# PRML 读书笔记一

李华阳

Oct 2017

## 目录



**摘要**

对 PRML 第一章 Introduction 部分进行总结。这份笔记以记录我认为惊艳之处为主，有些知识点可能就舍弃了（e.g. 第二节中的高斯分布）。另外本章关于信息论的部分的重点之前在博客上总结过，这里不再赘述。

## 1 以多项式曲线拟合为例

机器学习中的问题主要分为以下几类:

1. **监督学习（supervised learning）**
   (a) 数据形如 $\{(x_1, t_1), \ldots, (x_N, t_N)\}$，其中 $x_i$ 和 $t_i$ 分别表示第 $i$ 个数据的特征向量和目标向量。
   (b) 如果 $t$ 表示离散的类别标签，则该问题为**分类（classification）**问题,

(c) 如果 $t$ 表示连续的变量，则该问题为**回归（regression）**问题。

2. **非监督学习（unsupervised learning）**

   (a) 只有特征向量 $\{x_1, \ldots, x_N\}$，没有目标向量。

   (b) 其中，探究相似数据聚合情况的叫**聚类（clustering）**问题，

   (c) 确定输入数据分布的叫**密度估计（density estimation）**问题，

   (d) 从高维空间向低维空间映射的**可视化（visualization）**问题。

3. **强化学习（reinforcement learning）**

   (a) 其关注的问题是：在给定条件（situation）下寻找合适的行动（suitable actions），以期最大的回报（maximize a reward）

以一个回归问题为例：训练集中包含了一些样本 $\boldsymbol{x} \equiv (x_1, \ldots, x_N)^T$，以及在 $\sin(2\pi x)$ 的基础上添加基于高斯分布（Gaussian distribution）的随机噪声构造的对应目标集合 $\boldsymbol{t} \equiv (t_1, \ldots, t_N)^T$，我们的目的是通过学习训练集合上的数据，使得观测到一个新的数据 $\hat{x}$ 时，能够尽量准确的预测出对应目标 $\hat{t}$。

解决该问题的一种比较简单的思路是利用多项式进行曲线拟合（curve fitting）。

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1}$$

其中 $\boldsymbol{w}$ 表示多项式的系数，$y(x, \boldsymbol{w})$ 是关于 $x$ 的非线性函数，但是是关于 $\boldsymbol{w}$ 的线性函数。此外，还应该有一种衡量拟合效果的**误差函数（error function）** $\boldsymbol{w}$ 为调整 $\boldsymbol{w}$ 进而优化预测效果提供指导：

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{2}$$

这里的误差函数采用的是**最小二乘误差（least squares error）**，公式中的 $\frac{1}{2}$ 是为了之后的求导方便，关于误差函数的选择会在之后讨论。

曲线拟合问题可以通过最小化 $E(\boldsymbol{w})$ 进行求解，由于该公式是关于 $\boldsymbol{w}$ 的二次函数，因此一定可以找到一个使得 $E(\boldsymbol{w})$ 最小的解析解 $\boldsymbol{w}^*$，而 $y(x, \boldsymbol{w}^*)$ 就是我们想要的拟合曲线。如图 1 所示，$M$ 代表了多项式的最高阶数，可以看出 $M = 3$ 时曲线的拟合效果最好，$M = 0$ 或者 $M = 1$ 时拟合效果较差，

我们称这种现象为**欠拟合（under-fitting）**，而 $M = 9$ 时虽然 $E(\boldsymbol{w}^*) = 0$，但是该曲线对 $\sin(2\pi x)$ 的表达非常差泛化能力不足，这种现象称为**过拟合（over-fitting）**。

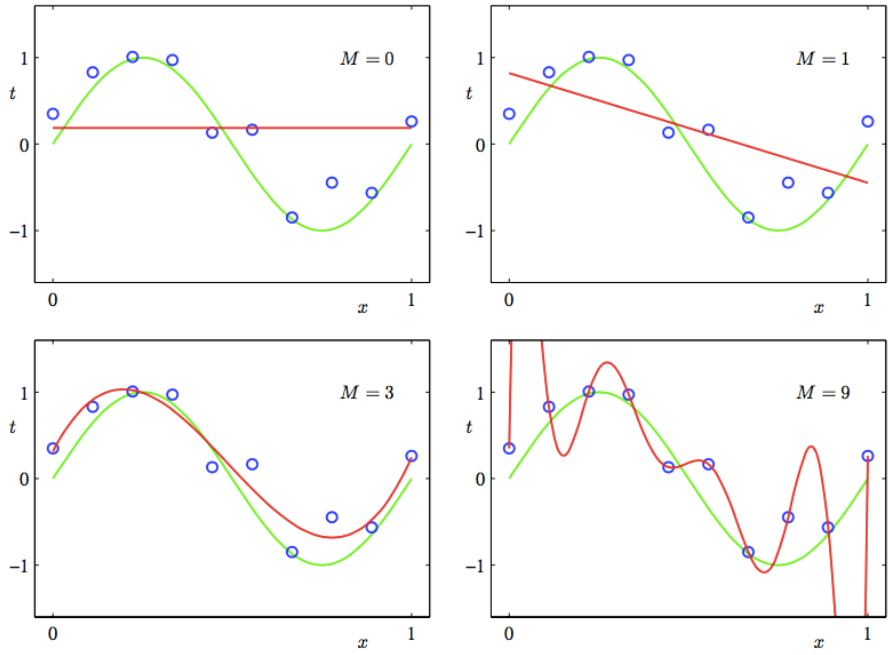


图 1: 不同阶数下的多项式曲线拟合效果

之后书中提出了两个非常有趣的问题：

1. $M = 3$ 的多项式只是 $M = 9$ 多项式的一个特例，按理说高阶多项式是可以包含阶较小的多项式的，但是为什么 $M = 9$ 时学习出的曲线效果这么差？

2. 对 $\sin(2\pi x)$ 进行多项式展开理论上是包含所有阶数的，但是为什么增大 $M$ 时学习效果反而变差了？

对于这两个问题，书中给出的解答是，直觉上看 $M$ 更大时更加灵活的多项式开始去拟合目标值 $t$ 中的随机噪声。但是这个解释并没有真的解决我的疑惑，希望在本书后半部分可以找到答案。虽然疑惑还在，但是一些方法可以用来改善过拟合的状况。

第一种解决过拟合问题的方法是增大训练集的大小。增大数据集的大小可以使我们承受得起更复杂的（或者说更灵活的）模型，一般来说数据集样本的数量要在模型参数个数的 5 到 10 倍（实际上参数个数并不能准确的代表模型的复杂度, 之后章节会讨论)。另一种方法是**正则化（regularization）**，将误差函数改写为如下形式:

$$\widetilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \tag{3}$$

其中 $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T\boldsymbol{w} = w_0^2 + w_1^2 + \cdots + w_M^2$，该正则化的方式也叫**岭回归（ridge regression）**。正则化主要是对拟合曲线的卷曲程度进行修正，由于 $w_0$ 作为截距项对拟合曲线的卷曲程度基本没有影响，所以一般不参与正则化。同时，对于 $\lambda$ 的选取也是一门学问，如果 $\lambda$ 过大会限制模型的灵活度，譬如 $\ln\lambda = 0$，从而造成欠拟合现象，如果 $\lambda$ 过小，如 $\ln\lambda = -\infty$，则相当于没有进行正则化。

## 2 概率论

此部分着重强调了从贝叶斯角度解释概率论的重要性, 已经从贝叶斯公式的层面上升到流派的层面。首先还是从频率学派的角度引出两条重要的概率统计规则:

**加法规则**
$$p(X) = \sum_{Y} p(X, Y) \tag{4}$$

**乘法规则**
$$p(X, Y) = P(Y|X)p(X) \tag{5}$$

由于**联合概率（joint probability）**中 $p(X, Y) = p(Y, X)$，所以对乘法规则稍作修改就得到了贝叶斯公式:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{6}$$

而对于贝叶斯公式的分母部分，应用加法规则可以得到:

$$p(X) = \sum_{Y} p(X|Y)p(Y) \tag{7}$$

其中 $p(Y)$ 叫做**先验概率（prior probability）**，$P(Y|X)$ 叫做**后验概率（posterior probability）**，后验概率是基于一些新的观测对先验概率进行

修正，使之更符合现实情况的概率分布。例如，样本足够大时，癌症患者在人群中的概率 $p(Y)$ 就是先验概率，在不知道更多信息的情况下，每个人得癌症的概率都是 $p(Y)$，但是有个人去医院拍摄了 X-光，基于这一事实可以对他得癌症的概率进行修正，修正后的概率就是 $p(Y|X)$。

离散型随机变量 $x$ 在给定值 $a$ 时 $p(x=a)$ 就是在值时的概率值。如果随机变量 $x$ 时连续的，则值在区间 $(a,b)$ 内的概率是对**概率密度（probability density）**函数 $p(x)$ 在该区间上的积分：

$$p(x \in (a,b)) = \int_a^b p(x)\mathrm{d}x \tag{8}$$

加法规则对于连续型随机变量同样适用，不过需要将求和符号换为积分符号。此外，**期望（expection）**和**方差（variance）**作为概率论的重点，他们的一些性质也在本章中进行论讨论。

首先对于期望，它是函数 $f(x)$ 在概率分布 $p(x)$ 下的均值：

**离散型随机变量**
$$\mathbb{E}[f] = \sum_x f(x)p(x) \tag{9}$$

**连续型随机变量**
$$\mathbb{E}[f] = \int f(x)p(x)\mathrm{d}x \tag{10}$$

如果数据集中有 $N$ 个样本，并且都依赖概率分布（密度）$p(x)$，那么有：

$$\mathbb{E}(f) \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{11}$$

并且当 $N \to \infty$ 时可以取等号。这一公式在之后会被广泛使用。

在讨论方差之前，我们可以先讨论**协方差（covariance）**的概念：

$$\begin{aligned} \mathrm{cov}[x,y] &= \mathbb{E}_{x,y}[x - \mathbb{E}[x]y - \mathbb{E}[y]] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{12}$$

它表达了 $x$ 和 $y$ 之间的相关性，如果两个随机变量是独立的，那么两者的协方差几乎为 0。当 $x$ 和 $y$ 是同一个随机变量是，协方差就变成了方差：

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{13}$$

频率学派对于概率的解释固然重要，但是当如果想要预测世纪末的时候南极的冰川是否会全部融化这种问题时，就没有办法像掷骰子一样进行随机重复试验了。而从**贝叶斯（Bayesian）**的角度，认为**概率是一种量化不确定**

**性的方法**。例如对于冰川融化的问题，可以利用概率来量化其不确定性，并且根据等不断更新的及时数据（e.g. 卫星观测站的观测、冰川融化速率等）进行修改，从而更准确的量化不确定性并作出相应的行动（e.g. 控制温室气体的排放速度等）。

在模式识别这块儿，我们可以用概率的机制来衡量模型参数 $\boldsymbol{w}$ 或者模型选择本身的不确定性。以之前曲线拟合为例，在观测数据之前，模型参数 $\boldsymbol{w}$ 服从先验概率 $p(\boldsymbol{w})$，则基于这一参数 $\boldsymbol{w}$ 观测到数据集 $\mathcal{D} = \{t_0, \ldots, t_N\}$ 的条件概率为 $p(\mathcal{D}|\boldsymbol{w})$，该条件概率常被看作参数 $\boldsymbol{w}$ 的函数，也被称为**似然函数（likelihood fuction）**。根据贝叶斯理论：

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \tag{14}$$

由于分母部分

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})\mathrm{d}\boldsymbol{w} \tag{15}$$

是一个常数，所以有：

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{16}$$

不论是概率学派还是贝叶斯学派，似然函数都起到了关键的作用，但是两个不同学派的角度差异主要表现在：

- 频率学派：$\boldsymbol{w}$ 是一个确定的参数，它的值通过某种估计方法来确定，并且通过考虑可能的数据集 $\mathcal{D}$ 的分布来衡量误差线（error bars）。
- 贝叶斯学派：观测数据集 $\mathcal{D}$ 只有一个，$\boldsymbol{w}$ 的概率分布表达了参数的不确定性。

频率学派最常用的估计方法就是**最大似然函数（maximum likelihood）**，寻找使 $p(\mathcal{D}|\boldsymbol{w})$ 值最大参数 $\boldsymbol{w}$。为了求导方便，也为了防治似然函数值过小在计算机中溢出，通常极小化 $-\log p(\mathcal{D}|\boldsymbol{w})$ 进行求解。确定误差线的方法通常采用 **bootstrap** 方法（不是前端里的 bootstrap 框架），该方法通过从原始数据集中有放回随机抽样出 $M$ 个数据集 $\{\mathcal{B}_0, \ldots, \mathcal{B}_M\}$，并且 $|\mathcal{B}_i| = |\mathcal{D}|$，然后利用各种统计量在这 $M$ 个数据集上求得的均值估计真实的统计量。

对比频率学派和贝叶斯学派，其实各有优劣。如果抛一枚公平的硬币，数据集中只有三个样本并且都是正面，则利用频率学派方法（最大似然估

计）训练的模型，在预测下一次抛掷硬币的正反面时基本只会预测正面，但是采用贝叶斯方法（具体方法之后会讨论）能够利用先验知识从而做出不那么极端的预测。贝叶斯方法也有被诟病的地方：

- 很多时候，先验概率只是为了数学上的计算方便，并不是事件真实先验的反应
- 先验知识较差的时候，给出的结果也较差。但是交叉验证等频率验证方法可以避免这些影响
- 贝叶斯公式在整个参数空间积分会带来计算困难（这一缺陷随着采样方法的进步已经得到了极大缓解）

书中说到，最大似然估计会低估分布的方差，从而造成**偏差（bias）**现象，这一现象与过拟合有关，但是为什么会低估分布的方差其实还没有弄明白，这块应该在书中第十章会详细讲。

现在来介绍采用贝叶斯方法是如何训练模型的，还是以曲线拟合为例。假设训练集合有 $N$ 个输入值 $\mathbf{x} = (x_1, \ldots, x_N)^T$，以及对应的目标值 $\mathbf{t} = (t_1, \ldots, t_N)^T$，我们用概率分布来表达目标值的不确定性。假设给定输入值 $x$，其对应的目标变量 $t$ 符合高斯分布（Gaussian distribution），该高斯分布的均值由公式 1 中的 $y(x, \boldsymbol{w})$ 给出，因此有：

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}) \tag{17}$$

其中，$\beta = \frac{1}{\delta^2}$。公式 17 的示意图如图 2 所示（这个图画的实在是太赞了!）

根据贝叶斯公式，似然函数依然发挥着重要作用，所以首先还是应该知道如何求解似然函数：

$$p(\boldsymbol{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_i|y(x_i, \boldsymbol{w}), \beta^{-1}) \tag{18}$$

最大化该似然函数等价于最大化：

$$\ln p(\mathbf{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{i=1}^{N} \{y(x_i, \boldsymbol{w}) - t_i\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \tag{19}$$

如果对 $\boldsymbol{w}$ 进行求导，可以看到 19 的最后两项都会消失，而 $\beta$ 只会起到一个放缩的作用，因此不妨令 $\beta = 1$，从而该问题的形式就变成了与公式 2 相同的形式。通过求解相同的问题，我们可以确定使得似然函数最大的 $\boldsymbol{w}$ 和

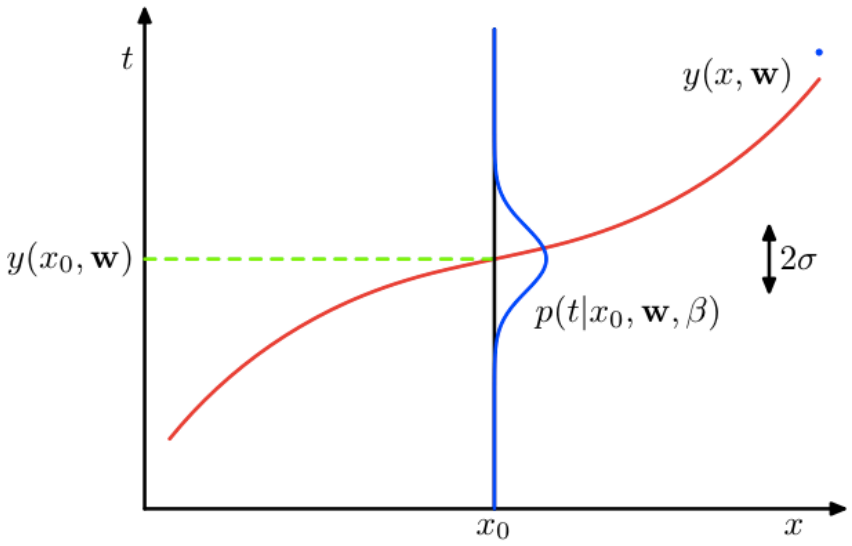


图 2: 给定 $x$ 时目标值 $t$ 的概率分布示意图

$\beta$，之前的问题到这里就已经结束了，但是对于贝叶斯公式来说，似然函数只是其中一部分。

我们还需要知道 $\boldsymbol{w}$ 的先验概率，现在假设

$$\begin{aligned} p(\boldsymbol{w}|\alpha) &= \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\mathbf{I}) \\ &= \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right\} \end{aligned} \tag{20}$$

其中 $\alpha$ 又叫做**超参数（hyperparameters）**，用来控制模型参数的分布。接下来就可以利用贝叶斯公式进行求解了：

$$p(w|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) \times p(\boldsymbol{w}|\alpha) \tag{21}$$

我们可以通过观测数据，选择可能性最大的 $\boldsymbol{w}$，或者说最大化后验概率的分布，这一技巧也叫做**最大后验概率（maximum posterior, MAP）**，在公式 21 外面包一层负对数函数，并结合公式 19 和公式 20，可以发现 MAP 等价于最小化：

$$\frac{\beta}{2}\sum_{i=1}^{N}\{y(x_i, \boldsymbol{w}) - t_i\}^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{22}$$

惊奇的发现经过不断的变化，最大化后验概率自然的带有 $L_2$ 正则项！所以相较于最大似然函数，最大后验概率更不容易过拟合。

# 3 模型选择

由于过拟合等问题，模型在训练集上的效果并不能很好的表示模型的效果，因此当数据量足够大的时候，通常是选出一部分独立的数据组成**验证集（validation set）**，然后在验证集上比较模型效果。一般来说，验证集主要是选择特征、调整超参数时对模型进行初步评估，如果数据量较小而模型迭代了较多次时有可能造成模型对验证集合过拟合，因此还需要第三个集合，也就是**测试集（test set）**，留作最终效果的评估。

如果数据量不大时留出的的验证集可能会受到较为严重的噪声干扰，这时可以进行**交叉验证（cross validation）**，将数据集分为 $S$ 份，每次选择 $\frac{S-1}{S}$ 份作为训练集合，留一个作为验证集。但是交叉验证也存在一些不足：

1. 一个模型训练的次数随着 $S$ 的增大而增大。
2. 参数模型复杂时，不同参数的组合会爆炸增长，采用交叉验证的训练成本比较大。

更好的验证方法会在之后讨论。

# 4 维度的诅咒

这一节的前面一部分讨论了一些在低维空间中的想法，在高维空间中也许并不适用。在高维空间中的困难也叫做**维度的诅咒（curse of dimensionality）**。关于维度诅咒的问题的确非常重要，但是它并没有阻止我们去发现在高维空间中适用的技术：

1. 真实数据总是限制在低维空间的，或者说，对目标变量有重要影响的维度总不会太多
2. 真实数据总是展现出光滑特性，对于大多数情况来说，一个小的改变对与目标变量的影响非常小。

# 5 决策论

联合概率分布 $p(\mathbf{x}, \mathbf{t})$ 提供了对这些变量不确定性的完整总结，给定数据集之后确定 $p(\mathbf{x}, \mathbf{t})$ 的过程也叫做**推断（inference）**阶段。尽管 $p(\mathbf{x}, \mathbf{t})$ 非常有用并且包含了很多信息量，但是我们还是要做出决策来选择最优的预测结果，这一阶段叫**决策（decision）**阶段，在解决推断问题之后，决策阶段往往是比较简单的。

假设现在需要对一个人进行诊断，确定其是否患癌症了，$\mathbf{x}$ 表示病人的 X-光数据，$p(\mathcal{C}_1)$ 表示他患癌症的概率，而 $p(\mathcal{C}_2)$ 表示他没有患癌症的概率。每一个类别 $\mathcal{C}_k$ 都对应一个区域 $\mathcal{R}_k$，该区域内的数据都赋予该类别。那么错误的概率就是：

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{R}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{R}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) \mathrm{d}x + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) \mathrm{d}x \end{aligned} \tag{23}$$

根据乘法规则 $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$，由于 $p(x)$ 是固定的，所以最小化犯错概率等价于最大化 $p(\mathcal{C}_k|\mathbf{x})$，也就是最大化每个数据真实类别的概率：

$$\begin{aligned} p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\ &= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) \mathrm{d}\mathbf{x} \end{aligned} \tag{24}$$

对于一些问题，分错的之后造成的后果是不一样的，所以给定一个损失矩阵，来减少平均损失。例如癌症诊断的问题，如果一个健康的人被诊断为癌症，最多只是会受些痛苦，但是如果一个癌症患者被诊断为无病损失的就可能是一个生命。因此可以定义平均损失函数为：

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \mathrm{d}x \tag{25}$$

其中，以癌症诊断为例的损失矩阵 $L$ 可以是：

| | 癌症 | 健康 |
|---|---|---|
| 癌症 | 0 | 1000 |
| 健康 | 1 | 0 |

之后可以通过最小化 $\mathbb{E}[L]$ 进行决策。

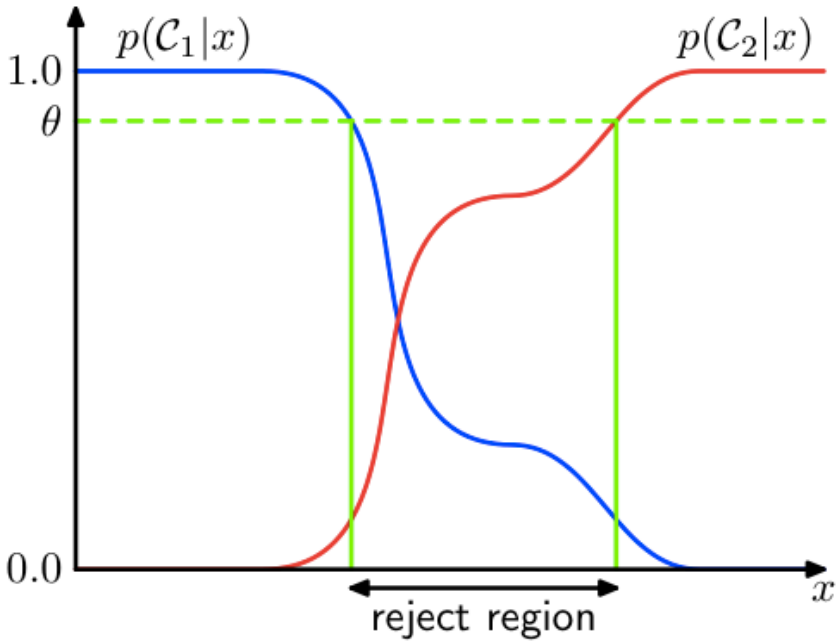


图 3: 对不确定性较大的区域进行驳回

对于一些不确定性较大的数据可以**选择驳回（reject option）**，如图 3 所示，这些数据可以利用更多的其他数据对不确定性进行修正。

现在我们把分类任务分成了两个阶段，推理阶段用来学习后验概率 $p(\mathcal{C}_k|\mathbf{x})$；决策阶段用来对最终的类别进行决策。当然也可以选择将两个阶段融合，直接学习出可以将数据 $\mathbf{x}$ 映射到决策的函数，这种函数叫做**判别函数（discriminant function）**。现在按照复杂程度给出三种解决决策问题的三种方法：

第一种方法，通过确定每个类别的条件概率 $p(\mathbf{x}|\mathcal{C}_k)$ 和先验概率 $p(\mathcal{C}_k)$ 来解决推断问题。然后利用贝叶斯公式学习后验概率

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{26}$$

分母部分可以对类别空间求和解得

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k) \tag{27}$$

类似的对于输入分布进行建模的方法叫做**生成模型（generative models）**。该方法的优点是利用 $p(\mathbf{x})$ 的分布可以用来发现并预测准确度较低的数据点，这也叫做异常值检测或新奇值检测（outlier detection or novelty detection）。但是该方法的缺陷是，(1) 需要大量的训练集；(2) 求解后验概率的过程中，求解 $p(\mathbf{x}, \mathcal{C}_k)$ 时的大量计算就有些浪费了；(3) 事实上 $p(\mathbf{x}|\mathcal{C}_k)$ 包含的很多结构对后验概率 $p(\mathcal{C}_k|\mathbf{x})$ 几乎没有什么影响。

第二种方法，通过直接确定后验概率 $p(\mathcal{C}_k|\mathbf{x})$ 来解决推断问题，然后利用决策理论为输入 $\mathbf{x}$ 赋予类别。直接对后验概率建模的方法也叫做**判别模型（discriminate models）**。该方法的优点是可以简化后验概率的计算开销。

第三种方法，找到一个决策函数 $f(\cdot)$，直接将输入 $\mathbf{x}$ 映射到类别标签。这种方法并不求解后验概率，所以计算度最低。

而利用后验概率的优势主要表现在：

- 面对损失矩阵 $L$ 不断更新的问题，而前两种方法利用后验概率，只需要稍作修改就可以更新预测模型。
- 利用后验概率可以进行驳回选择。
- 可以进行先验概率补偿。数据类别不均匀时一般要平衡数据集，然后在平衡数据集上学习后验概率，利用平衡数据集的后验概率除以平衡数据集中概类别的比例，再乘以原始数据集中该类别的比例，得到新的“后验概率”，之所以加引号是因为还需要进行归一化。
- 当利用一种数据没办法做出决策时，可能需要其它方面数据的支持，假设 $\mathbf{x}_1$ 时 X-光，$\mathbf{x}_2$ 是血液的数据，这两个放在一个模型里可能又不太合适，利用条件独立的假设，$p(\mathbf{x}_1, \mathbf{x}_2|\mathcal{C}_k) = p(\mathbf{x}_2|\mathcal{C}_k)p(\mathbf{x}_2|\mathcal{C}_k)$，在计算新的联合后验概率时可以利用之前的计算结果简化计算开销。条件假设是**朴素贝叶斯模型（naive Bayes model）**的一个例子。