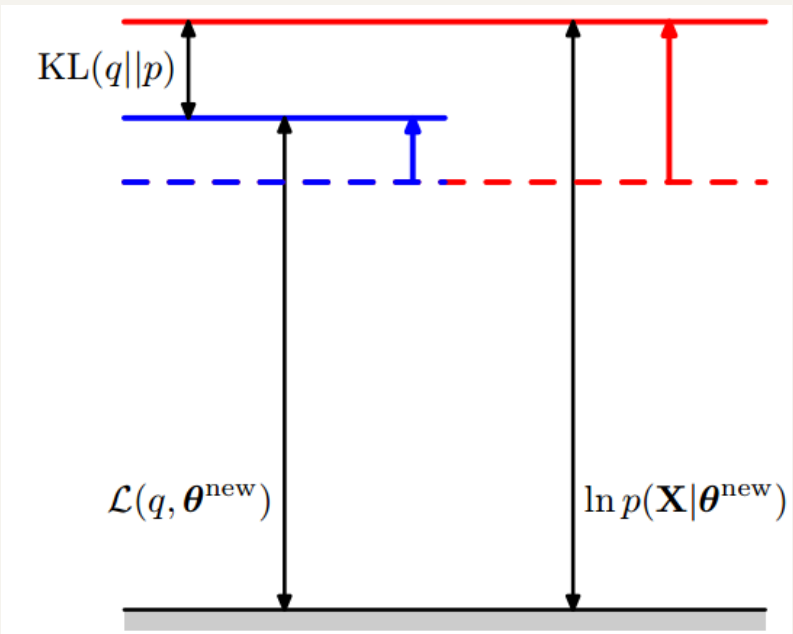


一般化的 EM 算法证明

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = L(q,\boldsymbol{\theta}) + KL(q||p)$$

$$= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}$$

M 步：固定 $q(\boldsymbol{Z})$，通过改变 $\boldsymbol{\theta}$ 的值来最大化下界。此时，由于 $\boldsymbol{\theta}^{new} \neq \boldsymbol{\theta}^{old}$，从而 $KL(q||p) > 0$。因此，$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ 增加的量比 $L(q,\boldsymbol{\theta})$ 大。



一般化的 EM 算法证明

$$\begin{aligned}\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= L(q,\boldsymbol{\theta}) + KL(q||p) \\ &= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} + \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\end{aligned}$$

M 步：固定 $q(\boldsymbol{Z})$，通过改变 $\boldsymbol{\theta}$ 的值来最大化下界。此时，由于 $\boldsymbol{\theta}^{new} \neq \boldsymbol{\theta}^{old}$，从而 $KL(q||p) > 0$。因此，$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ 增加的量比 $L(q,\boldsymbol{\theta})$ 大。

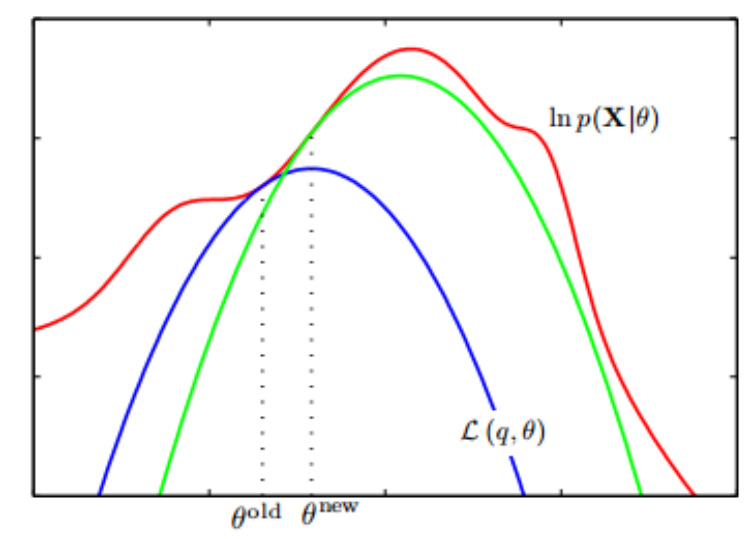
将 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$ 代入 $L(q,\boldsymbol{\theta})$ 得

$$\begin{aligned}L(q,\boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \\ &= Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) + const\end{aligned}$$

常数就是分布 q 的熵，与 $\boldsymbol{\theta}$ 无关。因此 M 步要最大化的实际是完全数据的对数似然在后验分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})$ 下的期望。

因为每经过一轮 E 步和 M 步，都会增大目标函数的值，因此当迭代收敛时，找到的一定是目标函数的极大值。

一般化的 EM 算法的证明



EM 算法的计算也可以看成是参数空间中的运算，红色曲线表示不完全数据的对数似然，它的最大值是我们想要的，蓝色曲线是下界。两条曲线在 $\boldsymbol{\theta}^{old}$ 处相切，即两条曲线梯度相同。

实例

高斯混合模型的 EM

完全数据的对数似然

完全数据为 $\{\boldsymbol{X}, \boldsymbol{Z}\}$，并且已知

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}, \quad p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

那么，完全数据的似然函数为

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= \prod_{n=1}^{N} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \\ &= \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) p(\boldsymbol{z}_n|\boldsymbol{\pi}) \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} \end{aligned}$$

高斯混合模型的 EM

完全数据的对数似然

两边取对数得

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} [\ln \pi_k + \ln N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

因此，完全数据对数似然在后验分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{old})$ 下的期望为

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) &= E_{\boldsymbol{Z}}[\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} [\ln \pi_k + \ln N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} E[z_{nk}] [\ln \pi_k + \ln N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned}$$

高斯混合模型的 EM

完全数据的对数似然

由于给定 $\boldsymbol{X}$ 时，各隐变量之间相互独立，故

$$
\begin{aligned}
E[z_{nk}] &= \frac{\sum_{z_n} z_{nk} \prod_{k'} [\pi_{k'} N(\boldsymbol{x_n}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{z_n} \prod_j [\pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}} \\
&= \frac{\pi_k N(\boldsymbol{x_n}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}{\sum_{j=1}^{K} \pi_j N(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\
&= \gamma(z_{nk})
\end{aligned}
$$

从而，

$$
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) [\ln \pi_k + \ln N(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]
$$

对上式求极大值，然并令导数为 0，可以直接地得到解析解。

混合高斯与 K 均值的关系

EM 视角下的 K 均值聚类算法

K 均值其实是高斯混合模型的一个特例，只需要对高斯混合模型增加一个简单的约束即可。具体的，考虑一个高斯混合模型，其中混合分量的协方差矩阵为 $\epsilon \boldsymbol{I}$，ϵ 是一个被所有分量共享的方差参数，$\boldsymbol{I}$ 是单位矩阵，从而

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\}$$

那么，

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\frac{1}{2\epsilon}||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2\}}{\sum_j \pi_j \exp\{-\frac{1}{2\epsilon}||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2\}}$$

现在，让 $\epsilon \to 0$，则上式分子分母中的指数都趋近于 0，但趋近速度不一样，其中，$||\boldsymbol{x}_n - \boldsymbol{\mu}_j||$ 最小的那项趋近的速度最慢。因此，只有当 $k = \arg\min_j ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2$ 时，$\gamma(z_{nk}) = 1$，k 为其他值时，$\gamma(z_{nk}) = 0$。即

$$\gamma(z_{nk}) = \begin{cases} 1, & \text{if } k = \arg\min_j ||\boldsymbol{x}_n - \boldsymbol{\mu}_j||^2 \\ 0, & \text{otherwise} \end{cases}$$

伯努利分布混合模型

伯努利分布

考虑 D 个二值变量 x_i 组成的集合，其中 $i = 1, \cdots, D$，每个变量都由一个参数为 μ_i 的伯努利分布控制，即

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

其中，$\boldsymbol{x} = (x_1, \cdots, x_D)^T$ 且 $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_D)^T$。

$$E[\boldsymbol{x}] = \boldsymbol{\mu}$$

$$cov[\boldsymbol{x}] = diag\{\mu_i(1-\mu_i)\}$$

伯努利分布混合模型

现在考虑这种分布的有限混合，即

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$

其中，$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k\}$，$\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$，且

$$p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

这个混合分布的均值和方差为

$$E[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k, \quad cov[\boldsymbol{x}] = \sum_{k=1}^{K} \pi_k \{\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\} - E[\boldsymbol{x}]E[\boldsymbol{x}]^T$$

其中，$\boldsymbol{\Sigma}_k = diag\{\mu_{ki}(1-\mu_{ki})\}$

伯努利分布混合模型

伯努利分布混合模型对数似然

假设有一个数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N\}$，那么这个模型的对数似然为

$$\ln p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\pi})=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)\}$$

由于求和位于对数运算内部，从而最大似然没有解析解。采用 EM 算法求解。

完全数据对数似然

已知

$$p(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{\mu})=\prod_{k=1}^{K}p(\boldsymbol{x}|\boldsymbol{\mu}_k)^{z_k},p(\boldsymbol{z}|\boldsymbol{\pi})=\prod_{k=1}^{K}\pi_k^{z_k}$$

那么

$$\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi})=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\{\ln\pi_k+\sum_{i=1}^{D}[x_{ni}\ln\mu_{ki}+(1-x_{ni})\ln(1-\mu_{ki})]\}$$

伯努利分布混合模型

期望

求完全数据对数似然函数关于隐藏变量后验概率分布的期望，得

$$
\begin{aligned}
& Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \\
&= E_{\mathbf{Z}}[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\}] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} E_{\mathbf{Z}}[z_{nk}]\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\} \\
&= \sum_{n=1}^{n}\sum_{k=1}^{K} \gamma(z_{nk})\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})]\}
\end{aligned}
$$

伯努利分布混合模型

解析解

求完全数据对数似然函数关于隐藏变量后验概率分布的期望的求导，并令导数为 0。

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{n=1}^{n} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})]\}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x}_n$$

$$\pi_k = \frac{N_k}{N}$$

推广 EM 算法

GEM 和 ECM

EM 算法的核心步骤在于求解 M 步的最大化问题，当 $\boldsymbol{x}$ 和 $\boldsymbol{z}$ 的联合分布属于指数分布族时，对数正好和指数相消，可以较容易求得解析解。但当联合分布是其他复杂的分布时，M 步往往不好直接求解，此时需要对标准 EM 算法进行扩展，其中 GEM(Generalized EM) 和 ECM(Expectation conditional maximization) 算法即是其中的例子。

在数据量比较大无法一次性载入内存进行计算时，还可以将标准 EM 算法转化为序列化算法，一次只需要载入一个数据点对参数进行序列化地 (增量式) 更新。

总结

总结

EM 算法适用场景：

- 数据集随机缺失
- 包含隐藏变量的极大似然求解

总结

EM 算法步骤：

- 随机初始化参数 $\boldsymbol{\theta}^{old}$
- E 步：用参数的当前值 $\boldsymbol{\theta}^{old}$ 求得隐藏变量的后验分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$
- M 步：用此后验分布计算完全数据似然的期望，并将此期望记作 $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$$

然后求得此期望最大值时的参数：

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$$

- 检查是否收敛，若收敛则算法终止，否则 $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$

参考：https://zhuanlan.zhihu.com/p/33365515