PRML 读书笔记

C. Lu

2018 年 3 月 7 日

目录

第一章 绪论 **5**

1.1 概率论 5

1.1.1 概率密度 5

1.1.2 期望和方差 7

1.2 信息论 7

第二章 概率分布 **9**

2.1 二元变量 9

2.2 多项式变量 9

2.3 高斯分布 9

2.4 指数族分布 9

第三章 核方法 **11**

第四章 稀疏核机 **13**

第五章 图模型 **15**

第六章 混合模型和 EM **17**

6.1 K-means 算法 17

6.2 混合高斯 18

6.3 EM 算法 20

6.4 EM 算法实例 22

6.4.1 用于混合高斯模型的 EM 22

6.4.2 伯努利分布的混合 22

第七章 近似推断 **23**

第八章 采样方法 **25**

第一章　绪论

1.1　概率论

概率论的两条基本法则：

$$\textbf{sum rule} \quad p(X) = \sum_{Y} p(X, Y) \tag{1.1}$$

$$\textbf{product rule} \quad p(X, Y) = p(Y \mid X)p(X) \tag{1.2}$$

其中 $p(X, Y)$ 是联合概率分布，表示是“X 且 Y 的概率”; $P(Y \mid X)$ 是条件概率，表示为“给定 X 的条件下，Y 发生的概率”。

根据对称性 $p(X, Y) = p(Y, X)$, 可以推导出：

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)} \tag{1.3}$$

公式1.3被称为 贝叶斯定理 (Bayes' theorem)。使用加和法则，贝叶斯定理中的分母可以用出现在分子中的项表示：

$$p(X) = \sum_{Y} p(X \mid Y)p(Y) \tag{1.4}$$

1.1.1　概率密度

对于连续型随机变量 x 位于区间 (a, b) 上的概率由下式给出：

$$p(x \in (a, b)) = \int_{a}^{b} p(x) \, \mathrm{d}x \tag{1.5}$$

由于概率是非负的，并且 x 的值一定位于实数轴的某个位置，因此概率密度一定满足以下两个条件：

$$p(x) \geqslant 0 \tag{1.6}$$

$$\int_{-\infty}^{+\infty} p(x)\ \mathrm{d}x = 1 \tag{1.7}$$

我们考虑一个随机变量的转换 $x = g(y)$，$p_x(x)$ 与 $p_y(y)$ 分别代表了 x 与 y 的概率密度函数，对于很小的 δ_x，落在区间 $(x,\ x+\delta_x)$ 内的观测值会被变换到 $(y,\ y+\delta_y)$ 中。其中 $p_x(x)\delta_x \simeq p_y(y)\delta_y$，因此

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))\left|g'(y)\right| \tag{1.8}$$

这个性质说明了概率密度最大值的位置，取决于变量的选择。

位于区间 $(-\infty,\ z)$ 的 x 的概率密度由累积分布函数 (c.d.f) 给出。定义为：

$$P(z) = \int_{-\infty}^{z} p(x)\ \mathrm{d}x \tag{1.9}$$

满足 $P'(x) = p(x)$。

如果有多个连续变量 $x_1, x_2, \ldots, x_D$，整体记作向量 $\boldsymbol{x}$，那么我们可以定义联合概率密度 $p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_D)$，使得 $\boldsymbol{x}$ 落在包含点 $\boldsymbol{x}$ 的无穷小体积 $\delta_{\boldsymbol{x}}$ 的概率由 $p(\boldsymbol{x})\delta_{\boldsymbol{x}}$ 给出。多便利那个概率密度必须满足

$$p(\boldsymbol{x}) \geqslant 0 \tag{1.10}$$

$$\int p(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = 1 \tag{1.11}$$

其中积分必须在整个 $\boldsymbol{x}$ 的空间上进行。

概率密度的加和规则、乘积规则以及贝叶斯规则同样可以应用到连续变量的概率密度函数上，也可以应用到离散变量和连续变量的混合的情形。若 x, y 是两个连续变量，那么加和规则和乘积规则为

$$p(x) = \int p(x, y)\ \mathrm{d}y \tag{1.12}$$

$$p(x, y) = p(y \mid x)p(x) \tag{1.13}$$

形式化证明连续变量的加和规则和乘积规则需要用到**测度论**的数学分支。

1.1.2 期望和方差

1.2 信息论

第二章　概率分布

2.1　二元变量

2.2　多项式变量

2.3　高斯分布

2.4　指数族分布

第三章　核方法

第四章　稀疏核机

第五章　图模型

第六章　混合模型和 EM

6.1　K-means 算法

假设有一个数据集 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$，它由 D 维欧几里得空间中的随机变量 $\boldsymbol{x}$ 的 N 次观测组成。我们的目的是要将数据划分成 K 个类别，假设 K 是给定的一个数。

引入一组 D 维向量 $\boldsymbol{\mu}_k$，其中 $k = 1, 2, \ldots, K$，且 $\boldsymbol{\mu}_k$ 是第 k 个聚类关联的一个代表，可以认为 $\boldsymbol{\mu}_k$ 是第 k 个聚类的中心。算法的目的是要找到每个数据点分别属于的类，以及一组向量 $\{\boldsymbol{\mu}_k\}$，使得每个数据点和它最近的向量 $\boldsymbol{\mu}_k$ 之间的距离的平方和最小。

现在，对于每个数据点 $\boldsymbol{x}_n$，引入一组对应的二值指示变量 $r_{nk} \in \{0, 1\}$，其中 $k = 1, 2, \ldots, K$，表示每个数据点 $\boldsymbol{x}_n$ 属于 K 个聚类中的。如果数据点 $\boldsymbol{x}_n$ 属于第 k 个聚类，那么 $r_{nk} = 1$，且对于 $j \neq k$, 有 $r_{nj} = 0$。定义目标函数，形式为:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{k} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \tag{6.1}$$

它表示每个数据点与被分配的向量 $\boldsymbol{\mu}_k$ 之间的距离的平方和。我们的目标是要找到 $\{r_{nk}\}$ 与 $\boldsymbol{\mu}_{nk}$ 的值，使得 J 达到最小值。

可以使用迭代的方法来达到目标。每次迭代分为两个步骤，分别对应 r_{nk} 的最优化和 $\boldsymbol{\mu}_k$ 的最优化。首先，为 $\boldsymbol{\mu}_k$ 选择一些初始值。然后，在第一阶段，关于 r_{nk} 最小化 J，保持 $\boldsymbol{\mu}_k$ 固定。第二阶段，关于 $\boldsymbol{\mu}_k$ 最小化 J，保持 r_{nk} 固定。不断重复这个二阶段优化直到收敛。

首先考虑确定 r_{nk}。公式6.1关于 r_{nk} 是线性的，且与不同的 n 相关的项是独立的，可以对每个 n 分别进行优化。只要 k 的值使 $\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2$ 的值最

小，就令 $r_{nk}=1$，换句话说，可以简单的将数据点的聚类设置为最近的聚类中心。形式化的表述为

$$r_{nk}=\begin{cases}1, & \text{如果}k=\arg\min_j \|\boldsymbol{x}_n-\boldsymbol{\mu}_j\|^2\\ 0, & \text{其他情况}\end{cases} \tag{6.2}$$

现在考虑 r_{nk} 固定时，关于 $\boldsymbol{\mu}_K$ 的最优化。目标函数 J 是一个二次函数，令它关于 $\boldsymbol{\mu}_k$ 的导数等于 0，即可达到最小值，即

$$2\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)=0 \tag{6.3}$$

解出 $\boldsymbol{\mu}_k$ 的值，结果为

$$\boldsymbol{\mu}_k=\frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}} \tag{6.4}$$

6.4中的分母等于聚类 k 中数据点的数量，因此这个结果的意义是：μ_k 为聚类 k 中所有数据点的均值。因此，此算法被称为 K-means 算法。

重新为数据点分配聚类的步骤以及重新计算聚类均值的步骤重复进行，直到聚类的分配不再改变。每个阶段都减小了目标函数 J，因此算法的收敛性得到保证。但是，*算法可能收敛到 J 的一个局部最小值而非全局最小。*

6.2 混合高斯

高斯混合模型的概率分布可以写成多个高斯分布的线形叠加，即

$$p(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{6.5}$$

引入一个 K 维的二值随机变量 $\boldsymbol{z}$, 采用“1-of-K”编码，其中一个特定的元素 z_k 等于 1，其余所有的元素都等于 0。于是 z_k 的值满足 $z_k\in\{0,1\}$ 且 $\sum_k z_k=1$，并且我们看到根据哪个元素非零，向量 $\boldsymbol{z}$ 有 K 个可能的状态。$\boldsymbol{z}$ 的边缘概率分布可以根据混合系数 π_k 进行赋值，即

$$p(z_k=1)=\pi_k \tag{6.6}$$

其中参数 $\{\pi_k\}$ 必须满足

$$0\leqslant\pi_k\leqslant 1 \tag{6.7}$$

以及

$$\sum_{k=1}^{K} \pi_k = 1 \tag{6.8}$$

由于 $\boldsymbol{z}$ 使用了“1-of-K”编码，也可以将这个概率分布写成

$$p(\boldsymbol{z}) = \prod_{k=1}^{k} \pi_k^{z_k} \tag{6.9}$$

对于 $\boldsymbol{z}$ 给定的一个值，$\boldsymbol{x}$ 的条件概率分布是一个高斯分布

$$p(\boldsymbol{x} \mid z_k = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6.10}$$

类似的也可以写成

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{6.11}$$

$\boldsymbol{x}$ 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

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6.12}$$

于是我们找到了一个将隐变量 $\boldsymbol{z}$ 显示写出的一个高斯混合分布一个等价公式。对联合概率分布 $p(\boldsymbol{x}, \boldsymbol{z})$ 而不是对 $p(\boldsymbol{x})$ 进行操作，会产生计算上极大的简化。

另一个有重要作用的量是给定 $\boldsymbol{x}$ 的情况下，$\boldsymbol{z}$ 的后验概率 $p(\boldsymbol{z} \mid \boldsymbol{x})$。用 $\gamma(z_k)$ 表示 $p(z_k = 1 \mid \boldsymbol{x})$，其值可由贝叶斯定理给出

$$\gamma(z_k) = p(z_k = 1 \mid \boldsymbol{x}) = \frac{p(z_k = 1) p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1) p(\boldsymbol{x} \mid z_j = 1)} \tag{6.13}$$

$$= \frac{\pi_k p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} \pi_j p(\boldsymbol{x} \mid z_j = 1)} \tag{6.14}$$

可以将 π_k 看成是 $z_k = 1$ 的先验概率，将 $\gamma(z_k)$ 看成是观测到 $\boldsymbol{x}$ 之后，对应的后验概率。

假设我们有观测数据集 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$，我们希望使用混合高斯来对数据建模。可以将这个数据集标示为 $N \times D$ 的矩阵 $\boldsymbol{X}$，其中第 n 行为 $\boldsymbol{x}_n^\top$。类似的，对应的隐变量被表示为一个 $N \times K$ 的矩阵 $\boldsymbol{Z}$，它的行为 $\boldsymbol{z}_n^\top$，可以使用图6.1 所示的图模型来表示独立同分布数据集的高斯混合模型。$\boldsymbol{X}$ 的对数似然函数为

$$\log p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{6.15}$$

最大化高斯混合模型的对数似然函数**??**eq:6-15 比单一的高斯分布的情形更加复杂。因为对 k 的求和出现在了对数内部；*如果令导数等于零，不会得到一个解析解*。使用基于梯度的优化方法可以得到解，但现在考虑另一种可行方法，称为 *EM 算法*。

6.3 EM 算法

期望最大化算法，也叫 EM 算法，是寻找潜在变量的概率模型的最大似然解的一种通用方法。考虑一个概率模型，其中所有的观测变量记作 $\boldsymbol{X}$，所有隐含变量记作 $\boldsymbol{Z}$。联合概率分布 $p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ 由一组参数 $\boldsymbol{\theta}$ 控制，目标是最大化似然函数

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta}) \tag{6.16}$$

这里，假设 $\boldsymbol{Z}$ 是离散的。直接优化 $p(\boldsymbol{X} \mid \boldsymbol{\theta})$ 比较困难，但是最优化完整数据似然函数 $p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ 就容易得很多。接下来，引入一个定义在隐变量 $\boldsymbol{Z}$ 上的分布 $p(\boldsymbol{Z})$。对任意 $p(\boldsymbol{Z})$，如下分解成立

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q \parallel p) \tag{6.17}$$

其中

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{6.18}$$

$$\mathrm{KL}(p \parallel q) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{6.19}$$

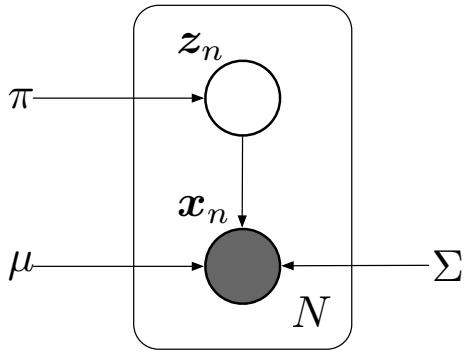

图 6.1: 一组 N 个独立同分布数据点 $\{\boldsymbol{x}_n\}$ 的高斯混合模型的图表示，对应的潜在变量为 $\{\boldsymbol{z}_n\}$，其中 $n = 1, 2, \ldots, N$。

$\mathcal{L}(q, \boldsymbol{\theta})$ 是概率分布 $q(\boldsymbol{Z})$ 的一个*范函*，并且是一个参数 $\boldsymbol{\theta}$ 的函数。因为 $\mathrm{KL}(p \parallel q) \geqslant 0$，当且仅当 $q(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ 时取得等号。因此，$\mathcal{L}(q, \boldsymbol{\theta}) \leqslant \log p(\boldsymbol{X} \mid \boldsymbol{\theta})$，即 $\mathcal{L}(q, \boldsymbol{\theta})$ 是 $\log p(\boldsymbol{X} \mid \boldsymbol{\theta})$ 是的一个下界。

EM 算法是一个两阶段迭代优化算法。

假设当前的参数 $\boldsymbol{\theta}^{\mathrm{old}}$，在 *E 步骤*中，下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ 关于 $q(\boldsymbol{Z})$ 最大化，而 $\boldsymbol{\theta}^{old}$ 保持固定。当 KL 散度为零时，即得到了最大化的解。换句话说，最大值出现在 $q(\boldsymbol{Z})$ 与后验概率分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})$ 相等时，KL 散度等于零，此时，下界等于最大似然函数。

在接下来的 *M 步骤*中，分布 $q(\boldsymbol{Z})$ 保持固定，下界 $\mathcal{L}(q, \boldsymbol{\theta})$ 关于 $\boldsymbol{\theta}$ 最大化，得到了某个新的值 $\boldsymbol{\theta}^{\mathrm{new}}$，这会使得下界 $\mathcal{L}$ 增大。同时也会使得对数似然增大，因为概率分布 q 由旧的参数值确定，并且在 *M 步骤*保持固定，因此不会等于新的后验分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})$，从而 KL 散度非零；而且对数似然的增加量大于下界 $\mathcal{L}(q, \boldsymbol{\theta})$ 的增加量。在 *E 步骤*之后，下界的形式为

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \\ &\quad - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \log p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) + \text{常数}\end{aligned}$$

其中常数是 q 的熵，与 $\boldsymbol{\theta}$ 无关。从而，在 *M 步骤*中，最大化的量是*完整数*

据对数似然函数的期望。完整的 EM 算法如算法**??**所示。

Algorithm 1 用于含有隐变量最大似然函数参数估计的 EM 算法

1: 选择参数的初始值 $\boldsymbol{\theta}^{(t)}, t = 0$

2: **repeat**

3: **E 步骤**：计算 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{(t)})$

4: **M 步骤**：计算 $\boldsymbol{\theta}^{(t+1)}$，由下式给出

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$

其中

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(t)}) \log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$$

5: **until** 对数似然函数收敛或者参数值收敛

6.4 EM 算法实例

6.4.1 用于混合高斯模型的 EM

6.4.2 伯努利分布的混合

第七章　近似推断

第八章　采样方法