模式识别与机器学习
读书笔记

C. Lu

2018 年 3 月 22 日

目录

第一章　绪论

1.1　概率论

概率论的两条基本法则：

$$\textbf{加法法则}\quad p(X) = \sum_{Y} p(X, Y) \tag{1.1}$$

$$\textbf{乘法法则}\quad p(X, Y) = p(Y \mid X)p(X) \tag{1.2}$$

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

在概率分布 $p(x)$ 下，函数 $f(x)$ 的期望

$$\mathbb{E}[f(x)] = \sum_x p(x)f(x) \tag{1.14}$$

连续变量的期望为

$$\mathbb{E}[f(x)] = \int_{-\infty}^{+\infty} p(x)f(x)\ \mathrm{d}x \tag{1.15}$$

在给定 N 个独立同分布的样本点时，可以通过求平均来得到期望的一个估计值

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{1.16}$$

有时候也需要考虑多变量的期望，例如

$$\mathbb{E}_x[f(x,y)] \tag{1.17}$$

表示 $f(x,y)$ 关于 x 的期望。$\mathbb{E}_x[f(x,y)]$ 是关于 y 的函数。同样地，也可以表示关于一个条件分布的期望，即

$$\mathbb{E}_x[f \mid y] = \sum_x p(x \mid y)f(x) \tag{1.18}$$

$f(x)$ 的方差被定义为

$$\mathrm{Var}[f(x)] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] \tag{1.19}$$

展开之后，可以得到

$$\mathrm{Var}[f(x)] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{1.20}$$

特别的，关于 x 的方差

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{1.21}$$

对于两个变量 x 和 y，协方差被定义为

$$\mathrm{Cov}(x,y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] \tag{1.22}$$

在两个随机向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的情形下，协方差是一个矩阵

$$\mathrm{Cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^\top] \tag{1.23}$$

1.2 信息论

随机变量 x 的熵

$$H[x] = -\sum_x p(x)\log_2 p(x) \tag{1.24}$$

对于连续型变量 x 的微分熵

$$H[x] = -\int p(x)\log p(x)\ \mathrm{d}x \tag{1.25}$$

最大微分熵的分布是高斯分布

$$p(x) = \frac{1}{(2\pi\sigma^2)}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.26}$$

对高斯分布求微分熵，可以得到

$$H[x] = \frac{1}{2}\{1+\log(2\pi\sigma^2)\} \tag{1.27}$$

可以看到，熵随着分布宽度 (即 σ^2) 的增加而增加。这个结果也表明，与离散熵不同，微分熵可以为负。

假设有一个联合概率分布 $p(\boldsymbol{x},\boldsymbol{y})$。我们从这个分布中抽取一对 $\boldsymbol{x}$ 和 $\boldsymbol{y}$。如果 $\boldsymbol{x}$ 的值已知，$\boldsymbol{y}$ 的条件熵为

$$H[\boldsymbol{y}\mid\boldsymbol{x}] = -\iint p(\boldsymbol{x},\boldsymbol{y})\log p(\boldsymbol{y}\mid\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \tag{1.28}$$

利用概率乘积规则，很容易得到，条件熵满足下面关系

$$H[\boldsymbol{x},\boldsymbol{y}] = H[\boldsymbol{y}\mid\boldsymbol{x}] + H[\boldsymbol{x}] \tag{1.29}$$

其中 $H[\boldsymbol{x},\boldsymbol{y}]$ 是概率分布 $p(\boldsymbol{x},\boldsymbol{y})$ 的微分熵，$H[\boldsymbol{x}]$ 是边缘分布 $p(\boldsymbol{x})$ 的微分熵。

1.2.1 相对熵 (KL 散度) 和互信息

考虑某个真实分布 $p(x)$，我们使用一个近似分布 $q(x)$ 来对它建模。衡量这两个分布之间的相似度，可以使用两个分布之间的相对熵，也称为 KL

散度

$$\begin{aligned}\mathrm{KL}(p \parallel q) &= -\int p(\boldsymbol{x}) \log q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \left(-\int p(\boldsymbol{x}) \log q(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\right) \\ &= -\int p(\boldsymbol{x}) \log \left\{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right\} \, \mathrm{d}\boldsymbol{x} \end{aligned} \tag{1.30}$$

注意，KL 散度不是对称量，即 $\mathrm{KL}(p \parallel q) \neq \mathrm{KL}(q \parallel p)$。另外 KL 散度满足 $\mathrm{KL}(p \parallel q) \geq 0$，当且仅当 $p(\boldsymbol{x}) = q()$ 时取得等号。可以使用 Jensen 不等式来证明。Jensen 不等式如下

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \tag{1.31}$$

其中 $\mathbb{E}[\cdot]$ 表示期望，$f(x)$ 为凸函数。

现在考虑由 $p(\boldsymbol{x}, \boldsymbol{y})$ 给出的两个变量 $\boldsymbol{x}$，$\boldsymbol{y}$ 组成的数据集。如果变量是独立的，那么联合概率分布可以分解为边缘概率的乘积 $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$。如果变量不是独立的，那么我们可以通过考察联合概率分布与边缘概率分布乘积之间的 KL 散度来判断它们是否接近于独立。此时，KL 散度为

$$\begin{aligned} I[\boldsymbol{x}, \boldsymbol{y}] &\equiv \mathrm{KL}(p(\boldsymbol{x}, \boldsymbol{y}) \parallel p(\boldsymbol{x})p(\boldsymbol{y})) \\ &= -\iint p(\boldsymbol{x}, \boldsymbol{y}) \log \left(\frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x}, \boldsymbol{y})}\right) \, \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \end{aligned} \tag{1.32}$$

这称为变量 $\boldsymbol{x}$ 与变量 $\boldsymbol{y}$ 之间的互信息。根据 KL 散度的性质，有 $I[\boldsymbol{x}, \boldsymbol{y}] \geq 0$，当且仅当 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 相互独立时取得等号。使用概率规则和乘积规则，可以看到互信息和条件熵之间的关系为

$$I[\boldsymbol{x}, \boldsymbol{y}] = H[\boldsymbol{x}] - H[\boldsymbol{x} \mid \boldsymbol{y}] = H[\boldsymbol{y}] - H[\boldsymbol{y} \mid \boldsymbol{x}] \tag{1.33}$$

互信息可以看作是知道了 $\boldsymbol{y}$ 值而造成的 $\boldsymbol{x}$ 的不确性减少 (反之亦然)。从贝叶斯的观点来看，可以把 $p(\boldsymbol{x})$ 看成是 $\boldsymbol{x}$ 的先验分布，把 $p(\boldsymbol{x} \mid \boldsymbol{y})$ 看成观测到新数据 $\boldsymbol{y}$ 之后的后验分布。互信息表示一个新观测数据 $\boldsymbol{y}$ 造成的 $\boldsymbol{x}$ 的不确定性减少。

第二章　概率分布

2.1　二元变量

首先考虑一个二元变量 $x \in \{0, 1\}$，其中 $x = 1$ 的概率记作参数 μ，

$$p(x = 1 \mid \mu) = \mu \tag{2.1}$$

其中 $0 \leq \mu \leq 1$。显然，$p(x = 0 \mid \mu) = 1 - \mu$。$x$ 的概率分布因此可以写成

$$\mathrm{Bern}(x \mid \mu) = \mu^x (1 - u)^{1-x} \tag{2.2}$$

这叫做伯努利分布。分布的均值和方差

$$\mathbb{E}[x] = \mu \tag{2.3}$$

$$\mathrm{Var}[x] = \mu(1 - \mu) \tag{2.4}$$

现在假设有一个观测数据集 $\mathcal{D} = \{x_1, \ldots, x_N\}$，且每次观测数据都是独立地从 $p(x \mid \mu)$ 中抽取的，因此可以构造关于 μ 的似然函数

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \tag{2.5}$$

以频率学家的观点来看，可以通过最大化似然函数来估计 μ 的值，或者，等价的，最小化交叉熵。在伯努利分布的形式下，对数似然函数为

$$\log p(\mathcal{D} \mid \mu) = \sum_{n=1}^{N} \log p(x_n \mid \mu) = \sum_{n=1}^{N} \{x_n \log \mu + (1 - x_n) \log(1 - \mu)\} \tag{2.6}$$

如果令 $\log p(\mathcal{D} \mid \mu)$ 关于 μ 的导数等于零，就得到了最大似然的估计值

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.7}$$

这样称为样本均值。如果我们把数据集里 $x = 1$ 的观测数量记作 m，那么可以把公式2.7写成下面的形式

$$\mu_{\mathrm{ML}} = \frac{m}{N} \tag{2.8}$$

我们也可以求解给定观测次数 N 的情况下，$x = 1$ 的观测次数 m 的概率分布。这个分布式**二项分布**。

$$\mathrm{Bin}(m \mid \mu, N) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.9}$$

其中

$$\binom{N}{m} = \frac{N!}{(N-m)!m!} \tag{2.10}$$

由于 $m = \sum_{n=1}^{N} x_n$，所以二项分布的均值和方差为

$$\mathbb{E}[m] = \sum_{m=0}^{N} m\mathrm{Bin}(m \mid \mu, N) = N\mu \tag{2.11}$$

$$\mathrm{Var}[m] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2\mathrm{Bin}(m \mid \mu, N) = N\mu(1-\mu) \tag{2.12}$$

2.1.1 Beta 分布

在最大似然框架中，会出现过拟合的现象，因此可以采用贝叶斯的观点。在伯努力分布中，引入一个关于 μ 的先验分布 $p(\mu)$。我们可以考虑一种简单的先验分布，这种分布与后验分布具有相同的性质，这被称为**共轭先验**。参数 μ 的先验被选择为 *Beta* **分布**，形式为

$$\mathrm{Beta}(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2.13}$$

其中 $\Gamma(x)$ 是由以下定义的函数

$$\Gamma(x) \equiv \int_0^{\infty} u^{x-1}e^{-u}\mathrm{d}u \tag{2.14}$$

Beta 分布的均值和方差为

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{2.15}$$

$$\mathrm{Var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \tag{2.16}$$

参数 a 和 b 通常被称为超参数。

现在，可以写出 μ 的后验概率：将 Beta 先验2.13 与二项似然函数2.5 相乘，之后归一化。可以得到后验概率分布的形式

$$p(\mu \mid m, l, a, b) = \frac{\Gamma(m+a+b+l)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1} \tag{2.17}$$

其中 m 和 l 分别为 N 次观测中，$x=1$ 与 $x=0$ 的次数。

2.2 多项式变量

2.3 高斯分布

2.4 指数族分布

我们目前所讨论的概率分布都属于*指数族分布*的特例。参数为 $\boldsymbol{\eta}$ 的变量 $\boldsymbol{x}$ 的指数族分布定义为具有如下形式的概率分布的集合

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^{\top}\boldsymbol{u}(\boldsymbol{x})\} \tag{2.18}$$

其中 $\boldsymbol{x}$ 可能是标量或者向量，可能是离散的或者连续的。函数 $g(\boldsymbol{x})$ 可以看成是系数，确保了概率分布的归一化，因此满足

$$g(\boldsymbol{\eta})\int h(\boldsymbol{x})\exp\{\boldsymbol{\eta}^{\top}\boldsymbol{u}(\boldsymbol{x})\}\,\mathrm{d}\boldsymbol{x} = 1 \tag{2.19}$$

考虑伯努利分布的指数族分布形式

$$p(x \mid \mu) = \mathrm{Bern}(x \mid \mu) = \mu^x(1-\mu)^{1-x} \tag{2.20}$$

将右侧写成对数的指数形式

$$\begin{aligned} p(x \mid \mu) &= \exp\{x\log\mu + (1-x)\log(1-\mu)\} \\ &= (1-\mu)\exp\left\{\log\left(\frac{\mu}{1-\mu}\right)\right\} \end{aligned} \tag{2.21}$$

与公式2.18比较，可以看出

$$\eta = \log\left(\frac{\mu}{1-\mu}\right) \tag{2.22}$$

从中解出 μ, 得 $\mu = \sigma(x)$，其中

$$\sigma(x) = \frac{1}{1+\exp(-\eta)} \tag{2.23}$$

这被称为 logistic sigmoid 函数。因此，伯努利分布的指数族分布形式为

$$p(x \mid \mu) = \sigma(-\eta)\exp(\eta x) \tag{2.24}$$

其中我们使用了 $1-\sigma(\eta) = \sigma(-\eta)$。

第三章　线性回归模型

给定一个由 N 个观测值 $\{\boldsymbol{x}_n\}$ 组成的数据集，其中 $n = 1, \ldots, N$，以及对应的目标值 $\{t_n\}$，我们的目标是预测给定的 $\boldsymbol{x}$ 值的情况下，t 的值。最简单的方法是直接建立一个适当的函数 $y(\boldsymbol{x})$，对于新的输入 $\boldsymbol{x}$，这个函数能够直接给出对应的 t 的预测。更一般地，从一个概率分布 $p(t \mid \boldsymbol{x})$ 建模，表示对于给定一个 $\boldsymbol{x}$，对于 t 的不确定性。

3.1　线性函数模型

最简单的模型是对输入变量的线性组合

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = \omega_0 + \omega_1 x_1 + \ldots + \omega_D x_D \tag{3.1}$$

其中 $\boldsymbol{x} = (x_1, \ldots, x_D)^\top$，这通常被简单的称为**线性回归**。这个模型的关键性质是它是参数 $\omega_0, \ldots, \omega_D$ 的一个线性函数。

另外，我们也可以扩展模型的类别：将输入变量固定的非线性函数进行组合，形式为

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = \omega_0 + \sum_{j=1}^{M-1} \omega_j \phi_j(\boldsymbol{x}) \tag{3.2}$$

其中 $\phi_j(\boldsymbol{x})$ 被称为基函数。通过把下标 j 的最大值记作 $M - 1$，模型中总参数个数为 M 个。参数 ω_0 被称为**偏置参数**。若定义一个额外的基函数 $\phi_0(\boldsymbol{x}) = 1$，这时

$$y(\boldsymbol{x}, \boldsymbol{\omega}) = \sum_{j=1}^{M-1} \omega_j \phi_j(\boldsymbol{x}) = \boldsymbol{\omega}^\top \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.3}$$

其中 $\boldsymbol{\omega} = (\omega_0, \ldots, \omega_{M-1})^\top$，$\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^\top$

3.1.1 最大似然与最小平方误差

假设目标变量 t 由确定的函数 $y(\boldsymbol{x}, \boldsymbol{\omega})$ 给出，这个函数被附加了高斯噪声，即

$$t = y(\boldsymbol{x}, \boldsymbol{\omega}) + \epsilon \tag{3.4}$$

其中，ϵ 是一个均值为零，精度 (方差的倒数) 为 β 的随机变量。因此，有

$$p(t \mid \boldsymbol{x}, \boldsymbol{\omega}, \beta) = \mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{\omega}), \beta^{-1}) \tag{3.5}$$

考虑有一个数据集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，对应的目标值为 $t_1, \ldots, t_N$，把目标值 $\{t_n\}$ 组成一个列向量，记作 $\boldsymbol{t}$。似然函数为

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{\omega}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{\omega}^\top \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \tag{3.6}$$

省略符号 $\boldsymbol{X}$, 并取似然对数，使用一元高斯分布的形式，有

$$\begin{aligned} \log p(\boldsymbol{t} \mid \boldsymbol{\omega}, \beta^{-1}) &= \sum_{n=1}^{N} \log \mathcal{N}(t_n \mid \boldsymbol{\omega}^\top \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \\ &= \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi) - \beta E_D(\boldsymbol{\omega}) \end{aligned} \tag{3.7}$$

其中平方和误差函数被定义为

$$E_D(\boldsymbol{\omega}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \boldsymbol{\omega}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.8}$$

之后，最大化似然函数。首先关于 $\boldsymbol{\omega}$ 最大化。在这里，最大化似然函数与最小化平方和误差等价。对 $\boldsymbol{\omega}$ 求梯度，得

$$\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{t} \mid \boldsymbol{\omega}, \beta) = \beta \sum_{n=1}^{N} \{t_n - \boldsymbol{\omega}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n)^\top \tag{3.9}$$

令这个梯度等于零，可得

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\boldsymbol{x}_n)^\top - \boldsymbol{\omega}^\top \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n) \phi(\boldsymbol{x}_n)^\top \right) \tag{3.10}$$

求解 $\boldsymbol{\omega}$，我们有

$$\boldsymbol{\omega}_{\mathrm{ML}} = (\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\top}\boldsymbol{t} \tag{3.11}$$

这里，$\boldsymbol{\Phi}$ 是一个 $N \times M$ 的矩阵，它的元素为 $\boldsymbol{\Phi}_{nj} = \phi_j(\boldsymbol{x}_n)$，即

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} \tag{3.12}$$

式中

$$\boldsymbol{\Phi}^{\dagger} \equiv (\boldsymbol{\Phi}^{\top}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi} \tag{3.13}$$

被称为 $\boldsymbol{\Phi}$ 矩阵的 Moore-Penrose *伪逆矩阵*，可以被看成是逆矩阵对非方阵的一种推广。实际上，若 $\boldsymbol{\Phi}$ 是方阵并且可逆，$\boldsymbol{\Phi}^{\dagger} = \boldsymbol{\Phi}^{-1}$。

3.1.2 最小平方误差的几何解释

3.1.3 正则化的最小平方误差

3.2 偏差-方差分解

3.3 贝叶斯线性回归

3.3.1 参数分布

3.3.2 预测分布

第四章　线性分类模型

第五章　核方法

第六章　稀疏核机

6.1　最大边缘分类器

在线性模型的二分类问题中，模型的形式为

$$y(\boldsymbol{x}) = \boldsymbol{\omega}^{\top}\phi(\boldsymbol{x}) + b \tag{6.1}$$

其中 $\phi(\boldsymbol{x})$ 表示一个固定的特征空间变换。训练数据由 N 个输入向量 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ 组成，对应的目标值为 $t_1, \ldots, t_N$，其中 $t_n \in \{-1, 1\}$，新的数据点 $\boldsymbol{x}$ 根据 $y(\boldsymbol{x})$ 的符号进行分类。

现在，假设数据在特征空间中是线性可分的，即至少存在一个参数 $\boldsymbol{\omega}$ 和 b，使得对于所有 $t_n = +1$ 的点，公式6.1都满足 $y(\boldsymbol{x}) > 0$，对于所由 $t_n = -1$ 的点，都满足 $y(\boldsymbol{x}) < 0$，从而对于所有的数据点 $t_n y(\boldsymbol{x}_n) > 0$。

有许多中方式可以将数据集分开，我们应该寻找泛化错误最小的那个解。**支持向量机** (SVM) 解决这个问题的方法是：选择使得边缘最大化的那个决策边界。

边缘的概念被定义为决策边界与任意样本之间的最小距离，如图所示

第七章　图模型

7.1　贝叶斯网络

第八章　混合模型和 EM

8.1　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:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{k} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2 \tag{8.1}$$

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8.1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

$$r_{nk}=\begin{cases}1, & \text{如果}k=\arg\min_j \|\boldsymbol{x}_n-\boldsymbol{\mu}_j\|^2\\ 0, & \text{其他情况}\end{cases} \tag{8.2}$$

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

$$2\sum_{n=1}^{N} r_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)=0 \tag{8.3}$$

解出 $\boldsymbol{\mu}_k$ 的值，结果为

$$\boldsymbol{\mu}_k=\frac{\sum_n r_{nk}\boldsymbol{x}_n}{\sum_n r_{nk}} \tag{8.4}$$

8.4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。*

8.2 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

$$p(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{8.5}$$

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

$$p(z_k=1)=\pi_k \tag{8.6}$$

其中参数 $\{\pi_k\}$ 必须满足

$$0\leqslant\pi_k\leqslant 1 \tag{8.7}$$

以及

$$\sum_{k=1}^{K} \pi_k = 1 \tag{8.8}$$

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

$$p(\boldsymbol{z}) = \prod_{k=1}^{k} \pi_k^{z_k} \tag{8.9}$$

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

$$p(\boldsymbol{x} \mid z_k = 1) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{8.10}$$

类似的也可以写成

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{8.11}$$

$\boldsymbol{x}$ 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

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{z}) p(\boldsymbol{x} \mid \boldsymbol{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{8.12}$$

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

$$\gamma(z_k) = p(z_k = 1 \mid \boldsymbol{x}) = \frac{p(z_k = 1) p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1) p(\boldsymbol{x} \mid z_j = 1)} \tag{8.13}$$

$$= \frac{\pi_k p(\boldsymbol{x} \mid z_k = 1)}{\sum_{j=1}^{K} \pi_j p(\boldsymbol{x} \mid z_j = 1)} \tag{8.14}$$

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8.1 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

$$\log p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{8.15}$$

最大化高斯混合模型的对数似然函数式8.15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*。

8.3 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

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\theta}) \tag{8.16}$$

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

$$\log p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q \parallel p) \tag{8.17}$$

其中

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{8.18}$$

$$\mathrm{KL}(p \parallel q) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \log \left\{ \frac{p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \right\} \tag{8.19}$$

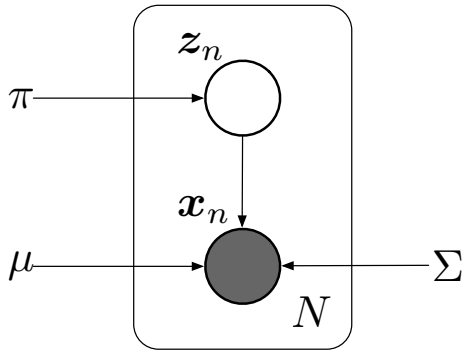

图 8.1: 一组 N 个独立同分布数据点 $\{\boldsymbol{x}_n\}$ 的高斯混合模型的图表示，对应的潜在变量为 $\{\boldsymbol{z}_n\}$，其中 $n = 1, 2, \ldots, N$。

$\mathcal{L}(q, \boldsymbol{\theta})$ 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 $\mathrm{KL}(p \,\|\, q) \geqslant 0$，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 $\log p(\boldsymbol{X} \mid \boldsymbol{\theta})$ 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

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \\ &\quad - \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \log p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) + \text{常数} \end{aligned} \tag{8.20}$$

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 如算法1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

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

5: **until** 对数似然函数收敛或者参数值收敛

8.4 EM 算法实例

8.4.1 用于混合高斯模型的 EM

现在考虑将 EM 算法的隐变量观点用于一个具体的例子，即高斯混合模型。我们的目标是是最到化对数似然函数8.15，这是使用观测数据集 $\boldsymbol{X}$ 计算的。这种情况比单一的高斯困难，因为求和出现在了对数运算内部。假设除了观测数据集 $\boldsymbol{X}$，还有对应的离散变量 $\boldsymbol{Z}$。现在考虑对完整数据 $\{\boldsymbol{X}, \boldsymbol{Z}\}$ 最大化。完整数据集的似然函数的形式为

$$p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \tag{8.21}$$

其中 z_{nk} 表示 $\boldsymbol{z}_n$ 的第 k 个分量。取对数，有

$$\log p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{8.22}$$

现在将完全数据的对数似然对 $\boldsymbol{Z}$ 的后验概率分布求期望。后验概率分布为

$$p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \tag{8.23}$$

在这个分布下, z_{nk} 的期望为

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}}[z_{nk}] &= \sum_{\boldsymbol{z}_1}\cdots\sum_{\boldsymbol{z}_N} z_{nk}p(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{z}_1} p(\boldsymbol{z}_1\mid\boldsymbol{x}_n,\boldsymbol{\theta})\cdots\sum_{\boldsymbol{z}_n} z_{nk}p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\theta})\cdots\sum_{\boldsymbol{z}_N} p(\boldsymbol{z}_N\mid\boldsymbol{x}_n,\boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{z}_n} z_{nk}p(\boldsymbol{z}_n\mid\boldsymbol{x}_n,\boldsymbol{\theta}) \\ &= p(z_{nk}=1\mid\boldsymbol{x}_n,\boldsymbol{\theta}) \end{aligned} \tag{8.24}$$

其中 $\boldsymbol{\theta}=(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})$。利用贝叶斯公式，有

$$\begin{aligned} p(z_{nk}=1\mid\boldsymbol{x}_n,\boldsymbol{\theta}) &= \frac{p(z_{nk}=1)p(\boldsymbol{x}_n\mid z_{nk}=1)}{\sum_{j=1}^{K}p(z_{nj}=1)p(\boldsymbol{x}_n\mid z_{nj}=1)} \\ &= \frac{\pi_k\mathcal{N}(\boldsymbol{x}_n\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x}_n\mid\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} \\ &\equiv \gamma(z_{nk}) \end{aligned} \tag{8.25}$$

$\gamma(z_{nk})$ 被定义为数据点 $\boldsymbol{x}_n$ 种含有来自于第 k 个高斯分布的“成分”。于是，完整数据的对数似然的期望值为

$$\mathbb{E}_{\boldsymbol{Z}}\left[\log p(\boldsymbol{X},\boldsymbol{Z}\mid\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})\right]=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\log\pi_k+\log\mathcal{N}(\boldsymbol{x}_n\mid\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\} \tag{8.26}$$

我们使用旧的参数 $\left\{\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old},\boldsymbol{\pi}^{old}\right\}$ 计算 $\gamma(z_{nk})(E$ 步骤$)$；之后保持 $\gamma(z_{nk})$ 不变，关于 $\boldsymbol{\mu}_k$,$\boldsymbol{\Sigma}_k$, π_k 最大化 (M 步骤)，得到新的 $\{\boldsymbol{\mu}^{new},\boldsymbol{\Sigma}^{new},\boldsymbol{\pi}^{new}\}$。

在进行 M 步骤之前，需要先参考一些关于矩阵求导数的运算，具体如下

$$\sum_{i=1}^{N}\boldsymbol{x}_i^{\top}\boldsymbol{S}\boldsymbol{x}_i=\mathrm{Tr}(\boldsymbol{S}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\top}) \tag{8.27}$$

$$\frac{\partial\mathrm{Tr}(\boldsymbol{A}\boldsymbol{B})}{\partial\boldsymbol{A}}=\boldsymbol{B}^{\top} \tag{8.28}$$

$$\frac{\partial}{\partial\boldsymbol{A}}\log|\boldsymbol{A}|=(\boldsymbol{A}^{-1})^{\top} \tag{8.29}$$

现在关于 π_k 最大化。注意到由于 $\sum_{k=1}^{K}\pi_k = 1$ 的限制，可以使用拉格朗日乘数法进行优化。构造拉格朗日函数为

$$\mathcal{L}(\boldsymbol{\pi}, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log\pi_k + \lambda(1 - \sum_{k=1}^{K}\pi_k) \tag{8.30}$$

对 π_k 求导，并令其等于零，有

$$\frac{1}{\lambda}\sum_{n=1}^{N}\gamma(z_{nk}) = \pi_k, \quad k = 1, \ldots, N \tag{8.31}$$

又由 $\sum_{k=1}^{K}\pi_k = 1$，得出 $\lambda = N$，所以更新后的 pi_k 为

$$\pi_k^{new} = \frac{N_k}{N} \tag{8.32}$$

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

关于 $\boldsymbol{\mu}_k$ 最大化。注意到，公式8.26中包含 $\boldsymbol{\mu}_k$ 的项是

$$\begin{aligned}&\sum_{n=1}^{N}\gamma(z_{nk})\log\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\\ &= \sum_{n=1}^{N}\gamma(z_{nk})\left\{-\frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}_k^{-1}| - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}_k)\right\}\end{aligned} \tag{8.33}$$

对 $\boldsymbol{\mu}_k$ 求导，并令其等于零，得

$$\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) = 0 \tag{8.34}$$

化简

$$\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_n = \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\boldsymbol{u}_k \tag{8.35}$$

两边同乘 $\boldsymbol{\Sigma}_k$，得

$$\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n = \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\mu}_k = N_k\boldsymbol{\mu}_k \tag{8.36}$$

所以，得到新的 $\boldsymbol{\mu}_k^{new}$ 为

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \tag{8.37}$$

关于 $\boldsymbol{\Sigma}_k$ 最大化。将公式8.33关于 $\boldsymbol{\Sigma}_k^{-1}$ 求导，并令其导数等于零。具体过程如下

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \frac{1}{N_k}\frac{\partial}{\partial\boldsymbol{\Sigma}_k^{-1}}\mathrm{Tr}\left(\boldsymbol{\Sigma}_k^{-1}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\right)\\ &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\end{aligned} \tag{8.38}$$

所以，新的 $\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})^\top$。

总结一下，高斯混合分布的参数估计如下

- 初始化均值 $\boldsymbol{\mu}_k$，协方差 $\boldsymbol{\Sigma}_k$ 和混合系数 π_k，计算对数似然的初始值
- *E 步骤*。使用当前参数，计算每个数据点的成分 $\gamma(z_{nk})$

$$\gamma(z_{nk}) = \frac{\pi_k\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{8.39}$$

- *M 步骤*。使用当前的 $\gamma(z_{nk})$ 重新估计参数。

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \tag{8.40}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})^\top \tag{8.41}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{8.42}$$

其中

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}) \tag{8.43}$$

- 计算对数似然函数

$$\log p(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) \right\} \tag{8.44}$$

检查参数或者对数似然函数的收敛性。若没有满足收敛条件，返回 E 步骤。

8.4.2 伯努利分布的混合

第九章　近似推断

第十章　采样方法

10.1　基本算法

10.1.1　概率分布采样

10.1.2　拒绝采样

10.1.3　重要性采样

10.2　马尔可夫链蒙特卡罗方法

10.2.1　马尔可夫链

10.2.2　Metropolis-Hastings 算法

10.3　吉布斯采样

第十一章　连续潜在变量

11.1　主成分分析

主成分分析，或者称为 PCA，是一种广泛使用的技术，应用的领域包括维度降低、有损数据压缩、特征抽取、数据可视化。有两种经常使用的 PCA 的定义：

- PCA 被定义为数据在被称为**主子空间**的低维线性空间上的投影，使得投影数据的方差最大化
- 也可以被定义为使得平均投影代价最小的线性投影，平均投影代价是指数据点和它们的投影之间的平均平方距离

11.1.1　最大方差形式

考虑一组观测数据集 $\boldsymbol{x}_n$，其中 $n = 1, \ldots, N$，$\boldsymbol{x}_n$ 是一个 D 维欧几里得空间中的变量。考虑将数据投影到维度是 $M < D$ 的空间中。

首先，考虑在一维空间 $(M = 1)$ 上的投影。可以使用 D 维向量 $\boldsymbol{u}_1$ 定义投影方法。为了计算方便和不失一般性，假定选择一个单位向量，而且 $\boldsymbol{u}_1^\top \boldsymbol{u}_1 = 1$，这样每个数据点 $\boldsymbol{x}_n$ 被投影到一个标量值 $\boldsymbol{u}_1^\top \boldsymbol{x}_n$ 上。被投影数据的均值是 $\boldsymbol{u}_1^\top \overline{\boldsymbol{x}}$，其中 $\overline{\boldsymbol{x}}$ 是样本集合的均值，形式为

$$\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \tag{11.1}$$

投影数据的方差为

$$\frac{1}{N}\sum_{n=1}^{N}\{\boldsymbol{u}_1^\top \boldsymbol{x}_n - \boldsymbol{u}_1^\top \overline{\boldsymbol{x}}\}^2 = \boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 \tag{11.2}$$

其中 $\boldsymbol{S}$ 是协方差矩阵，定义为

$$\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \overline{\boldsymbol{x}})(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top \tag{11.3}$$

现在关于 $\boldsymbol{u}_1$ 最大化投影方差 $\boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1$，包含约束条件 $\|\boldsymbol{u}_1\| = 1$。可以采用拉格朗日乘数法，转化为如下的无约束优化问题

$$\boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 + \lambda_1(1 - \boldsymbol{u}_1^\top \boldsymbol{u}_1) \tag{11.4}$$

通过令其关于 $\boldsymbol{u}_1$ 导数等于零，可以看到驻点满足

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \tag{11.5}$$

这表明 $\boldsymbol{u}_1$ 一定是 $\boldsymbol{S}$ 的一个特征向量。如果左乘 $\boldsymbol{u}_1^\top$，可以看到方差为

$$\boldsymbol{u}_1^\top \boldsymbol{S}\boldsymbol{u}_1 = \lambda_1 \tag{11.6}$$

因此当我们将 $\boldsymbol{u}_1$ 设置为与具有最大特征值 λ_1 的特征向量时，方差会达到最大值。这个特征向量被称为**第一主成分**。

可以使用类似的方法定义额外的主成分，方法是：在与所有已经考虑过的方向正交的方向中，选择使得投影数据方差最大化的方向。例如，我们考虑 M 维投影空间的情况，那么最大化投影数据方差的方向由协方差矩阵 $\boldsymbol{S}$ 的 M 个特征向量 $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_M$ 定义，对应的 M 个特征值是 $\lambda_1, \lambda_2, \ldots, \lambda_M$。

总结一下，主成分分析涉及计算数据集的均值 $\overline{\boldsymbol{x}}$ 和协方差矩阵 $\boldsymbol{S}$，然后寻找与 $\boldsymbol{S}$ 对应的 M 个最大特征值的特征向量。寻找一个 $D \times D$ 矩阵的完整的特征向量分解的代价为 $\boldsymbol{O}(D^3)$。若只将数据投影到前 M 个主成分中，那么只需要寻找前 M 个特征向量和特征值，这可以使用更高效的算法，时间复杂度为 $\boldsymbol{O}(MD^2)$，或者也可以使用 EM 算法。

11.1.2 最小误差形式

PCA 的另一种形式，基于最小投影误差。先引入一个完备的单位正交集合 $\{\boldsymbol{u}_i\}$，其中 $i = 1, \ldots, D$，且满足

$$\boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij} \tag{11.7}$$

由于是完备的，因此每个数据点可以唯一的表示为一个线性组合，即

$$\boldsymbol{x}_n = \sum_{i=1}^{D} \alpha_{ni} \boldsymbol{u}_i \tag{11.8}$$

其中，系数 α_{ni} 对于不同的数据点来说是不同的。这相当于将坐标系旋转到了一个由 $\{\boldsymbol{u}_i\}$ 表示的新坐标系。我们有 $\alpha_{nj} = \boldsymbol{x}_n^\top \boldsymbol{u}_j$，不失一般性，我们有

$$\boldsymbol{x}_n = \sum_{i=1}^{D} (\boldsymbol{x}_n^\top \boldsymbol{u}_i) \boldsymbol{u}_i \tag{11.9}$$

我们的目标是限定数量 $M < D$ 个变量的一种表示来近似数据点，这对应于在低维子空间的一个投影。因此，我们可以用下式来近似每个数据点 $\boldsymbol{x}_n$

$$\tilde{\boldsymbol{x}}_n = \sum_{i=1}^{M} z_{ni} \boldsymbol{u}_i + \sum_{i=M+1}^{D} b_i \boldsymbol{u}_i \tag{11.10}$$

其中 $\{z_{nj}\}$ 依赖于特定的数据点，而 $\{b_i\}$ 是常数，对于所有的数据点都相同。可以选择 $\{\boldsymbol{u}_i\}$，$\{z_{ni}\}$，$\{b_i\}$，从而最小化由维度降低带来的失真。失真的度量，可以使用原始数据点与它近似点 $\boldsymbol{x}_n$ 之间的平方距离。因此目标函数是

$$J = \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 \tag{11.11}$$

考虑关于 $\{z_{ni}\}$ 的最小化。代入 $\tilde{\boldsymbol{x}}_n$ 的表达式，然后令它关于 z_{nj} 的导数等于零，有

$$z_{nj} = \boldsymbol{x}_n^\top \boldsymbol{u}_j \tag{11.12}$$

其中 $j = 1, \ldots, M$；类似地，令 J 关于 b_i 的导数等于零，我们有

$$b_j = \overline{\boldsymbol{x}}^\top \boldsymbol{u}_j \tag{11.13}$$

其中，$j = M+1, \ldots, D$。将 z_{nj} 与 b_j 代入到式11.10中，使用式11.9中 $\boldsymbol{x}_n$ 的展开式，可以得到

$$\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n = \sum_{i=M+1}^{D} \left\{ (\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top \boldsymbol{u}_i \right\} \boldsymbol{u}_i \tag{11.14}$$

然后就得到了失真度量 J 的表达式，它是一个纯粹的关于 $\{\boldsymbol{u}_i\}$ 的函数，形式为

$$J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} (\boldsymbol{x}_n^\top \boldsymbol{u}_i - \overline{\boldsymbol{x}}^\top \boldsymbol{u}_i)^2 = \sum_{i=M+1}^{D} \boldsymbol{u}_i^\top \boldsymbol{S} \boldsymbol{u}_i \tag{11.15}$$

11.1.3 高维数据的 PCA

在 PCA 的一些应用中，会出现数据维度远大于数据样本的个数。例如将 PCA 应用到由几百张图片构成的数据集合，此时，每个图片的维度为几百万维，远大于样本个数。若数据维度为 D，样本个数为 N，$(N << D)$，N 个数据点定义了一个最多为 $N-1$ 维的子空间，运行 PCA 会发现至少 $N-D+1$ 个特征值为零；另外，寻找 $D \times D$ 矩阵的特征向量算法的计算复杂度为 $\boldsymbol{O}(D^3)$，对于高维度的数据来说，在计算上也不可行。

可以使用下面的方法解决这个问题。$\boldsymbol{X}$ 定义为 $(D \times D)$ 维中心数据矩阵，它的第 n 行为 $(\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top$。这样，数据的协方差矩阵可以写作 $\boldsymbol{S} = \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X}$，对应的特征向量方程变成了

$$\frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{11.16}$$

将两侧左乘一个 $\boldsymbol{X}$，可得

$$\frac{1}{N} \boldsymbol{X} \boldsymbol{X}^\top (\boldsymbol{X} \boldsymbol{u}_i) = \lambda_i (\boldsymbol{X} \boldsymbol{u}_i) \tag{11.17}$$

若定义 $\boldsymbol{v}_i = \boldsymbol{X} \boldsymbol{u}_i$，就有

$$\frac{1}{N} \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \tag{11.18}$$

它是 $N \times N$ 矩阵 $\frac{1}{N} \boldsymbol{X} \boldsymbol{X}^\top$ 的一个特征向量方程，且这个矩阵与原始的协方差矩阵有相同的 $N-1$ 个特征值，原始协方差矩阵本身还有额外 $D-N+1$ 个

值为零的特征值。我们可以在低维空间中解决这个问题，计算代价是 $\boldsymbol{O}(N^3)$ 而不是 $\boldsymbol{O}(D^3)$。

为了得到原始协方差矩阵的特征向量，可以将公示11.18两侧左乘 $\boldsymbol{X}^\top$，得

$$(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{v}_i) = \lambda_i(\boldsymbol{X}^\top\boldsymbol{v}_i) \tag{11.19}$$

从中，我们可以看到，$(\boldsymbol{X}^\top\boldsymbol{v}_i)$ 是原始数据协方差矩阵的一个特征向量，对应的特征值是 λ_i。但是，这个特征向量的长度不一定为 1，即 $\boldsymbol{u}_i \propto \boldsymbol{X}^\top\boldsymbol{v}_i$。将此向量归一化，就可以得到 $\boldsymbol{u}_i$，且 $\|\boldsymbol{u}_i\| = 1$。假设 $\boldsymbol{v}_i$ 是单位向量，可以通过以下方式归一化

$$\boldsymbol{u}_i = \frac{1}{(N\lambda_i)^{\frac{1}{2}}}\boldsymbol{X}^\top\boldsymbol{v}_i \tag{11.20}$$

总结一下，使用这种方法，首先计算 $\boldsymbol{X}\boldsymbol{X}^\top$，然后找到它的特征向量和特征值，之后使用公式11.20计算原始数据协方差矩阵的特征向量。

11.2 概率 PCA

11.3 核 PCA

11.4 非线形隐变量模型

11.4.1 独立成分分析

11.4.2 自编码器

11.5 非线性流行建模

第十二章　顺序数据

12.1　马尔可夫模型

考虑一组数据 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$，它们之间的联合概率分布使用乘积规则可以表示为如下

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1) \prod_{n=2}^{N} p(\boldsymbol{x}_n \mid x_1, \ldots, \boldsymbol{x}_{n-1}) \tag{12.1}$$

如果假设右侧的每个条件概率分布只与最近的一次观测有关，独立于其他所有之前的观测，那么就得到了**一阶马尔科夫链**。这个模型中, N 此数据观测的联合概率分布为

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1) \prod_{n=2}^{N} p(\boldsymbol{x}_n \mid x_{n-1}) \tag{12.2}$$

在这种模型的大部分应用中，条件概率分布 $p(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1})$ 被限制为相等的，这个模型被称为**同质马尔可夫链**。

虽然这比数据间独立的模型要一般一些，但是还是非常受限。对一些序列数据来说，连续若干个观测变量会对下一次预测提供重要的信息。一种让更早的观测产生影响的方法是使用更高阶的马尔可夫链，例如**二阶马尔可夫链**，其联合概率分布为

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = p(\boldsymbol{x}_1) p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) \prod_{n=3}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}, \boldsymbol{x}_{n-2}) \tag{12.3}$$

假如我们想构造任意阶数不受马尔可夫假设限制的序列模型时，同时能够使用较少的参数，可以引入额外的隐变量来使得复杂的模型能够从简

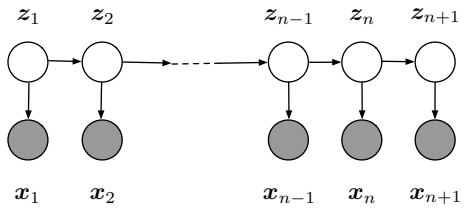

图 12.1: 使用隐变量的马尔可夫链来表示顺序数据，每个观测值都一个对应的隐变量的状态为条件。

单的模型中构建。对于每个观测值 $\boldsymbol{x}_n$，引入一个隐变量 $\boldsymbol{z}_n$(类型和维度可以与 $\boldsymbol{x}_n$ 不同)。现在假设隐变量 $\boldsymbol{z}_n$ 之间构成马尔可夫链，得到的图模型如图12.1所示。它满足如下的条件独立性质：在给定 $\boldsymbol{z}_n$ 的条件下，$\boldsymbol{z}_{n-1}$ 与 $\boldsymbol{z}_{n+1}$ 条件独立

$$\boldsymbol{z}_{n-1} \perp \boldsymbol{z}_{n+1} \mid \boldsymbol{z}_n \tag{12.4}$$

这个模型的联合概率分布为

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) = p(\boldsymbol{z}_1) \left[\prod_{n=2}^{N} p(\boldsymbol{z}_n \mid \boldsymbol{z}_{n-1}) \right] \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \boldsymbol{z}_n) \tag{12.5}$$

对顺序数据来说，图12.1描述了两个重要模型。如果隐变量是离散的，那么我们得到了隐马尔可夫模型或者 HMM。在 HMM 中，观测变量可以是离散的或者连续的，并且可以使用许多不同条件概率分布进行建模。如果隐变量和观测变量都是高斯变量 (节点的条件概率分布对于父节点的依赖是线性高斯的形式)，那么就得到了线性动态系统 (linear dynamical system)。

12.2 隐马尔可夫模型

12.3 线性动态系统

第十三章　组合模型