模式识别与机器学习

Pattern Recognition & Machine Learning

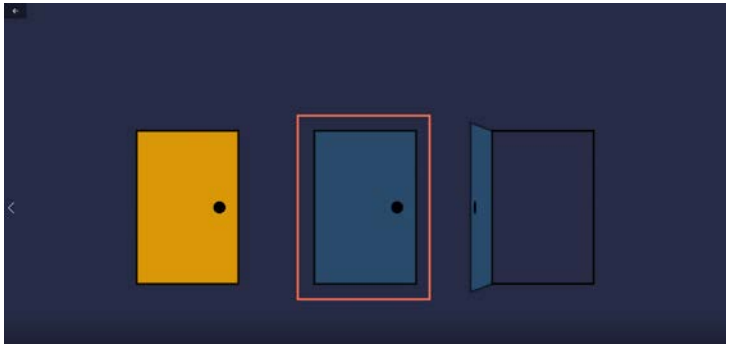
第2讲 贝叶斯学习基础

- 本节学习目标

  - ✓掌握贝叶斯公式在机器学习中的应用思路
  - ✓能够熟练运用贝叶斯决策方法
  - ✓明确分类器相关的基本概念
  - ✓掌握基于高斯分布的贝叶斯分类器
  - ✓理解朴素贝叶斯分类器
  - ✓能够熟练运用各种参数估计方法

目录

- 有趣的概率



- 贝叶斯公式

例：假设某个动物园里的雌性和雄性熊猫的比例是4：6，雌性熊猫中90%的熊猫是干净整洁的，雄性熊猫中20%是干净整洁的。

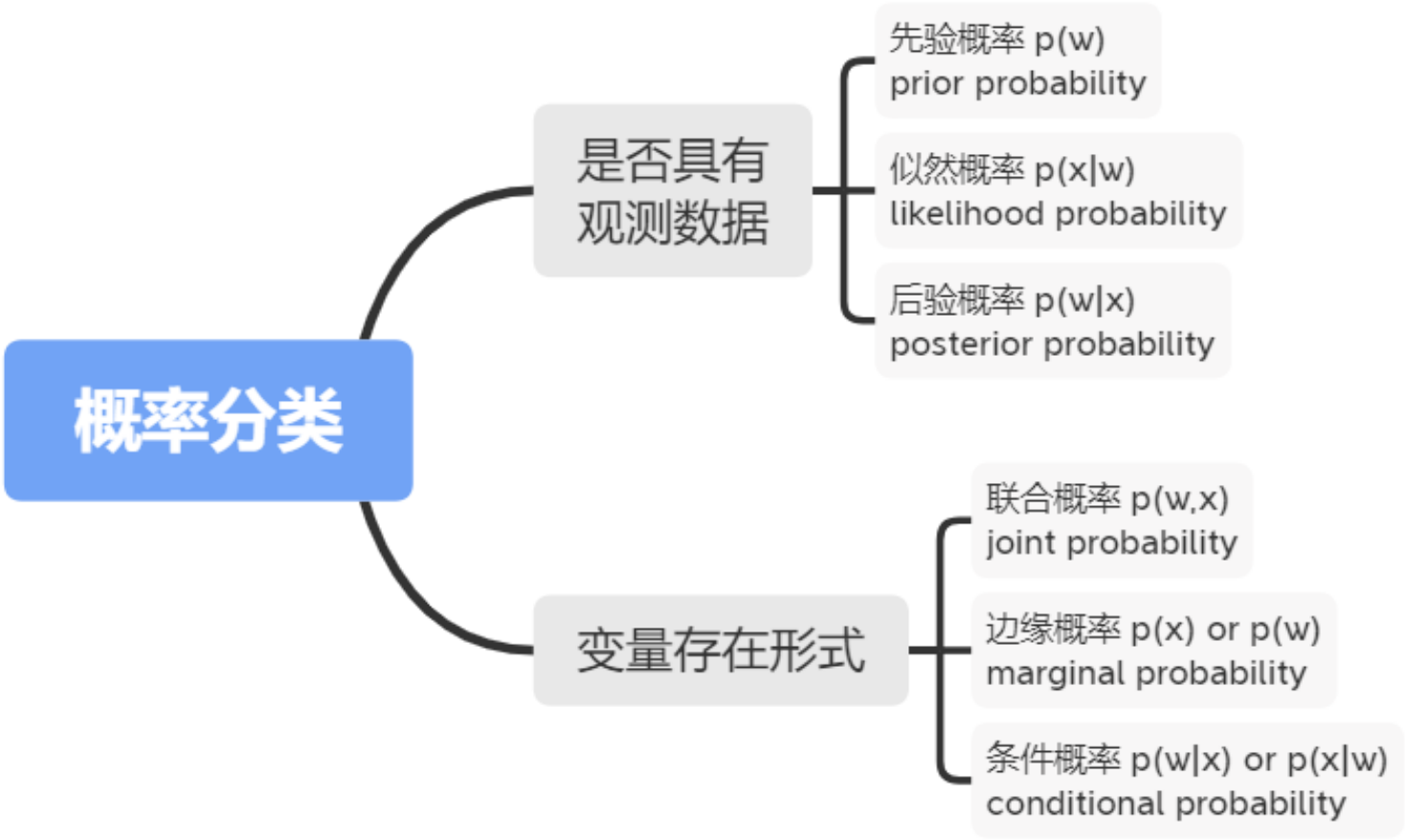
1. 求解“**正向概率**”：

在动物园中看到一只干净整洁的雄性熊猫的概率是多少？

2. 求解“**逆向概率**”：

如果看到一只熊猫是干净整洁的，它是雄性的概率是多少？

假设x表示观测变量，w表示模型参数：



假设x表示观测变量，w表示模型参数：

联合概率 = 条件概率 X 边缘概率

$$p(w,x) = p(x|w)p(w) = p(w|x)p(x)$$

某变量的边缘概率等于

$$p(x) = \sum_{w} p(w,x) \qquad p(w) = \sum_{x} p(w,x)$$

贝叶斯公式

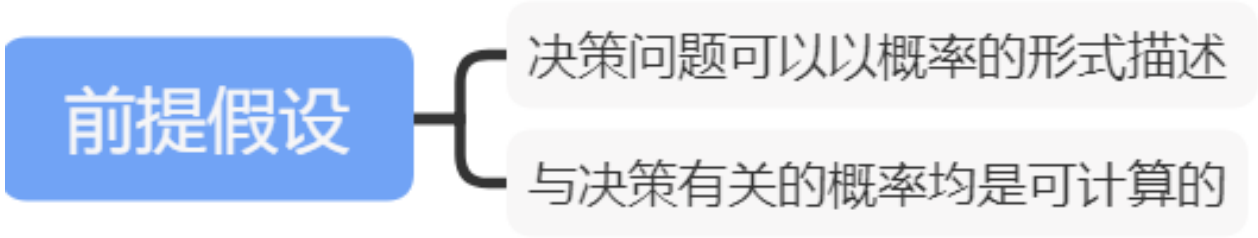
$$p(w|x) = \frac{p(w,x)}{p(x)} = \frac{p(x|w)p(w)}{\sum_{w} p(w,x)}$$

目录

- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计

- 贝叶斯决策

贝叶斯决策（Bayesian decision）是概率框架下实施决策的基本方法，它通过综合考虑决策的后验分布和错误决策的损失来做出决策。其中，贝叶斯公式被用于计算后验分布。贝叶斯决策的前提是假设：

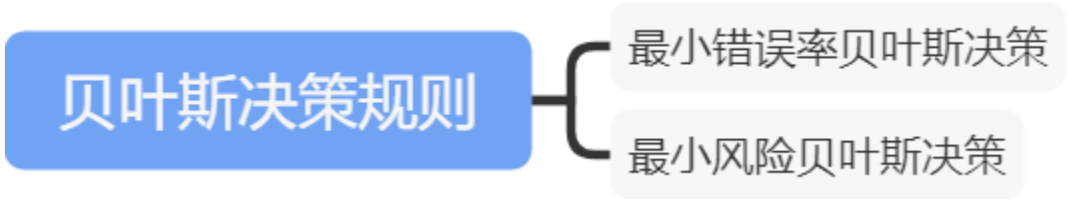


例：根据熊猫的形态特征来判断熊猫的性别。
设w表示性别，$w=1$表示雌性，$w=2$表示雄性。
熊猫为雌性的先验概率为$p(w=1)$，为雄性的先验概率为$p(w=2)$，则

$$p(w=1)+p(w=2)=1$$

假设x表示观测变量，刻画熊猫的形态特征，$x=1$表示熊猫是干净整洁的，反之$x=0$.

在给定决策问题的概率描述（先验概率和似然概率）之后，贝叶斯决策使用贝叶斯公式推导出性别变量w的后验分布$p(w|x)$，然后通过决策规则做出决策。



- 最小错误率贝叶斯决策

决策的平均错误率尽可能地小

熊猫分类问题**分类错误**：

- 某样本类别是雄性$w=2$，但被分为雌性$w=1$;
- 某样本类别是雌性$w=1$，但被分为雄性$w=2$;

$$p(error|x)=\begin{cases}p(w=1|x) & \text{如果}x\text{被判定为雄性}w=2\\ p(w=2|x) & \text{如果}x\text{被判定为雌性}w=1\end{cases}$$

决策的平均错误率：

$$p(error) = \int_{-\infty}^{\infty} p(error|x)p(x)dx$$

$$= \int_{R_1} p(x,w=2)dx + \int_{R_2} p(x,w=1)dx$$

- 最小错误率贝叶斯决策

【例】已知池中有两种鱼，比例为7：3，若随机捞上一条，按照70%和30%概率随机猜测其种类，则整体误差接近于多少？

参考答案：

P(error) = 0.3*0.7+0.7*0.3

对于二分类问题，最小错误率贝叶斯决策：

$$\begin{cases} x\text{被判定为第一类} & \text{如果} p(w=1|x) > p(w=2|x) \\ x\text{被判定为第二类} & \text{如果} p(w=1|x) < p(w=2|x) \end{cases}$$

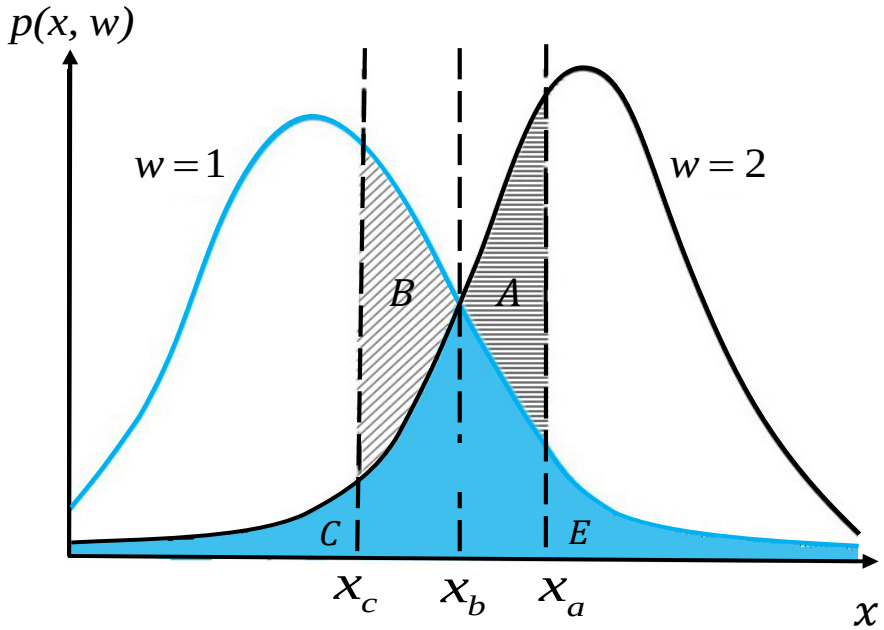


图2-1 二分类的最小错误率贝叶斯决策示意图

（图中近似三角形阴影区域A和B的面积分别表示相比于$x = x_b$为决策边界，$x = x_a$和$x = x_c$作为决策边界所增加的平均错误率）

考虑更一般的多分类问题，假设存在C个类别，将特征空间分为区域$R_1, R_2, \cdots, R_C$。每一类都会错分成其他类，那么决策的平均错误率可表示为

$$
\begin{aligned}
&p(error) \\
&= \left[\int_{R_2} p(x, w=1)dx + \int_{R_3} p(x, w=1)dx + \cdots + \int_{R_C} p(x, w=1)dx\right] \\
&+ \left[\int_{R_1} p(x, w=2)dx + \int_{R_3} p(x, w=2)dx + \cdots + \int_{R_C} p(x, w=2)dx\right] + \cdots \\
&+ \left[\int_{R_1} p(x, w=C)dx + \int_{R_2} p(x, w=C)dx + \cdots + \int_{R_{C-1}} p(x, w=C)dx\right] \\
&= \int_{i=1}^{C} \sum_{j=1, j\neq i}^{C} \int_{R_j} p(x, w=i)dx
\end{aligned}
$$

可能错分的情况存在$C \times (C-1)$种，涉及到的计算很多，所以通常采样计算平均正确率$p(correct)$来计算$p(error)$

$$
\begin{aligned}
&p(error) \\
&= 1 - p(correct) \\
&= 1 - \left[\int_{R_1} p(x, w = 1)dx + \int_{R_2} p(x, w = 2)dx + \cdots + \int_{R_C} p(x, w = C)dx \right] \\
&= 1 - \int_{c=1}^{C} \sum_{R_c} p(x, w = c)dx
\end{aligned}
$$

对于更一般化的多类分类问题，最小错误率决策表示为最大化平均正确率，平均正确率$p(correct)$的计算如下：

$$p(correct) = \int_{c=1}^{C} \sum_{R_c} p(x,w)dx$$

由上式可以看出，最大化$p(correct)$等价于将x判别为联合概率$p(x,w)$最大类别，即决策输出$h(x)$表示为

$$h(x) = \text{argmax}_c\, p(w = c|x)$$

在实际分类应用中，往往不必计算后验概率。根据贝叶斯公式，后验概率可以表示为联合概率除以边缘概率$p(x)$，对于所有类别，分母都是相同的，所以决策时实际上只需比较分子即可。因此只需要计算$p(x|w)p(w)$，将样本判别为其值最大类别。

- 思考与计算

“假设某个动物园里的雌性和雄性熊猫的比例是4：6，雌性熊猫中90%的熊猫是干净整洁的，雄性熊猫中20%是干净整洁的”。计算在该动物园中看到一只干净整洁的雄性熊猫的概率是多少？如果看到一只熊猫是干净整洁的，它是雄性的概率是多少？

- 思考与计算

“假设某个动物园里的雌性和雄性熊猫的比例是4：6，雌性熊猫中90%的熊猫是干净整洁的，雄性熊猫中20%是干净整洁的”。计算在该动物园中看到一只干净整洁的雄性熊猫的概率是多少？如果看到一只熊猫是干净整洁的，它是雄性的概率是多少？

$$p(F)=0.4 \quad p(M)=0.6 \quad p(C|F)=0.9 \quad p(C|M)=0.2$$

$$p(C,M)=p(C|M)\times p(M)=0.2\times 0.6=0.12$$

$$p(M|C)=\frac{p(C|M)\times p(M)}{p(C|M)\times p(M)+p(C|F)\times p(F)}=\frac{0.12}{0.12+0.36}=0.25$$

- 最小风险贝叶斯决策

【例】假设患病白细胞浓度服从均值为2000，标准差为1000的正态分布，未患病白细胞浓度服从均值为7000，标准差为3000的正态分布，患病的人数比例为0.5%，问当白细胞浓度为3100时，应该做出什么决策？

思考：“患病被误诊为正常”与“正常被误诊为患病”哪个损失更大？

- 最小风险贝叶斯决策

最小化决策带来的平均损失，也叫做最小化风险（risk）。

考虑一个多类分类问题,样本的真实类别为第j类，但是被误判为第i类的损失为

$$\lambda_{ij} = \lambda(h(x) = i|w = j)$$

对于C类分类问题，损失矩阵是一个$C \times C$的矩阵$(\lambda_{ij})_{C\times C}$.

根据损失的定义可知，损失矩阵的对角元素通常为0。

平均损失的两重含义：

1. 获得观测值后，决策造成的损失对实际所属类别的各类可能的平均，称为条件风险（conditional risk）：

$$R(h(x)|x) = \sum_i \lambda(h(x)|w = i)p(w = i|x)$$

2. 条件风险对x的数学期望，称为总体风险：

$$R\big(h(x)\big) = \mathbb{E}\big(R(h(x)|x)\big) = \int R(h(x)|x)p(x)dx$$

决策函数：

$$h(x) = \mathrm{argmin}_{\mathrm{i}} \sum_i \lambda(h(x) = j|w = i)p(w = i|x)$$

以二分类问题为例。
标记α_1表示把样本判别为第一类,α_2表示把样本判别为第二类。
二分类问题中的损失矩阵λ_{ij}是一个2×2的矩阵，条件风险为：

$$R(\alpha_1|x)=\lambda_{11}p(w=1|x)+\lambda_{12}p(w=2|x)$$
$$R(\alpha_2|x)=\lambda_{21}p(w=1|x)+\lambda_{22}p(w=2|x)$$

根据最小风险贝叶斯决策规则，如果满足

$$(\lambda_{21}-\lambda_{11})p(w=1|x)>(\lambda_{12}-\lambda_{22})p(w=2|x)$$

或者满足

$$(\lambda_{21}-\lambda_{11})p(x|w=1)p(w=1)>(\lambda_{12}-\lambda_{22})p(x|w=2)p(w=2)$$

则x将被判别为第一类，否则被判别为第二类。

- 最小风险贝叶斯决策&最小错误率贝叶斯决策

假设决策损失定义为$0-1$损失，即

$$\lambda(\alpha_i|w=j)=\begin{cases}0 & i=j\\ 1 & i\neq j\end{cases}$$

此时，条件风险=条件错误率

$$\begin{aligned}R(\alpha_i|x)&=\sum_{j=1}^{C}\lambda(\alpha_i|w=j)p(w=j|x)\\&=\sum_{j\neq i}p(w=j|x)\\&=1-p(w=i|x)\end{aligned}$$

- 最小风险贝叶斯决策&最小错误率贝叶斯决策

【例】假设患病白细胞浓度服从均值为2000，标准差为1000的正态分布，未患病白细胞浓度服从均值为7000，标准差为3000的正态分布，患病的人数比例为0.5%，问当白细胞浓度为3100时，应该做出什么决策？（假设将患病判别为正常的损失是将正常判别为患病的损失的100倍，分别考虑最小错误率和最小风险贝叶斯决策）

• 最小风险贝叶斯决策&最小错误率贝叶斯决策

【例】假设患病白细胞浓度服从均值为2000，标准差为1000的正态分布，未患病白细胞浓度服从均值为7000，标准差为3000的正态分布，患病的人数比例为0.5%，问当白细胞浓度为3100时，应该做出什么决策？（假设将患病判别为正常的损失是将正常判别为患病的损失的100倍，分别考虑最小错误率和最小风险贝叶斯决策）

设 w 表示是否患病，x 表示白细胞浓度，根据题意可以得到

$$p(w=1)=0.5\% \tag{2-3}$$

$$p(w=2)=99.5\% \tag{2-4}$$

$$p(x\mid w=1)=\mathcal{N}(2000,1000^2) \tag{2-5}$$

$$p(x\mid w=2)=\mathcal{N}(7000,3000^2) \tag{2-6}$$

（1）若进行最小错误率贝叶斯决策，则需要根据贝叶斯公式计算随机变量 w 的后验分布，计算结果如下：

$$p(w=1\mid x)=\frac{p(x\mid w=1)p(w=1)}{p(x)}=1.9\% \tag{2-7}$$

$$p(w=2\mid x)=\frac{p(x\mid w=2)p(w=2)}{p(x)}=98.1\% \tag{2-8}$$

其中

$$p(x)=p(x\mid w=1)p(w=1)+p(x\mid w=2)p(w=2) \tag{2-9}$$

贝叶斯最小错误率决策会选择后验概率最大的类别，即 $h(x)=2$。

（2）若进行最小风险贝叶斯决策，需考虑不同决策的损失，假设决策损失矩阵为（只是假设，合理的数值应该视真实情况而定）

$$\Lambda=\begin{bmatrix}0 & 100\\ 1 & 0\end{bmatrix} \tag{2-10}$$

其中 λ_{ij} 表示将第 i 类数据判别为第 j 类的损失，也可以用 $\lambda(h(x)=j\mid w=i)$ 表示。在该例子中，λ_{12} 表示将患病判别为正常的损失，λ_{21} 表示将正常判别为患病的损失。最小风险决策综合考虑各种决策的损失，选择条件风险最小的类别，其中条件风险的计算如下：

$$R(h(x)\mid x)=\sum_i \lambda(h(x)\mid w=i)p(w=i\mid x) \tag{2-11}$$

可以得到将 x 判别为不同类别的条件风险为

$$\begin{aligned}&R(h(x)=1\mid x)\\&=\lambda(h(x)=1\mid w=1)p(w=1\mid x)+\lambda(h(x)=1\mid w=2)p(w=2\mid x)\\&=98.1\%\end{aligned} \tag{2-12}$$

$$\begin{aligned}&R(h(x)=2\mid x)\\&=\lambda(h(x)=2\mid w=1)p(w=1\mid x)+\lambda(h(x)=2\mid w=2)p(w=2\mid x)\\&=190\%\end{aligned} \tag{2-13}$$

最小风险贝叶斯决策会选择条件风险最小的类别，即 $h(x)=1$。

目录

二类分类问题： 要机器来判断一张图像是大熊猫还是小熊猫
多类分类问题： 区分一张图片是大熊猫、小熊猫还是棕熊

（a）大熊猫

（b）小熊猫

（c）棕熊

分类器是一个计算系统，它通过计算出一系列判别函数的值做出分类决策，实现对输入数据进行分类的目的。
判别函数是一个从输入特征映射到决策的函数，其结果可以直接用于做出分类决策。
分类问题中，分类器会把输入空间划分成多个决策区域，这些决策区域之间的边界称作**决策面**或**决策边界**。

例：对于一个C类图像识别任务，分类器将提取的特征$\mathbf{x}$作为输入向量（例如使用图像的SIFT特征表示一张图像），然后输出一个对应的类标签。

首先，分类器计算出C个判别函数

$$\mathrm{g}_i(\mathbf{x}), i = 1,2,\cdots,C.$$

其次，分类器会把一个特征向量$\mathbf{x}$划分为类别c，如果满足：

$$\mathrm{g}_c(\mathbf{x}) > \mathrm{g}_{c'}(\mathbf{x}), \forall c' \neq c.$$

在输入空间中，使得$\mathrm{g}_c(\mathbf{x}) = \mathrm{g}_{c'}(\mathbf{x}), \forall c' \neq c$成立的超平面就是决策面。

分类器的构建方法有很多种，常用的方法大致可以分为三大类，这里按照复杂度依次降低的顺序罗列。其中生成式模型和判别式模型都是基于概率框架，生成式模型构建所有观测的联合分布，而判别式模型只关心给定输入数据时输出数据的条件分布。

在得到一个训练好的分类器后，我们需要去评估这个分类器的性能好坏。当分类器确定后，其错误率亦随之确定了。分类器的错误率可以用于比较对于同一问题设计的多种分类器的优劣。分类器的错误率计算通常有三种方法：

1. 根据错误率的定义按照公式进行计算。
2. 计算错误率的上界。
3. 通过在测试数据上进行分类实验来估计错误率。

$$\text{Error} = \frac{1}{M}\sum_{i=1}^{M} I[h(x_i) \neq y_i]$$

其中$I[\cdot]$表示单位函数，当且仅当括号中的条件满足时取值为1，否则取值为0。把$1 - \text{Error}$称作精度Accuracy

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M} I[h(x_i) = y_i]$$

目录

- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计

- 高斯密度函数/正态密度函数

一元高斯密度函数：

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

其中μ是均值，σ^2是方差，分别表示为

$$\mu = \mathbb{E}[x] = \int xp(x)dx$$

$$\sigma^2 = \mathbb{E}[(x-\mu)^2] = \int (x-\mu)^2 p(x)dx$$

多元高斯密度函数：

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

其中$\boldsymbol{\mu}$是均值，Σ是协方差。

- 基于高斯分布的贝叶斯决策

假设类条件概率分布为高斯分布：

$$p(\mathbf{x}|w=i)=\mathcal{N}(\boldsymbol{\mu},\Sigma_i),i=1,2,\cdots,C$$

贝叶斯决策得到的判别函数为

$$\begin{aligned}\mathrm{g_i}(\mathbf{x})&=\ln p(\mathbf{x}|w=i)+\ln p(w=i)\\&=-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln p(w=i)\\&-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_i})^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu_i})\end{aligned}$$

通过判别函数可以得到决策面$\mathrm{g_i}(x)=\mathrm{g_j}(x)$为

$$-\frac{1}{2}\left[(x-\boldsymbol{\mu_i})^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu_i})-(\mathbf{x}-\boldsymbol{\mu_j})^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu_j})\right]+\ln\frac{p(w=i)}{p(w=j)}-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}=0$$

考虑当所有类别的协方差矩阵都相等的情况下，即

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_C = \Sigma$$

则判别函数可化简为

$$\mathrm{g_i}(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_i})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu_i}) + \ln p(w=i)$$

忽略与i无关的项$\mathbf{x}^{\mathbf{T}}\Sigma^{-1}\mathbf{x}$，判别函数进一步简化为

$$\mathrm{g_i}(\mathbf{x}) = (\Sigma^{-1}\boldsymbol{\mu}_i)^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\mathbf{u}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu_i} + \ln p(w=i)$$

此时判别函数是$\mathbf{x}$线性函数，决策面是一个超平面。

当决策区域R_i与R_j相邻时，决策面满足

$$\mathrm{g_i}(\mathbf{x}) - \mathrm{g_j}(\mathbf{x}) = \mathbf{0}$$

即

$$\left[\Sigma^{-1}(\boldsymbol{\mu_i} - \boldsymbol{\mu_j})\right]^{\mathrm{T}}(\mathbf{x} - \mathbf{x_0}) = 0$$

当各类别的先验概率相等时，可以得到

$$\mathbf{x_0} = \frac{1}{2}(\boldsymbol{\mu_i} + \boldsymbol{\mu_j})$$

即为$\boldsymbol{\mu_i}$与$\boldsymbol{\mu_j}$连线的中点。

图2-3展示了此时基于非对角协方差矩阵的二维高斯分布的贝叶斯决策面。

当各类别的先验概率不相等时，则$\mathbf{x_0}$不在$\boldsymbol{\mu_i}$与$\boldsymbol{\mu_j}$连线的中点上，并且向先验概率小的方向偏移。

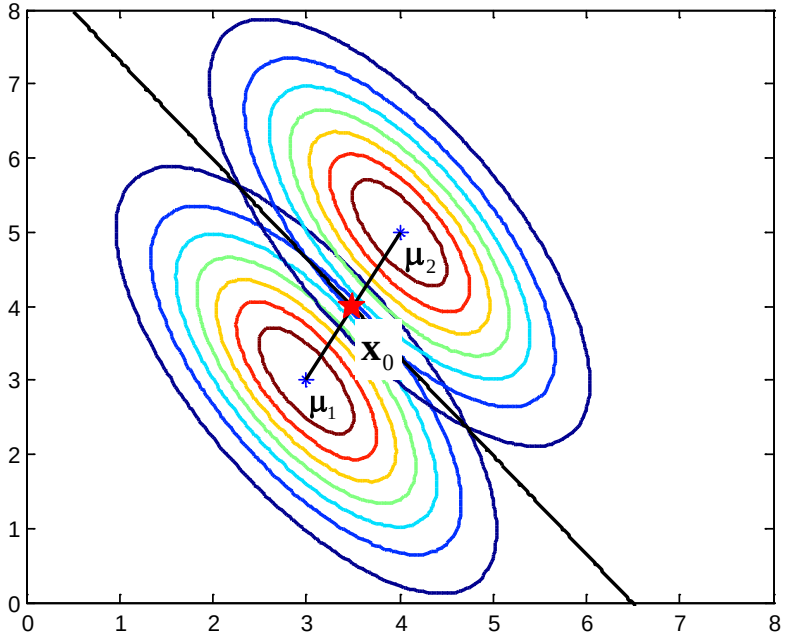


图2-3

（图中展示了当类别先验概率相等，两个类条件概率分布均为高斯分布且具有相等的非对角协方差矩阵时的贝叶斯决策的决策面，图中椭圆形的环表示类条件概率密度等高线。）

目录

- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计

- 朴素贝叶斯分类器

朴素贝叶斯（naïve Bayes）分类器对条件概率分布提出了**特征条件独立的**假设。

例：对于一张大熊猫图像，它的词袋特征可以表示为一个D维的向量，朴素贝叶斯假设向量的D个元素之间相互独立，其联合分布可以写成 个独立的概率分布相乘。

基于此假设，类别w的后验概率为：

$$p(w|\mathbf{x}) = \frac{p(w)p(\mathbf{x}|w)}{p(\mathbf{x})} \propto p(w)\prod_{d=1}^{D} p(\mathbf{x}_d|w)$$

其中D为特征的个数，x_d为第d个特征上的值。因此，基于朴素贝叶斯分类器的分类结果为

$$\text{argmax}_{\text{w}}\, p(w)\prod_{d=1}^{D} p(x_d|w)$$

目录

- 贝叶斯公式
- 贝叶斯决策
- 分类器的相关概念
- 基于高斯分布的贝叶斯分类器
- 朴素贝叶斯分类器
- 参数估计

- 最大似然估计（maximum likelihood estimation）

最大似然估计是一种给定观测时估计模型参数的方法，它试图在给定观测的条件下，找到最大化似然函数的参数值。

例：假设数据的分布是联合高斯分布的，那么似然函数就是所有观测数据以均值与协方差为参数的联合高斯密度函数，此时$p(\mathcal{D}|\theta) = \mathcal{N}(\mathcal{D}|\boldsymbol{\mu}, \Sigma)$。最大似然方法找到使得似然函数$p(\mathcal{D}|\theta)$最大的模型参数的值$\hat{\theta}_{ml}$，即

$$\hat{\theta}_{ml} = \text{argmax}_{\theta}\, p(\mathcal{D}|\theta)$$

为了计算方便，通常使用似然函数的自然对数作为优化目标，称作对数似然（log-likelihood），那么

$$\hat{\theta}_{ml} = \text{argmax}_{\theta}\, \ln p(\mathcal{D}|\theta)$$

如果数据是独立同分布的且样本个数为N，那么所有训练数据的对数似然函数表示为

$$\ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(\mathbf{x_i}|\theta)$$

考虑基于高斯分布的贝叶斯分类器，给出高斯分布的最大似然估计。假设某类别具有N个样本，则类条件密度/似然密度函数的对数为

$$\sum_{i=1}^{N} \ln p(\mathbf{x_i}|\theta) = \sum_{i=1}^{N} \ln \mathcal{N}(\mathbf{x_i}|\boldsymbol{\mu}, \Sigma)$$

关于均值和协方差进行求导，对上式进行最大化，以得到均值与协方差的估计值：

$$\boldsymbol{\mu_{ml}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x_i}, \qquad \Sigma_{ml} = \frac{1}{N}(\mathbf{x_i} - \boldsymbol{\mu_{ml}})(\mathbf{x_i} - \boldsymbol{\mu_{ml}})^{\mathrm{T}}$$

- 思考与计算

推导高斯分布的均值与协方差的最大似然估计。$\arg\max_{\boldsymbol{\mu},\Sigma} \sum_{i=1}^{N} \ln N(\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma)$

$$\frac{d\sum_{i=1}^{N} \ln N(\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma)}{d\boldsymbol{\mu}} = \frac{d\sum_{i=1}^{N} \ln \frac{1}{2\pi \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\}}{d\boldsymbol{\mu}}$$

$$= \frac{-\frac{1}{2}\sum_{i=1}^{N} d(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{d\boldsymbol{\mu}}$$

$$= \frac{-\frac{1}{2}\sum_{i=1}^{N} (d(\mathbf{x}_i - \boldsymbol{\mu})^{\top}) \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + (\mathbf{x}_i - \boldsymbol{\mu})^{\top} d(\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}))}{d\boldsymbol{\mu}}$$

$$= \frac{-\frac{1}{2}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}) \Sigma^{-1} (d(\mathbf{x}_i - \boldsymbol{\mu})) + (\mathbf{x}_i - \boldsymbol{\mu})^{\top} d(\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}))}{d\boldsymbol{\mu}}$$

$$= \frac{\frac{1}{2}\sum_{i=1}^{N} 2(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} d\boldsymbol{\mu}}{d\boldsymbol{\mu}} = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} = \mathbf{0}^{\top}$$

$$\boldsymbol{\mu}_{m\ell} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

$$\frac{d\sum_{i=1}^{N} \ln N(\mathbf{x}_i \mid \boldsymbol{\mu}, \Sigma)}{d\Sigma} = \frac{d\sum_{i=1}^{N} \ln \frac{1}{2\pi \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})\}}{d\Sigma}$$

$$= \frac{-\frac{N}{2} d \ln \mid \Sigma \mid -\frac{1}{2}\sum_{i=1}^{N} d((\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}))}{d\Sigma}$$

$$= \frac{-\frac{N}{2} \mathrm{Tr}[\Sigma^{-1} d\Sigma] + \frac{1}{2}\sum_{i=1}^{N} d\mathrm{Tr}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (d\Sigma) \Sigma^{-1}]}{d\Sigma}$$

$$= \frac{-\frac{N}{2} \mathrm{Tr}[\Sigma^{-1} d\Sigma] + \frac{1}{2}\sum_{i=1}^{N} d\mathrm{Tr}[\Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} (d\Sigma)]}{d\Sigma}$$

$$= -N\Sigma^{-1} + \sum_{i=1}^{N} \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\top} \Sigma^{-1} = 0$$

$$\Sigma_{m\ell} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_{m\ell})(\mathbf{x}_i - \boldsymbol{\mu}_{m\ell})^{\top}$$

- 最大后验估计

最大后验估计是在最大似然估计的基础上考虑参数的先验分布，通过贝叶斯公式获得参数的后验分布，并以后验分布作为估计的优化目标。参数θ的最大后验估计$\hat{\theta}_{\text{map}}$表示为

$$\hat{\theta}_{\text{map}} = \text{argmax}_\theta \ln p(\theta|\mathcal{D}) = \text{argmax}_\theta \ln \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$
$$= \text{argmax}_\theta \ln p(\mathcal{D}|\theta) + \ln p(\theta)$$

考虑基于高斯分布的贝叶斯分类器，假设协方差已知情况下给出对均值的最大后验估计。首先假设均值是服从高斯分布的，如$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mu)$，则其对数后验概率为

$$\ln p(\boldsymbol{\mu}|\mathcal{D}) = \sum_{i=1}^{N} \ln \mathcal{N}\left(\mathbf{x_i}|\boldsymbol{\mu}, \Sigma_\mu\right) + \ln \mathcal{N}\left(\boldsymbol{\mu}|\mathbf{0}, \Sigma_\mu\right) + \text{const}$$

- 思考与计算

推导高斯分布的均值的最大后验估计。

$$\ln p(\boldsymbol{\mu}\mid\mathcal{D}) = \sum_{i=1}^{N}\ln p(\mathbf{x}_i\mid\boldsymbol{\mu}) + \ln p(\boldsymbol{\mu}) + \text{const}$$
$$= \sum_{i=1}^{N}\ln\mathcal{N}(\mathbf{x}_i\mid\boldsymbol{\mu},\Sigma) + \ln\mathcal{N}(\boldsymbol{\mu}\mid 0,\Sigma_{\boldsymbol{\mu}}) + \text{const},$$

$$\frac{d\sum_{i=1}^{N}\ln\mathcal{N}(\mathbf{x}_i\mid\boldsymbol{\mu},\Sigma)+\ln\mathcal{N}(\boldsymbol{\mu}\mid 0,\Sigma_{\boldsymbol{\mu}})}{d\boldsymbol{\mu}}$$
$$= \frac{d\sum_{i=1}^{N}\ln\frac{1}{2\pi\mid\Sigma\mid^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\}+d\ln\frac{1}{2\pi\mid\Sigma_{\boldsymbol{\mu}}\mid^{1/2}}\exp\{-\frac{1}{2}\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}\}}{d\boldsymbol{\mu}}$$
$$= \frac{-\frac{1}{2}\sum_{i=1}^{N}d(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})-\frac{1}{2}d\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}}{d\boldsymbol{\mu}}$$
$$= \frac{-\frac{1}{2}\sum_{i=1}^{N}(d(\mathbf{x}_i-\boldsymbol{\mu})^{\top})\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))-\frac{1}{2}(d\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^{\top}d\Sigma_{\boldsymbol{\mu}}^{-1}\boldsymbol{\mu})}{d\boldsymbol{\mu}}$$
$$= \frac{-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(d(\mathbf{x}_i-\boldsymbol{\mu}))+(\mathbf{x}_i-\boldsymbol{\mu})^{\top}d(\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}))-\frac{1}{2}(\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}d\boldsymbol{\mu}+\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}d\boldsymbol{\mu})}{d\boldsymbol{\mu}}$$
$$= \frac{\frac{1}{2}(\sum_{i=1}^{N}2(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}-2\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1})d\boldsymbol{\mu}}{d\boldsymbol{\mu}} = \sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}+\boldsymbol{\mu}^{\top}\Sigma_{\boldsymbol{\mu}}^{-1}=\mathbf{0}^{\top}$$

$$\boldsymbol{\mu}_{\text{map}} = N(N\Sigma^{-1}+\Sigma_{\boldsymbol{\mu}}^{-1})^{-1}\Sigma^{-1}\boldsymbol{\mu}_{\text{m}\ell}$$
$$= N\Sigma_{\boldsymbol{\mu}}(N\Sigma_{\boldsymbol{\mu}}+\Sigma)^{-1}\boldsymbol{\mu}_{\text{m}\ell}.$$

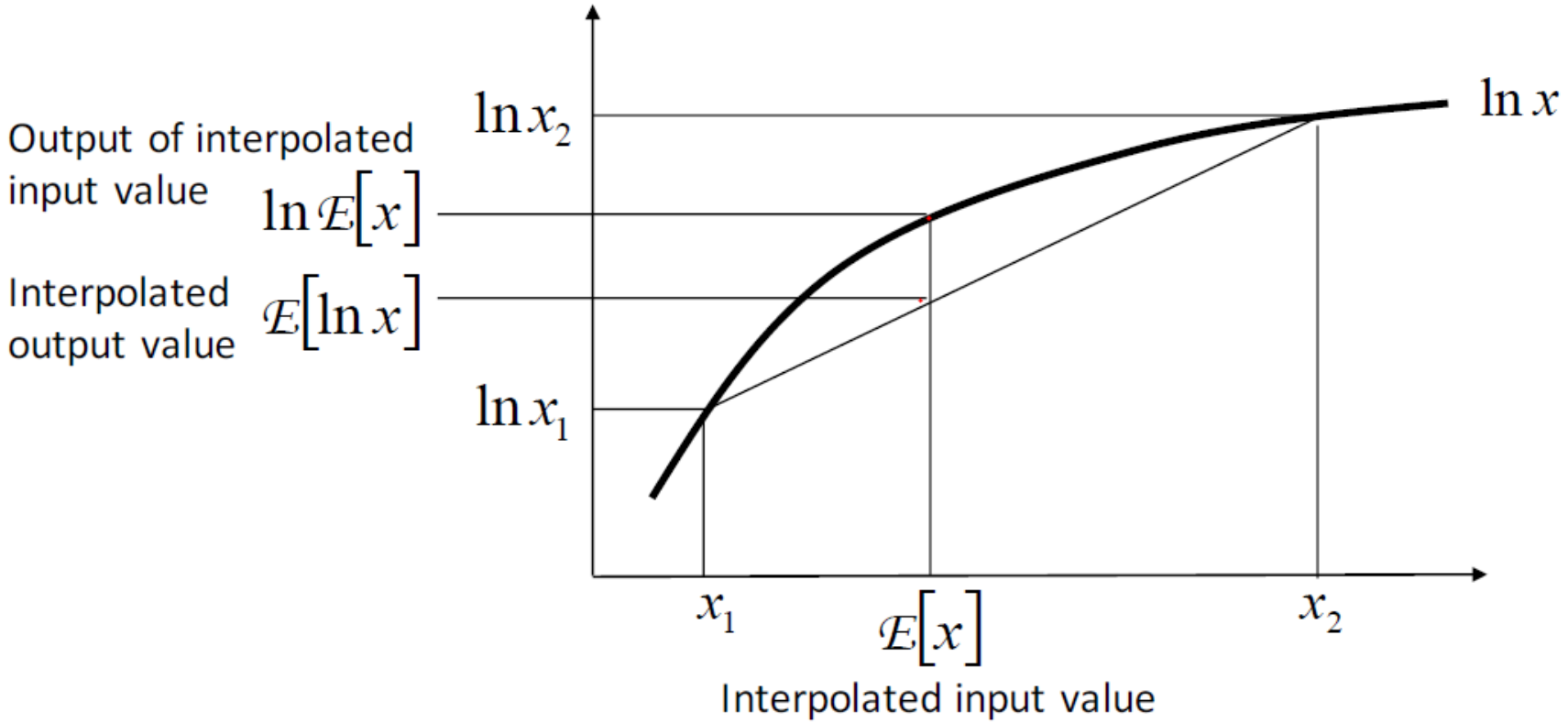
- 期望最大化算法（expectation maximization，EM）

例：对不完整数据建模时，使用隐变量定义缺失数据；
对复杂的观测数据建模时，使用隐变量定义潜在因素。
考虑一个概率模型，X表示观测变量集，Z表示隐变量集，θ表示模型参数，目标是最大化观测变量X对参数θ的对数似然函数：

$$L(\theta) = \ln p(X|\theta) = \ln \int p(X, Z|\theta) dZ$$

=>交换积分运算与对数运算的顺序

$$\ln \sum_n p_n x_n \geq \sum_n p_n \ln x_n \quad \text{subject to } p_n \geq 0, \sum_n p_n = 1$$



Lower bound:

$$\mathcal{L}(q,\lambda)=\sum_{n=1}^{N}\sum_{z_n} q(z_n)\ln\frac{p(x_n,z_n\mid\lambda)}{q(z_n)}$$

$$=\frac{p(x_n\mid\lambda)p(z_n\mid x_n,\lambda)}{q(z_n)}$$

$$=\sum_{n=1}^{N}\left\{\sum_{z_n} q(z_n)\ln p(x_n\mid\lambda)+\sum_{z_n} q(z_n)\ln\frac{p(z_n\mid x_n,\lambda)}{q(z_n)}\right\}$$

$$=\sum_{n=1}^{N}\left\{\ln p(x_n\mid\lambda)-\sum_{z_n} q(z_n)\ln\frac{q(z_n)}{p(z_n\mid x_n,\lambda)}\right\}$$

$$=\underbrace{\ln p(X\mid\lambda)}_{\text{Log-scaled likelihood}}-\underbrace{\sum_{n=1}^{N}\mathrm{KL}(q(z_n)\|\,p(z_n\mid x_n,\lambda))}_{\text{KL divergence}}$$

“Lower bound” = “log-scaled likelihood” – “KL divergence”

$$\ln\sum_n p_n x_n \geq \sum_n p_n \ln x_n \quad \text{subject to } p_n\geq 0,\ \sum_n p_n = 1$$

- KL散度:

Discrete variable case:

$$\mathrm{KL}(p\|q) = \sum_z p(z)\ln\frac{p(z)}{q(z)}$$

Continuous variable case:

$$\mathrm{KL}(p\|q) = \int p(x)\ln\frac{p(x)}{q(x)}\mathrm{d}x$$

$$\mathrm{KL}(p\|q) = \int p(x)\left(\ln p(x) - \ln q(x)\right)\mathrm{d}x$$

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x} \geqslant -\ln\int q(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0$$

$$\mathrm{KL}(p\|q) \geq 0 \qquad \text{If } q = p,\ \mathrm{KL}(p\|q) = 0$$

- 进一步分析对数似然的下界:

Lower bound:

$$\mathcal{L}(q,\boldsymbol{\lambda}) = \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln\frac{p(\boldsymbol{x}_n,\boldsymbol{z}_n \mid \boldsymbol{\lambda})}{q(\boldsymbol{z}_n)}$$

$$= \underbrace{\sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln p(\boldsymbol{x}_n,\boldsymbol{z}_n \mid \boldsymbol{\lambda})}_{\text{Auxiliary function}} \underbrace{-\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln q(\boldsymbol{z}_n)}_{\text{Constant}}$$

"Lower bound" ∝ "Auxiliary function"

- 最大化对数似然的下界（分别关于额外引入的分布q和模型参数λ）：

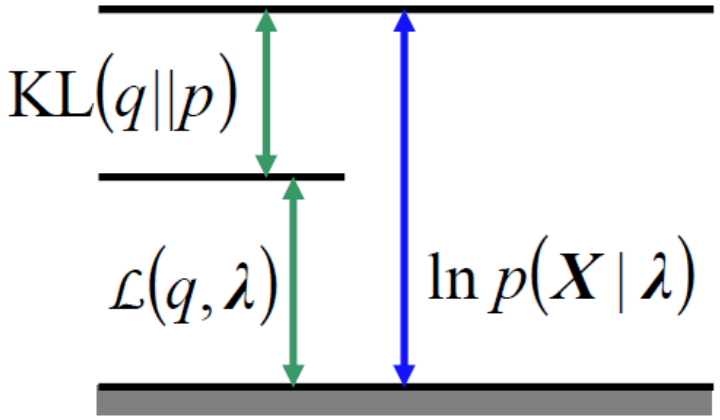
Lower bound: functional of q and function of $\boldsymbol{\lambda}$

E-step: Maximize lower bound with respect to q while fixing $\boldsymbol{\lambda}$:

$$\mathcal{L}(q,\boldsymbol{\lambda}) = \ln p(\boldsymbol{X}\,|\,\boldsymbol{\lambda}) - \sum_{n=1}^{N} \mathrm{KL}(q(\boldsymbol{z}_n)\,\|\,p(\boldsymbol{z}_n\,|\,\boldsymbol{x}_n,\boldsymbol{\lambda}))$$

M-step: Maximize lower bound with respect to $\boldsymbol{\lambda}$ while fixing q

$$\mathcal{L}(q,\boldsymbol{\lambda}) \propto \sum_{n=1}^{N}\sum_{\boldsymbol{z}_n} q(\boldsymbol{z}_n)\ln p(\boldsymbol{x}_n,\boldsymbol{z}_n\,|\,\boldsymbol{\lambda})$$



* Need to set initial model parameters $\boldsymbol{\lambda}$

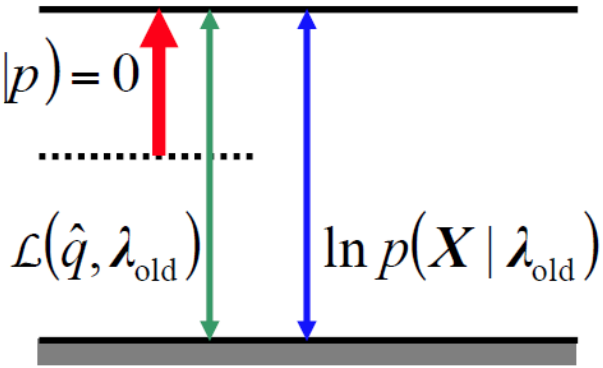
- E步：

- Set KL divergence to 0 under the fixed model parameters $\boldsymbol{\lambda}_{\text{old}}$

$$\sum_{n=1}^{N} \mathrm{KL}\left(q(\boldsymbol{z}_n) \| p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}})\right) = 0 \quad \Longrightarrow \quad \hat{q}(\boldsymbol{z}_n) = p(\boldsymbol{z}_n \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}})$$

Calculate posterior probabilities of latent variables for each sample

$$p(z_{n,m} = 1 \mid \boldsymbol{x}_n, \boldsymbol{\lambda}_{\text{old}}) = \frac{p(z_{n,m} = 1 \mid \boldsymbol{\lambda}_{\text{old}}) p(\boldsymbol{x}_n \mid z_{n,m} = 1, \boldsymbol{\lambda}_{\text{old}})}{\sum_{m=1}^{M} p(z_{n,m} = 1 \mid \boldsymbol{\lambda}_{\text{old}}) p(\boldsymbol{x}_n \mid z_{n,m} = 1, \boldsymbol{\lambda}_{\text{old}})}$$



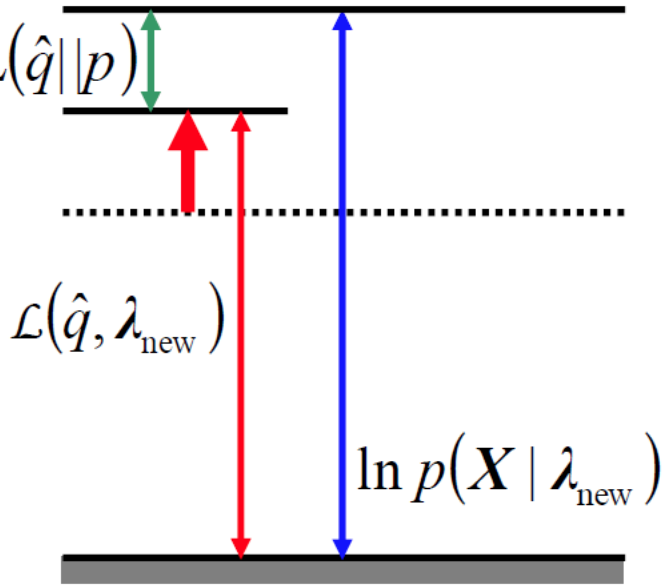
- M步：

- Maximize auxiliary function with respect to model parameters λ_{new}

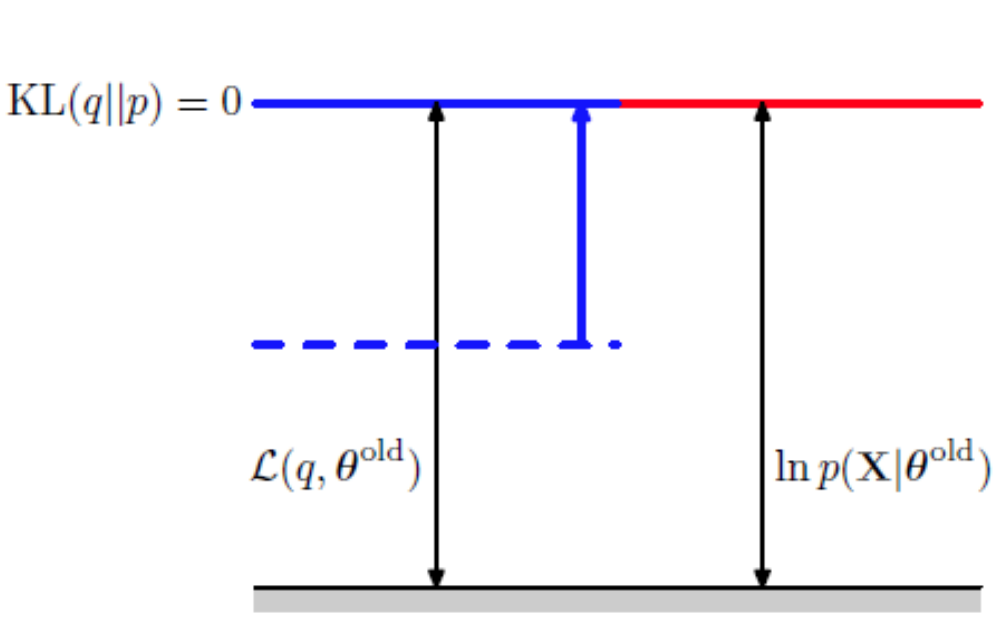
$$\mathcal{L}(\hat{q},\lambda) \propto \sum_{n=1}^{N}\sum_{z_n}\hat{q}(z_n)\ln p(x_n,z_n\mid\lambda)$$

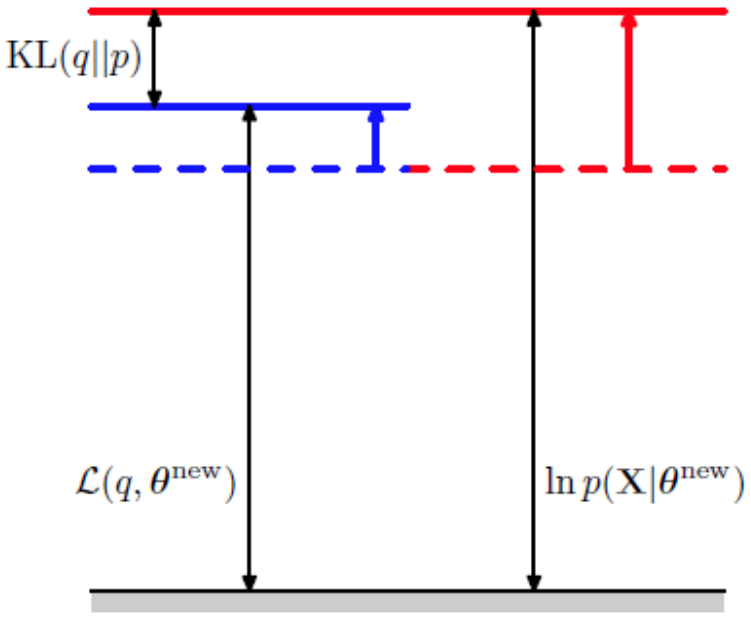
Auxiliary function

$$Q(\lambda_{\text{new}},\lambda_{\text{old}}) = \sum_{n=1}^{N}\sum_{z_n} p(z_n\mid x_n,\lambda_{\text{old}})\ln p(x_n,z_n\mid\lambda_{\text{new}})$$



- 期望最大化算法（expectation maximization，EM）





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

- 期望最大化算法（expectation maximization，EM）

例：对不完整数据建模时，使用隐变量定义缺失数据；
对复杂的观测数据建模时，使用隐变量定义潜在因素。
考虑一个概率模型，X表示观测变量集，Z表示隐变量集，θ表示模型参数，目标是最大化观测变量X对参数θ的对数似然函数：

$$L(\theta) = \ln p(X|\theta) = \ln \int (X, Z|\theta) dZ$$

EM算法是一种迭代算法，常用于求解带有隐变量的概率模型的最大似然或者最大后验估计。
E步：根据给定观测变量X和当前参数θ推理隐变量Z的后验概率分布，并计算观测数据X和隐变量Z的对数联合概率关于Z的后验概率分布的期望；
M步：最大化E步求得的期望，获得新的参数θ。

算法 2-1 期望最大化算法（expectation maximization，EM）

输入：观测数据 X

1: 初始化参数 $\theta^{(1)}$，

2: REPEAT

3: E 步：记 $\theta^{(t)}$ 为第 t 次迭代参数的估计值，计算对数联合概率分布 $\ln p(X,Z|\theta)$ 关于隐变量 Z 的后验概率分布 $p(Z|X,\theta^{(t)})$ 的期望：

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}} \ln p(X,Z|\theta) = \int p(Z|X,\theta^{(t)}) \ln p(X,Z|\theta) dZ.$$

4: M 步：求解使 $Q(\theta|\theta^{(t)})$ 最大化的 θ，得到第 $t+1$ 次迭代的参数估计：

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}).$$

5: UNTIL 满足收敛条件：

$$\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon_1 \text{ 或 } \|Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t-1)})\| < \varepsilon_2,$$

其中 $\varepsilon_1, \varepsilon_2$ 是非常小的正数

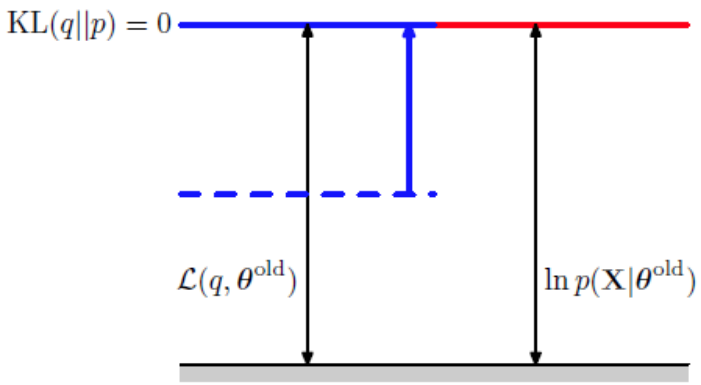
输出：参数 $\theta^{(t+1)}$

- EM算法总结

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

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$



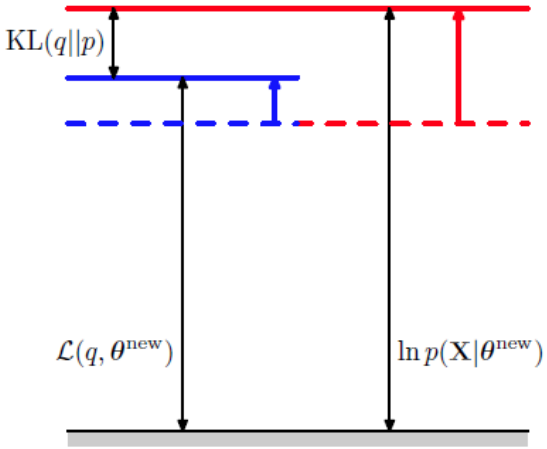
E step Evaluate $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\mathrm{old}})$.

M step Evaluate $\boldsymbol{\theta}^{\mathrm{new}}$ given by

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

where

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



- 思考与计算

1、最大后验也可以用EM算法吗？
2、如果后验分布无法计算呢？

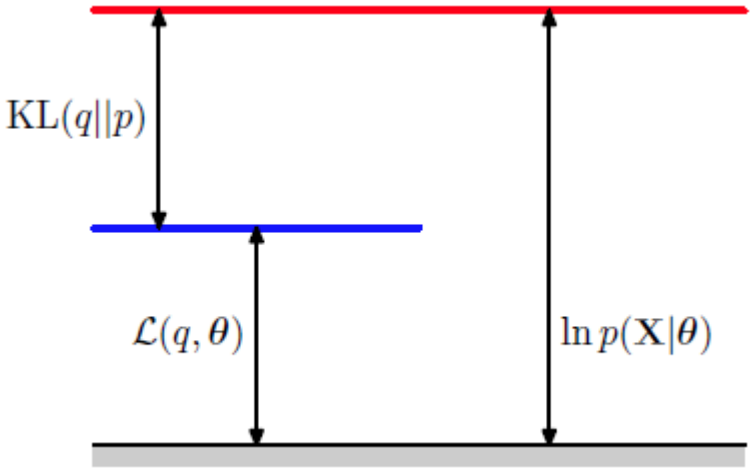
- 变分推理

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

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$\mathrm{KL}(p\,\|\,q) \geq 0$ If $q = p$, $\mathrm{KL}(p\,\|\,q) = 0$



- 平均场近似

$$\mathcal{L}(q) = \int (\textstyle\prod_i q_i)\{\ln p(X,Z) - \sum_i \ln q_i\} dZ$$

$$= \int q_j \{ \int \ln p(X,Z) \prod_{i \neq j} q_i dZ_i \} dZ_j - \int q_j \ln q_j dZ_j + \text{const}$$

$$= \int q_j \ln \tilde{p}(X,Z_j) dZ_j - \int q_j \ln q_j dZ_j + \text{const},$$

$$\ln \tilde{p}(X,Z_j) = \int \ln p(X,Z) \prod_{i \neq j} q_i dZ_i + \text{const}.$$

$$\mathbb{E}_{i \neq j}[\ln p(X,Z)] = \int \ln p(X,Z) \prod_{i \neq j} q_i dZ_i.$$

最大化下界等价于最小化一个相应的KL散。变分下界在

$$q_j(Z_j) = \tilde{p}(X,Z_j)$$

时取得最大值。于是得到最优解：

$$\ln q_j^*(Z_j) = \mathbb{E}_{i \neq j}[\ln p(X,Z)] + \text{const}.$$

- 贝叶斯参数估计

贝叶斯参数估计不直接估计参数的值，而是通过贝叶斯公式推理出参数的后验分布。因此贝叶斯参数估计得到的是参数θ在给定观测数据集$\mathcal{D}$的后验分布

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

贝叶斯参数估计从训练数据X学习出参数的后验分布$p(\theta_c|X, w = c)$。在训练完成后，利用该后验分布可以得到测试样本 的类条件概率分布为

$$p(\mathbf{x}_*|w = c) = \int p(\mathbf{x}_*|w = c, \theta_c)p(\theta_c|X, w = c)d\theta_c$$

考虑基于高斯分布的贝叶斯分类器，假设协方差已知，且$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mu)$，则均值参数的后验分布为

$$p(\boldsymbol{\mu}|x) = \mathcal{N}(N\Sigma_\mu\left(\mathrm{N}\Sigma_\mu + \Sigma\right)^{-1}\boldsymbol{\mu}_{ml}, \left(\Sigma_{\boldsymbol{\mu}}^{-1} + N\Sigma^{-1}\right)^{-1})$$

- 思考与计算

推导高斯分布的均值的贝叶斯估计。

$$
\begin{aligned}
p(\boldsymbol{\mu}\,|\,\mathcal{D}) &= p(\boldsymbol{\mu})\prod_{i=1}^{N} p(\mathbf{x}_i\,|\,\boldsymbol{\mu}) \,/\, p(\mathcal{D}) \\
&= \mathcal{N}(\boldsymbol{\mu}\,|\,0,\Sigma_{\boldsymbol{\mu}})\prod_{i=1}^{N} \mathcal{N}(\mathbf{x}_i\,|\,\boldsymbol{\mu},\Sigma) \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi\,|\,\Sigma_{\boldsymbol{\mu}}\,|^{1/2}}\exp\{-\frac{1}{2}\boldsymbol{\mu}^{\top}{\Sigma_{\boldsymbol{\mu}}}^{-1}\boldsymbol{\mu}\}\prod_{i=1}^{N}\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\exp\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\} \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi\,|\,\Sigma_{\boldsymbol{\mu}}\,|^{1/2}}\left(\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\right)^{N}\exp\{-\frac{1}{2}\boldsymbol{\mu}^{\top}{\Sigma_{\boldsymbol{\mu}}}^{-1}\boldsymbol{\mu}-\sum_{i=1}^{N}\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi\,|\,\Sigma_{\boldsymbol{\mu}}\,|^{1/2}}\left(\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\right)^{N}\exp\left\{-\frac{1}{2}\boldsymbol{\mu}^{\top}{\Sigma_{\boldsymbol{\mu}}}^{-1}\boldsymbol{\mu}-\frac{1}{2}\sum_{i=1}^{N}\mathbf{x}_i^{\top}\Sigma^{-1}\mathbf{x}_i-2\boldsymbol{\mu}^{\top}\Sigma^{-1}\mathbf{x}_i+\boldsymbol{\mu}^{\top}\Sigma^{-1}\boldsymbol{\mu}\right\} \,/\, p(\mathcal{D}) \\
&= \frac{1}{2\pi\,|\,\Sigma_{\boldsymbol{\mu}}\,|^{1/2}}\left(\frac{1}{2\pi\,|\,\Sigma\,|^{1/2}}\right)^{N}\exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}^{\top}({\Sigma_{\boldsymbol{\mu}}}^{-1}+N\Sigma^{-1})\boldsymbol{\mu}-2N\boldsymbol{\mu}^{\top}\Sigma^{-1}\boldsymbol{\mu}_{\mathrm{m}\ell}+\sum_{i=1}^{N}\mathbf{x}_i^{\top}\Sigma^{-1}\mathbf{x}_i\right]\right\} \,/\, p(\mathcal{D}) \\
&= \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}^{\top}({\Sigma_{\boldsymbol{\mu}}}^{-1}+N\Sigma^{-1})\boldsymbol{\mu}-2N\boldsymbol{\mu}^{\top}\Sigma^{-1}\boldsymbol{\mu}_{\mathrm{m}\ell}+\sum_{i=1}^{N}\mathbf{x}_i^{\top}\Sigma^{-1}\mathbf{x}_i\right]\right\}\times\mathrm{const} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}-N(\Sigma_{\boldsymbol{\mu}}^{-1}+N\Sigma^{-1})\Sigma^{-1}\boldsymbol{\mu}_{\mathrm{m}\ell}\right)^{\top}(\Sigma_{\boldsymbol{\mu}}^{-1}+N\Sigma^{-1})^{-1}\left(\boldsymbol{\mu}-N(\Sigma_{\boldsymbol{\mu}}^{-1}+N\Sigma^{-1})\Sigma^{-1}\boldsymbol{\mu}_{\mathrm{m}\ell}\right)\right\}\times\mathrm{const} \\
&= \mathcal{N}\left(N\Sigma_{\boldsymbol{\mu}}(N\Sigma_{\boldsymbol{\mu}}+\Sigma)^{-1}\boldsymbol{\mu}_{\mathrm{m}\ell},(\Sigma_{\boldsymbol{\mu}}^{-1}+N\Sigma^{-1})^{-1}\right).
\end{aligned}
$$

1. Deng J, Dong W, Socher R, et al. Imagenet: A Large-Scale Hierarchical Image Database[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.
2. Lowe D G. Object Recognition from Local Scale-Invariant Features[C]//Proceedings of the 17th IEEE International Conference on Computer Vision. New York: IEEE, 1999: 1150-1157.
3. Sun S. Multi-view Laplacian Support Vector Machines[C]//Advanced Data Mining and Applications. Berlin: Springer, 2011: 209-222.
4. Sun S, Shawe-Taylor J, Mao L. PAC-Bayes Analysis of Multi-view Learning[J]. Information Fusion, 2017, 35(5): 117-131.
5. 张学工. 模式识别[M]. 第三版. 北京：清华大学出版社, 2009.
6. Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1), 1-22.