Chapter *3* 生成模型

Siheng Zhang
zhangsiheng@cvte.com

2021 年 6 月 2 日

本章对应于 **UML 第 24、31 章，PRML 第 1、2 章**，主要讨论以下问题：

- 贝叶斯最优准则需要估计特征的联合分布，这对实际应用带来了不可计算的困难，解决这个问题的关键是特征独立假设。
- 进一步地，为了估计类条件概率，本章讨论了参数化方法，非参数化的方法相对独立，因此留到其它章节。
- 通过估计潜在分布进行判别的模型，我们称之为生成式模型，包括朴素贝叶斯、混合高斯模型等等。注意到，估计概率密度是机器学习中最为一般化也更难的问题。判别式模型则通过优化目标函数来避免这个问题。
- 但是，生成式模型和判别式模型之间也存在着紧密的关联。本章的最后将会从贝叶斯分类器推导出线性判别器。而再下一章，我们也会指出，为判别式模型添加约束项（通常是为了防止过拟合）本质上与某些先验假设下的生成模型等价。

目录

1　朴素贝叶斯（Naive Bayes，NB）

回顾贝叶斯最优准则（*第 1 章，Ex6*）：$h_{\text{Bayes}}(\boldsymbol{x}) = \arg\max_{y\in\{0,1\}} p(Y=y|X=\boldsymbol{x})$。为了刻画后验概率函数，我们需要 2^d 个参数，这意味着，所需样本的数量随着特征维数指数倍地增加。为了避免这个问题，需要假设给定标签时，各个特征相互独立，即：$p(X=\boldsymbol{x}|Y=y) = \prod_{i=1}^d p(X_i = x_i|Y=y)$。

结合贝叶斯公式，贝叶斯最优准则可以简化为：

$$h_{\text{Bayes}}(\boldsymbol{x}) = \arg\max_{y\in\{0,1\}} p(Y=y)\prod_{i=1}^{d} p(X_i = x_i|Y=y) \tag{1}$$

其中待估计的参数为 $2d+1$ 个。我们使用极大似然法估计这些参数，得到的分类器称为朴素贝叶斯分类器。

2　参数密度估计——极大似然法（Maximum Likelihood Estimation，MLE）

参数密度估计假设类条件概率的分布形式已知（当然，如果选取的分布与实际数据的真实分布相去甚远，则结果也是错的。因此，为了对数据分布做尽可能少的假设，非参数估计就大有用途。但是本章暂不讨论这部分），问题就在于估计分布的参数。给定一个独立同分布的训练集 $S = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$，$S$ 的似然可以由 θ 表示，即 $L(S;\theta) = \prod_{i=1}^m p(\boldsymbol{x}_i;\theta)$。通常我们优化其对数形式，

$$\log L(S;\theta) = \sum_{i=1}^{m} \log p(\boldsymbol{x}_i;\theta) \tag{2}$$

下面对于常见分布给出参数估计的例子。推导过程略显繁琐，结论却浅显且符合直觉。

1 伯努利（Bernoulli）分布，最大似然估计结果等于样本均值，$\theta_{\text{ML}} = \sum_{i=1}^m x_i/m$，

伯努利分布刻画了 0-1 变量 x 的概率，$x=1$ 的概率记为 θ，$x=0$ 的概率为 $1-\theta$，即 $p(x;\theta) = \theta^x(1-\theta)^{(1-x)}$。对应的对数似然函数为

$$\log L(S;\theta) = \sum_{i=1}^{m} \log p(x_i;\theta) = \sum_{i=1}^{m} x_i \log\theta + (1-x_i)\log(1-\theta)$$

对 θ 求导并令导函数为 0，可以得到:

$$\frac{\partial \log L(S;\theta)}{\partial\theta} = \sum_{i=1}^{m} \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta} = \sum_{i=1}^{m} \frac{x_i - \theta}{\theta(1-\theta)} = 0 \Longrightarrow \theta_{\text{ML}} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

2 多项式（Multinomial）分布，参数 $\theta = \boldsymbol{\mu}$ 的最大似然估计结果等于样本均值，$\boldsymbol{\mu}_{\text{ML}} = \sum_{i=1}^m \boldsymbol{x}_i/m$，

多项式分布所刻画的随机变量有 d 个可能的值，用 d 维独热（one-hot，即有且仅有一个元素为 1，其它为 0）向量 $\boldsymbol{x}$ 表示。记 $x_j = 1$ 的概率为 μ_j，则有

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{d} \mu_j^{x_j} \quad s.t. \quad \sum_{j=1}^{d} \mu_j = 1,\ \forall j,\ \ \mu_j \geq 0$$

对应的对数似然函数为

$$\log L(S;\theta) = \sum_{i=1}^{m} \log p(\boldsymbol{x}_i;\theta) = \sum_{i=1}^{m}\sum_{j=1}^{d} x_{ij}\log\mu_j$$

使用拉格朗日乘子 λ，最大化对数似然等价于最大化如下函数：$L' = \log L(S;\theta) + \lambda\left(\sum_{j=1}^d \mu_j - 1\right)$。对 μ_j 求导并令导函数为 0，可以得到:

$$\frac{\partial L'}{\partial \mu_j} = \sum_{i=1}^{m} \frac{x_{ij}}{\mu_j} + \lambda = 0 \Longrightarrow \mu_{j,\text{ML}} = -\sum_{i=1}^{m} x_{ij}/\lambda$$

注意到，$\sum_{j=1}^d \mu_j = -m/\lambda = 1$，可以得到 $\lambda = -m$，从而得到结果。

3 高斯（Gaussian）分布，参数 $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$，最大似然估计分别为样本均值与方差。

高斯分布函数为

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

对应的对数似然函数为

$$\log L(S;\theta) = \sum_{i=1}^{m} \log p(\boldsymbol{x}_i;\theta) = \frac{-md}{2}\log(2\pi) - \frac{m}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

对 $\boldsymbol{\mu}$ 求导并令导函数为 0，可以得到 $\boldsymbol{\mu}_{\text{ML}} = \sum_{i=1}^{m} \boldsymbol{x}_i/m$。

对 $\boldsymbol{\Sigma}$ 求导并令导函数为 0，可以得到（此处不严格证明 $\boldsymbol{\Sigma}$ 为对称阵）：

$$\frac{\partial \log L(S;\theta)}{\partial \boldsymbol{\Sigma}} = -\frac{m}{2}(\boldsymbol{\Sigma}^{-1})^\top + \frac{1}{2}\sum_{i=1}^{m}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1} \Rightarrow \boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{ML}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{ML}})^\top$$

注 *1*：估计 $\boldsymbol{\Sigma}$ 需要用到以下性质：

- $tr[\boldsymbol{ABC}] = tr[\boldsymbol{CAB}] = tr[\boldsymbol{BCA}]$，标量的迹是它本身，因此 $\boldsymbol{x}^\top\boldsymbol{Ax} = tr[\boldsymbol{x}^\top\boldsymbol{Ax}] = tr[\boldsymbol{xx}^\top\boldsymbol{A}]$；
- $\partial tr[\boldsymbol{AB}]/\partial\boldsymbol{A} = \boldsymbol{B}^\top$；$\partial\log|\boldsymbol{A}|/\partial\boldsymbol{A} = (\boldsymbol{A}^{-1})^\top$；$\partial tr(\boldsymbol{AX}^{-1}\boldsymbol{B})/\partial\boldsymbol{X} = -(\boldsymbol{X}^{-1}\boldsymbol{BAX}^{-1})^\top$

4 指数分布族（Exponential family）的极大似然估计结果由特征函数的均值给出。

上述例子都是指数分布族的特例。满足如下形式的分布均属于指数分布族：

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\{\boldsymbol{\eta}^\top\boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\} \tag{3}$$

对数似然函数为：

$$\log L(S;\theta) = \sum_{i=1}^{m}\log h(\boldsymbol{x}_i) + \boldsymbol{\eta}^\top\sum_{i=1}^{m}u(\boldsymbol{x}_i) - \sum_{i=1}^{m}A(\boldsymbol{\eta})$$

对 $\boldsymbol{\eta}$ 求导并令导函数为 0，得到 $\frac{\partial A(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \sum_{i=1}^{m}u(\boldsymbol{x}_i)/m$。结合下面的注，在具体分布中，求得配分函数 $A(\boldsymbol{\eta})$ 和特征函数 $u(\boldsymbol{x})$，即可得到对应的 $\boldsymbol{\eta}_{\text{ML}}$。

此外，利用分布函数积分为 1 的性质进行求导，可以得到：

$$\int h(\boldsymbol{x})\exp\{\boldsymbol{\eta}^\top\boldsymbol{u}(\boldsymbol{x}) - A(\boldsymbol{\eta})\}\left(\boldsymbol{u}(\boldsymbol{x}) - \frac{\partial A(\boldsymbol{\eta})}{\partial\eta}\right) = 0 \Rightarrow \frac{\partial A(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \mathbb{E}[u(\boldsymbol{x})]$$

因此，$\sum_i u(\boldsymbol{x}_i)$ 称为充分统计量。此外，$u(\boldsymbol{x})$ 的方差可以通过 $A(\boldsymbol{\eta})$ 二阶导表示，其它高阶统计矩亦然。事实上，对于指数分布族，只要我们找到配分函数对其归一化，就能通过微分计算其统计矩。

注 *2*：伯努利分布：$p(x|\theta) = \theta^x(1-\theta)^{1-x} = \exp\left\{\log\left(\frac{\theta}{1-\theta}\right)x + \log(1-\theta)\right\}$，对比可得：$h(x) = 1$，$u(x) = x$，$\eta = \log\frac{\theta}{1-\theta}$，$A(\eta) = \log(1+\exp(\eta))$。

注 *3*：多项式分布：注意到，由于 $\sum_{j=1}^{d}\mu_d = 1$，多项式分布事实上有 $d-1$ 个参数，

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{d}\mu_j^{x_j} = \exp\left\{\sum_{j=1}^{d}x_j\log\mu_j\right\} &= \exp\left\{\sum_{j=1}^{d-1}x_j\log\mu_j + \left(1-\sum_{j=1}^{d-1}x_j\right)\log\left(1-\sum_{j=1}^{d-1}\mu_j\right)\right\} \\ &= \exp\left\{\sum_{j=1}^{d-1}x_j\log\left(\frac{\mu_j}{1-\sum_{k=1}^{d-1}\mu_k}\right) + \log\left(1-\sum_{j=1}^{d-1}\mu_j\right)\right\} \end{aligned}$$

令 $\eta_j = \log\left(\mu_j/(1-\sum_{k=1}^{d-1}\mu_k)\right)$，则有 $\mu_j = \left(\exp\eta_j/(1+\sum_{k=1}^{d-1}\exp\eta_k)\right)$，且 $1-\sum_{j=1}^{d-1}\mu_j = \left(1/(1+\sum_{k=1}^{d-1}\exp\eta_k)\right)$。对比可得：$h(\boldsymbol{x}) = 1, u(\boldsymbol{x}) = \boldsymbol{x}, A(\boldsymbol{\eta}) = \log(1+\sum_{k=1}^{d-1}\exp\eta_k)$.

注 *4*：高斯分布：对比可得 $h(\boldsymbol{x}) = (2\pi)^{-d/2}, u(\boldsymbol{x}) = (1, \boldsymbol{x}, \boldsymbol{xx}^\top)^\top, \boldsymbol{\eta} = (-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\boldsymbol{\Sigma}|, \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1})^\top$。

3 从 MLE 到贝叶斯推理

从上述结果直观来看，MLE 对小数据集容易过拟合。定义参数 θ 关于样本 $\boldsymbol{x}$ 的经验损失为负对数似然 $l(\theta,\boldsymbol{x}) = -\log\mathcal{P}_\theta(\boldsymbol{x})$，则 MLE 等价于 ERM。根据真实分布 $\mathcal{P}$，参数 θ 的真实风险为：

$$\mathbb{E}[l(\theta,\boldsymbol{x})] = -\sum_{\boldsymbol{x}}\mathcal{P}(\boldsymbol{x})\log\mathcal{P}_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{x}}\mathcal{P}(\boldsymbol{x})\log\left(\frac{\mathcal{P}(\boldsymbol{x})}{\mathcal{P}_\theta(\boldsymbol{x})}\right) + \sum_{\boldsymbol{x}}\mathcal{P}(\boldsymbol{x})\log\frac{1}{\mathcal{P}(\boldsymbol{x})} \geq \sum_{\boldsymbol{x}}\mathcal{P}(\boldsymbol{x})\log\frac{1}{\mathcal{P}(\boldsymbol{x})}$$

等号成立当且仅当 $\mathcal{P} = \mathcal{P}_\theta$。某些情况下，容易证明 MLE 可以达到较低的真实误差，例如，在已知高斯分布的方差情况下估计其均值，有

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[l(\boldsymbol{\mu}_{\text{ML}},\boldsymbol{x}) - l(\boldsymbol{\mu},\boldsymbol{x})] = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\log\left(\frac{\mathcal{P}_{\boldsymbol{\mu}}(\boldsymbol{x})}{\mathcal{P}_{\boldsymbol{\mu}_{\text{ML}}}(\boldsymbol{x})}\right) = \frac{1}{2}(\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{\text{ML}} - \boldsymbol{\mu})$$

可以看到，MLE 估计得到的参数与真实分布的风险之间的差距是有界的。

但是，另一方面，我们也想知道最坏的情况下 MLE 估计得到的参数真实风险如何。考虑服从伯努利分布的随机变量，假设参数 θ 是一个较小的非零值。连续采样 m 个样本，取值都是 0 的概率为 is $(1-\theta)^m \geq e^{-2m\theta}$。在这种情况下，$\theta_{\mathsf{ML}}=0$，其真实风险为 $\mathbb{E}[l(\boldsymbol{\mu}_{\mathsf{ML}}, x)] = \theta \log l(\boldsymbol{\theta}_{\mathsf{ML}}, 1) + (1-\theta) \log l(\boldsymbol{\theta}_{\mathsf{ML}}, 0) = \theta \log(1/\theta_{\mathsf{ML}}) = \infty$。

为了解决这个问题，贝叶斯推理引入关于参数的先验分布 $p(\theta)$。为了计算的简便，**通常希望后验分布与先验分布的函数形式相同**，称为**共轭性**，其先验称为**共轭先验**。

1 伯努利分布的共轭先验是 Beta 分布

已知服从伯努利分布的数据集的似然函数与 $\mu^x(1-\mu)^{1-x}$ 成比例，因此先验分布应该与 μ 和 $1-\mu$ 的幂次成比例，才能保证先验与似然函数相乘之后，后验分布也与 μ 和 $1-\mu$ 的幂次成比例。Beta 分布正好满足这一点。

$$\mathsf{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \tag{4}$$

其中，$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ 保证了分布的归一化。后验分布具备如下形式：

$$p(\mu|S) \propto p(S|\mu)\mathsf{Beta}(\mu|a,b) = \mu^{a+\sum_{i=1}^m x_i - 1}(1-\mu)^{m-\sum_{i=1}^m x_i + b - 1}$$

为了保证其归一化，后验分布必须为 $\mathsf{Beta}(a+\sum_{i=1}^m x_i, b+m-\sum_{i=1}^m x_i)$。

beta 分布的均值为 $\mathbb{E}(\mu) = \frac{a}{a+b}$，因此预测 $x=1$ 的概率（基于数据集 S）由后验分布的均值给出，即

$$p(x=1|S) = \int_0^1 p(x=1|\mu)p(\mu|S)\mathrm{d}\mu = \int_0^1 \mu p(\mu|S)\mathrm{d}\mu = \mathbb{E}(\mu|S) = \frac{a+\sum_{i=1}^m x_i}{b+m}$$

注意到，S 可以包含无穷样本，即 $m \to \infty$，因此上述结果收敛于 $\frac{\sum_{i=1}^m x_i}{m}$，与 MLE 相同。

2 多项式分布的共轭先验是迪利克雷（Dirichlet）分布

观察似然函数的形式，先验函数应该形如：$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{j=1}^d \mu_j^{\alpha_j - 1}$，其中 $0 \leq \mu_k \leq 1$。令 $\alpha_0 = \sum_{j=1}^d \alpha_j$，其归一化形式即为迪利克雷分布：

$$\mathsf{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_d)} \prod_{j=1}^d \mu_j^{\alpha_j - 1}$$

后验分布为：

$$p(\boldsymbol{\mu}|S) \propto p(S|\boldsymbol{\mu})\mathsf{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \prod_{i=1}^m \prod_{j=1}^d \mu_j^{x_{ij}} \prod_{j=1}^d \mu_j^{\alpha_j - 1} \overset{m_j := \sum_{i=1}^m x_{ij}}{\Longrightarrow} \prod_{j=1}^d \mu_j^{m_j + \alpha_j - 1} = \mathsf{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\boldsymbol{m})$$

3 高斯分布

对于高斯分布，仅讨论一维空间中已知方差估计期望的情况，期望的共轭先验也是高斯分布 $\mathcal{N}(\mu|\mu_0, \Sigma_0)$。其后验分布可以从对数形式给出：

$$\log p(\mu|S) \propto \mu^2(m\sigma^{-1} + \sigma_0^{-1}) - 2\mu\left(\sigma^{-1}\sum_{i=1}^m x_i + \sigma_0^{-1}\mu_0\right) + \frac{\mu_0^2}{\sigma_0} + \frac{\sum_{i=1}^m x_i^2}{\sigma} \Longrightarrow \mathcal{N}\left(\frac{\sigma^2}{m\sigma_0^2+\sigma^2}\mu_0 + \frac{\sigma_0^2}{m\sigma_0^2+\sigma^2}\sum_{i=1}^m x_i, \frac{1}{m\sigma^{-1}+\sigma_0^{-1}}\right)$$

4 局部观测数据的极大似然——最大化期望（Expectation Maximization，EM）

到现在为止，我们讨论的都是标签已知的情况，$S = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$。标签 y_i 可以视为隐变量，它决定了 x_i 从哪个分布采样得到，如果它不能被观测到，那么极大化样本序列 $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$ 的对数似然函数应为：

$$\log L(S;\theta) = \sum_{i=1}^m \log \sum_{j=1}^k p_\theta(\boldsymbol{x}_i, y_j) = \sum_{i=1}^m \log \sum_{j=1}^k p_\theta(\boldsymbol{x}_i|y_j)p_\theta(y_j) \tag{5}$$

对于这类问题，采用最大化期望算法迭代求解。其中，在求期望的阶段（E-step），使用现在的参数 θ^{old} 计算隐变量后验，即 $p(\boldsymbol{Y}|\boldsymbol{X}, \theta^{\mathsf{old}})$，根据后验进一步计算样本对数似然的期望；在求最大值的阶段（M-step），

4.1 EM 算法求解高斯混合模型（Gaussian Mixture Model，GMM）

GMM 是一类典型问题，其待估计参数包括混合系数以及各个类的期望与方差。

结合拉格朗日乘子法，其优化目标为：

$$\sum_{i=1}^{m} \log \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + \lambda \left(\sum_{j=1}^{k} \pi_j - 1 \right) \tag{6}$$

对 $\boldsymbol{\mu}_k$ 求导并令导函数为 0，可以得到

$$\sum_{i=1}^{m} \underbrace{\frac{\pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}}_{z_{ij}} \boldsymbol{\Sigma}_k(\boldsymbol{x}_i - \boldsymbol{\mu}_j) \Longrightarrow \boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m} z_{ij} \boldsymbol{x}_i}{\sum_{i=1}^{m} z_{ij}} \tag{7}$$

in which $z_{ij} = p(y_j = 1|\boldsymbol{x}_i)$ is the posterior probability. Similarly,

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{m} z_{ij} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top}{\sum_{i=1}^{m} z_{ij}} \tag{8}$$

Then, take derivatives with regard to each π_j and set it to zero

$$\sum_{i=1}^{m} \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_l \pi_l \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda = \sum_{i=1}^{m} \frac{z_{ij}}{\pi_j} + \lambda \Longrightarrow \pi_j = -\frac{\sum_{i=1}^{m} z_{ij}}{\lambda}$$

With the constraint that $\sum_{j=1}^{k} \pi_j = -\sum_{i=1}^{m} \sum_{j=1}^{k} z_{ij}/\lambda = -m/\lambda = 1$, then $\lambda = -m$, and hence

$$\pi_j = \frac{\sum_{i=1}^{m} z_{ij}}{m} \tag{9}$$

It means that the mixing coefficient for the k-th component is given by the average posterior which that component takes for explaining the data points. Notes that the calculation above drops into a circle form: $\boldsymbol{\mu}, \boldsymbol{\Sigma} \to z_{ij} \to \boldsymbol{\mu}, \boldsymbol{\Sigma}$

下面是对 GMM 的 EM 算法伪代码：

- fix k, the number of Gaussian components;
- initialize: $\forall j = 1, \cdots, k, z_{ij} = \frac{1}{k}$, and $\pi_j = \frac{1}{k}$;
- M-step, solve $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ according to *Eq.*7 and *Eq.*8;
- E-step, solve z_{ij}, π_i according to *Eq.*9.
- Repeat E-M step until convergence.

5 与判别式模型的比较

生成式模型为数据的潜在分布假定一个参数形式，将学习问题转化为参数估计。但是，在判别式模型中，学习目标是直接估计判别函数的参数。

显然，如果参数估计成功，利用生成式模型直接获得贝叶斯分类器是可靠的。但问题是，逼近潜在分布往往比学习一个判别器难得多（因为逼近潜在分布更加靠近学习问题的本质）。Vladimir Vapnik 因此说：

在解决问题的时候，不应该将一个更加一般化的问题作为其中间步骤。

"When solving a given problem, try to avoid a more general problem as an intermediate step."

但是，生成式模型在某些情况下计算复杂度比判别器学习更低。生成式模型能够先从数据中估计参数，以供未来特定任务下使用，节省了实时的计算。

此外，前沿的生成式模型研究有一个更宏伟的目标，就是从潜在分布中采样得到与真实世界无异的数据。这一路线背后的信念来自于 Richard Feynman 的著名格言：

如果我不能创造某些事物，那么说明我并不真正理解它们。

"What I cannot create, I do not understand."

5.1 从 NB 到线性分类器

考虑二分类情况，假设类概率密度 $p(X = \boldsymbol{x}|Y = y)$ 服从高斯分布，分别记为 $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$。$h_{\text{Bayes}}(\boldsymbol{x}) = 1$，当且仅当

$$
\begin{aligned}
&\frac{p(Y=0)p(X=\boldsymbol{x}|Y=0)}{p(Y=1)p(X=\boldsymbol{x}|Y=1)} > 1 \\
\Longleftrightarrow\ & \log\frac{p(Y=0)}{p(Y=1)} + \log p(X=\boldsymbol{x}|Y=0) - \log p(X=\boldsymbol{x}|Y=1) > 0 \\
\Longleftrightarrow\ & \boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} + 2(\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1})\boldsymbol{x} + \underbrace{\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} + 2\log\frac{p(Y=0)}{p(Y=1)}}_{\text{b}} > 0
\end{aligned}
$$

which is a quadratic discriminant function. Further, if we assume that $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, the classifier can be simplified to be a linear discriminant function $\boldsymbol{w}\cdot\boldsymbol{x}+b$, with $\boldsymbol{w} = 2(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}$ and $b = \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 + 2\log\frac{p(Y=0)}{p(Y=1)}$. If the prior probability is equal, namely $p(Y=0) = p(Y=1)$, the bias term can be further simplified.

6 Exercises and solutions

Ex1 K-means (see *UML Chapter 22.2, PRML Chapter 9.1*). K-means is a simple but important clustering algorithm. In fact, GMM is sometimes called *soft* K-means. As a hard version, K-means assigns the most probable cluster label to an example (*i.e.*, $z_{ij} = 1$ for one of $j \in 1, \cdots, k$ but 0 for others), and calculate the mean and covariance based on the in-cluster instead of global data. Formally, its procedure is as below,

- fix k, the number of clusters;
- randomly choose initial clustering centers $\boldsymbol{\mu}_1^0, \cdots, \boldsymbol{\mu}_k^0$
- loop from $t = 0$ to max_iter
- 1. $\forall i \in \{1, \cdots, m\}$, determine $j = \arg\min_j d(\boldsymbol{x}_i, \boldsymbol{\mu}_j^t)$ and set $z_{ij}^t = 1$;
- 2. $\forall j \in \{1, \cdots, k\}$, update $\boldsymbol{\mu}_j^{t+1} = \frac{\sum_{i=1}^m \boldsymbol{x}_i z_{ij}^t}{\sum_{i=1}^m z_{ij}^t}$;

in which $d(\cdot,\cdot)$ can be arbitrary distance function. Note that the step 1. corresponds to M-step of GMM, and step 2 corresponds to E-step. For GMM, the objective is to maximize likelihood, and for k-means, the objective can be viewed as minimizing the sum of in-cluster distance (if we choose the distance to be Euclidean distance, the loss is also called Sum of in-cluster Square Error, *a.k.a.*, SSE):

$$
C = \min_{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k} \sum_{j=1}^{k} \sum_{i=1,z_{ij}=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{\mu}_j)
$$

Now, prove that: each iteration of the k-means algorithm does not increase the objective.

Solution: According to the iteration,

$$
C^t = \sum_{j=1}^{k} \sum_{i=1,z_{ij}^t=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{t+1}) \leq \sum_{j=1}^{k} \sum_{i=1,z_{ij}^t=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{t}) \leq \sum_{j=1}^{k} \sum_{i=1,z_{ij}^{t-1}=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{\mu}_j^{t}) = C^{t-1}
$$

Ex2 Simplex of Dirichlet distribution Because of the summation constraint, the distribution over the space of the $\{\mu_j\}$ is confined to a simplex of dimensionality $d-1$.

Ex3 Sequential estimation (see *PRML Chapter 2.3.5*).

Ex4 Sequential estimation under the perspective of Bayesian reasoning (see *PRML Chapter 2.3.5*).

Ex5 **Unbiased estimation** (UML Ex24.1) θ_{ML}, in intrinsic, is a function of observed random variables, and hence has its expectation. If the expectation of an estimation is exactly the parameter in theory, we say that the estimation is unbiased. In the case of exponential family,

$$\mathbb{E}(\mu_{\text{ML}}) = \mathbb{E}\left(\frac{\sum_{i=1}^m x_i}{m}\right) = \sum_{i=1}^m \frac{\mathbb{E}(x_i)}{m} = \mathbb{E}(x) = \mu$$

Hence, we say that the MLE for mean parameter is unbiased. Now, prove that the maximum likelihood estimator of the variance of a Gaussian variable is biased.

Solution:

$$\mathbb{E}(\mathbf{\Sigma}_{\text{ML}}) = \sum_{i=1}^m \frac{\mathbb{E}((\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{ML}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\text{ML}})^\top)}{m} = \sum_{i=1}^m \frac{\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top) + \mathbb{E}(\boldsymbol{\mu}_{\text{ML}}\boldsymbol{\mu}_{\text{ML}}^\top) - 2\mathbb{E}(\boldsymbol{\mu}_{\text{ML}}\boldsymbol{x}_i^\top)}{m}$$

Consider each term in the numerator, note that each pair of samples is independent,

$$\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i^\top) = \mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

$$\mathbb{E}(\boldsymbol{\mu}_{\text{ML}}\boldsymbol{\mu}_{\text{ML}}^\top) = \frac{1}{m^2}\mathbb{E}\left(\sum_{i=1}^m\sum_{j=1}^m \boldsymbol{x}_i\boldsymbol{x}_j^\top\right) = \frac{1}{m^2}\mathbb{E}\left(\sum_{i=1}^m\sum_{j=1}^m (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_j - \boldsymbol{\mu})^\top + 2\boldsymbol{\mu}\sum_{i=1}^m(\boldsymbol{x}_i - \boldsymbol{\mu})^\top + \sum_{i=1}^m\sum_{j=1}^m \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) = \frac{\mathbf{\Sigma}}{m} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

$$\mathbb{E}(\boldsymbol{\mu}_{\text{ML}}\boldsymbol{x}_i^\top) = \mathbb{E}\left(\frac{1}{m}\sum_{j=1}^m \boldsymbol{x}_j\boldsymbol{x}_i^\top\right) = \frac{1}{m}\mathbb{E}\left(\sum_{j=1}^m (\boldsymbol{x}_j - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top + 2\boldsymbol{\mu}\sum_{j=1}^m(\boldsymbol{x}_j - \boldsymbol{\mu})^\top + \sum_{j=1}^m \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) = \frac{\mathbf{\Sigma}}{m} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

Hence, $\mathbb{E}(\mathbf{\Sigma}_{\text{ML}}) = \frac{m-1}{m}\mathbf{\Sigma}$ which is biased.

Ex6 **The connection between smoothing and regularized MLE** (UML Ex24.2) Consider the following regularized loss minimization for parameter estimation in the case of Bernoulli distribution:

$$\min \frac{1}{m}\sum_{i=1}^m -\log \mathcal{P}_{\boldsymbol{\mu}}(\boldsymbol{x}_i) + \frac{1}{m}(\log(1/\mu) + \log(1/(1-\mu)))$$

6.1 Show that the preceding objective is equivalent to the usual empirical error had we added two pseudo-examples to the training set.

6.2 Derive a high probability bound on $|\mu' - \mu|$, and use this to bound the true risk.

Solution:

6.1 The regularized loss can be written as

$$-\frac{1}{m}\sum_{i=1}^m x_i \log\mu + (1-x_i)\log(1-\mu) - \frac{1}{m}(\log(\mu) + \log(1-\mu))$$

Take derivatives with regard to μ and set it to zero leads to $\mu' = \frac{1+\sum_{i=1}^m x_i}{m+2}$. It's equivalent to adding two pseudo-examples $\{0, 1\}$ into the training set, which is called 'add-1' smoothing.

6.2 Using triangle inequality,

$$|\mu' - \mu| = |\mu' - \mathbb{E}(\mu') + \mathbb{E}(\mu') - \mu| \leq |\mu' - \mathbb{E}(\mu')| + |\mathbb{E}(\mu') - \mu|$$

Since $\mathbb{E}(\mu') = \frac{1+m\mu}{m+2}$, we have that $|\mathbb{E}(\mu') - \mu| \leq \frac{1}{m+2}$, and $|\mu' - \mathbb{E}(\mu')| = \frac{m}{m+2}|\frac{1}{m}\sum_{i=1}^m x_i - \mu|$. Following Hoeffding's inequality, for any $\epsilon > 0$,

$$P\left(|\mu' - \mu| \geq \frac{1}{m+2} + \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right)$$

Chapter 4. Linear models for classification and regression, penalization

Chapter 5. Decision stumps, ensemble learning, Bayes PAC

Chapter 6. Perceptron, MLP, deep learning, Generalization bounds on deep learning.