第一章

笔记本:	PRML读书笔记		
创建时间:	2018/7/22 23:02	**更新时间:**	2018/8/5 18:15
作者:	王		
标签:	简介		
URL:	https://baike.baidu.com/item/%E7%9B%B8%E5%AF%B9%E7%86%B5/4233536?fr=aladdin		

1.1
模式识别所领域所关心的问题：通过计算机算法自动发现数据之间的规则，并应用这些规则（比如分类，回归预测等）

在做多项式拟合问题的时候，增加样本数据也是减少过拟合的方法，但是我们不能根据数据量的多少来决定模型参数多少，而是应该根据要解决的问题的复杂度来决定参数量

1.2
var[x] = E[x^2] − E[x]^2
conv(x,y)=E[(x-E[x])(y-E[y])]=E[xy]-E[x]E[y]

频率学派与贝叶斯学派：
　简单来说，当我们在解决一个问题的时候，一般是构建一个问题模型来求解，模型是由许多参数组成的，我们的目的就是找到这些参数，频率学派与贝叶斯学派的最大不同之处就在于对于参数空间的认知上面。频率学派认为一个模型的参数是固定的，求解问题时要做的就是要找到在参数空间中最有可能的参数，所以有极大似然估计和置信区间。而贝叶斯学派则认为参数空间中的值都是可能的，每个值都拥有各自的概率，所以有先验概率和后验概率。

贝叶斯学派的优点：采用先验概率更加自然，显得不那么极端。比如抛硬币，连续三次出现正面，频率学派会将出现正面的概率视为1，而贝叶斯学派却不会。
贝叶斯学派的缺点：依赖于先验知识的选择，先验选择的不好，结果就可能会很差

单变量高斯分布函数：x为单变量

$$\mathcal{N}\left(x|\mu,\sigma^2\right)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad (1.46)$$

多变量高斯分布函数：x为D维向量， Σ为D*D的协方差矩阵， |Σ|为协方差矩阵的行列式

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \qquad (1.52)$$

极大似然估计
极大似然估计存在问题：样本方差的计算是一个有偏估计，因为它有系统性误差，无论怎样进行抽样，样本方差值总是小于理论方差值。之所以会出现这种系统性误差，是由于在计算均值的时候采用的是样本均值而不是总体均值，样本均值比理论上的总体均值更加靠近这一组样本中心。

求解过程：

$$\mathcal{N}\left(x|\mu,\sigma^2\right)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad (1.46)$$

$$p(\mathbf{x}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}\left(x_n|\mu,\sigma^2\right). \qquad (1.53)$$

根据上述两个式子，我们可以得到求最大似然估计的对数概率函数：

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right)=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi). \qquad (1.54)$$

分别进行求导计算之后得到优化目标函数：

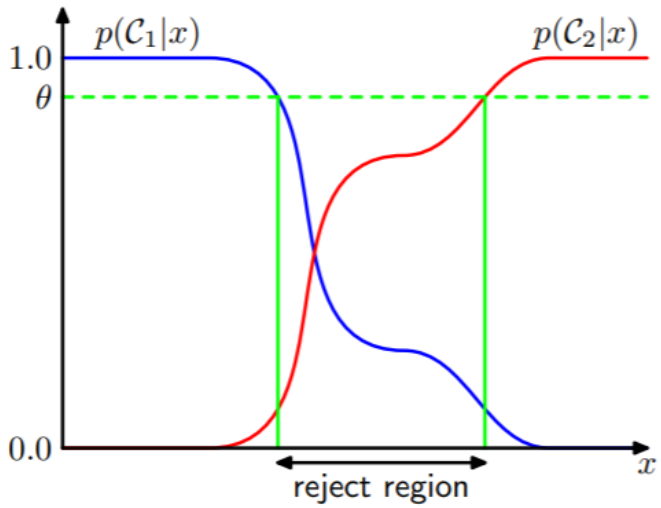
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.55}$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2 \tag{1.56}$$

决策论

reject option: 当采用决策方法进行分类时，对于单个样本分属于每个class的概率很低的时候，或者两个概率很接近的时候，就需要设置一个reject option作为一个threshhold，再进一步进行判定（比如人工干预）。

Figure 1.26 Illustration of the reject option. Inputs x such that the larger of the two posterior probabilities is less than or equal to some threshold θ will be rejected.



信息论

当用位来表示一个随机变量时，信息熵表示这个随机变量的长度下界

将N个物品放到一些箱子中，每个箱子可以存放的物品个数是固定的，那么可以有多少放置的方式？

$$W = \frac{N!}{\prod_i n_i!} \tag{1.94}$$

Kullback–Leibler divergence（KL散度，又称为相对熵）：

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x})\ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x})\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}\mathrm{d}\mathbf{x}. \qquad (1.113)\end{aligned}$$

KL散度是量化两种概率分布之间差异的方式，在概率学或者统计学中我们会用一种更加简单的近似概率分布来替代原有的分布，KL散度所做的就是使用一个分布来近似另一个分布时损失的信息

1.非对称性　　2.非负性

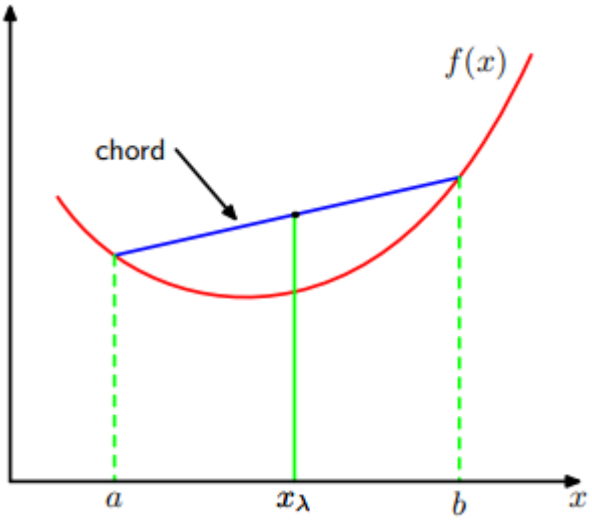
互信息：

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \end{aligned} \tag{1.120}$$

表示两个变量之间的相关程度，即给定一个变量，另一个随机变量不确定性的削弱程度，最小值为0，意味着一个变量对另一个变量没有关系，最大值为随机变量的熵，意味着给定一个随机变量，能够完全消除另一个随机变量的不确定性。

凸函数：

A convex function $f(x)$ is one for which every chord (shown in blue) lies on or above the function (shown in red).



满足条件：

$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda)f(b). \tag{1.114}$$