第二章 概率分布

本章主要介绍概率分布，几个特定的比较常用的概率分布，这些分布主要用于构建更加复杂的模型，讨论一些重要的统计学概念例如贝叶斯推断。

本章通过概率分布讨论密度估计问题，后者指给定一个有限数据集，构建这个数据集对应的分布模型。这里假定数据集的每个数据都是独立同分布。密度估计问题是没有明确唯一答案的，一个有限的数据集可以对应无限种概率分布。

本章讨论参数分布中的二项分布，多项式分布，高斯分布。可以通过共轭先验概率更好地计算这些分布。

最后讨论无参密度估计方法来估计分布，这种分布只取决于数据集的大小，参数只控制模型的复杂度，不控制这个分布。方法有直方图，近邻法，核方法。

2.1 二项式变量

抛硬币实验，硬币不均匀，

$$p(x=1|\mu)=\mu$$

用μ表示硬币坏掉这个性质体现在概率上的变化。

开始进行贝叶斯观点的推断。**第一步**，给定数据集D = {x1, . . . , xN}，构建他的似然函数，即在参数的条件下得到数据集的**概率**，似然函数同时是μ的函数。

$$p(\mathcal{D}|\mu)=\prod_{n=1}^{N}p(x_n|\mu)=\prod_{n=1}^{N}\mu^{x_n}(1-\mu)^{1-x_n}.$$

在频率学家观点里，把似然函数最大化来估计μ，因为D已经出现，所以最大化D的概率得到的μ对于D是最准确的，得到μ以后得到D的分布。相当于通过已知的数据推测未知的数据。贝叶斯观点μ不固定？或者说现在没法固定，通过现有的数据，模拟μ的分布来解决问题

最大化对数的似然函数得到

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

被称为最大似然估计，也被称为样本均值。

这种估计从给定数据集角度来讲很准确，但是有可能发生过拟合的现象。如果给定数据集太少而且缺乏代表性，比如三次都是硬币向上，这种方法就会把以后的结果都预测为朝上，这明显不合理。

接下来扩展到N次试验中m次硬币向上的概率，μ为控制参数，二项分布

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

$$\binom{N}{m}\equiv\frac{N!}{(N-m)!m!}$$

继续使用贝叶斯方法推断。**第二步**，引入一个和似然函数共轭（参数成比例）的**先验分布**，是参数μ的**分布**。a和b是超参数，即参数μ的分布的参数。

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

贝叶斯**第三步**，先验分布乘以似然函数（概率）得到后验分布。没有直接乘（**可能因为参数和概率混在一起，算起来很麻烦？**），而是先得到

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

得到变量的部分。通过和Beta分布比较，还有后验分布正则化的条件，得到系数。m是硬币正面朝上次数，l是正面朝下次数。

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

所以和先验分布比较，在一次抛硬币之后a增加了m，b增加了l，超参数变化对应观测的影响，叫观测的影响数。

这个方法可以把得到的后验分布当做先验分布，然后再加入一组观测数据的影响。可以不断观测，不断优化这个函数，直到接近真实的分布。可以用于流式数据的学习还有在获得所有数据前就要给出预测的学习。

之前得到的后验分布只是μ的分布，实际要的是预测x，所以还有**第四步**，

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})\,\mathrm{d}\mu = \int_0^1 \mu p(\mu|\mathcal{D})\,\mathrm{d}\mu = \mathbb{E}[\mu|\mathcal{D}].$$

p(μ|D)就是刚才得到的后验分布。第一个p(x = 1|D)是概率，因为μ是变量，又得到了他的分布，所以第二个式子是在所有可能μ的情况下对应p(x = 1|D)的值，利用概率乘以概率，结果可以理解为μ根据已有数据集D所有可能分布的均值，x=1 根据这个均值的概率。
后验均值总是在先验均值和最大似然估计之间，当数据集趋向于无穷时，后验均值和最大似然估计相等。

由此可以得到贝叶斯学习的一个性质，当数据集数量增大时，关于参数μ的后验分布的不确定性也开始减小，μ的范围也越来越小。
他给出了具体证明。

$$\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]\right]$$

$$\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta}] = \mathbb{E}_{\mathcal{D}}\left[\mathrm{var}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]\right] + \mathrm{var}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}|\mathcal{D}]\right].$$

θ是参数，D是数据集，一系列计算得到两式。。第一项是后验方差的平均值，第二项是平均值的方差，由于方差是正数，可以得到平均后验方差总是比先验方差小。当然这里是平均，在局部有可能出现比先验方差大的情况。

2.2 多项式变量

一个变量不止两种可能，有多种可能。可以用一种向量形式表示

$$\mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}.$$

x = (x1, x2, x3, x4, x5, x6)T 。六种可能，当xk发生时对应的值为1，其他为0。
对应的大写x的分布如下，μk代表第k个状态对应参数。连乘里面没有发生的状态幂次为0直接消掉。（连乘是想出来的，之前没见过。。）

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{x_k}$$

开始贝叶斯推断**第一步**，给定数据集D，包含向量**x1**到**xN**，构建似然函数，既在已有μ条件下的概率

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}.$$

$$m_k = \sum_n x_{nk}$$

mk决定这个概率，是充分统计数。
似然函数取对数后，通过拉格朗日乘数法和μk加和为一的条件，算出

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

同2.1，扩展到多次结果联合的概率，叫做多项式分布

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K}\prod_{k=1}^{K}\mu_k^{m_k}$$

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1!m_2!\ldots m_K!}.$$

$$\sum_{k=1}^{K} m_k = N.$$

第二步，引入共轭先验，狄利克雷分布（这里说通过参考多项式分布的式子引入，之前说是参考似然函数分布的式子引入，应该前面的写法是对的。总体感觉像是凭空引入的，为了方便拼凑出来，没什么物理意义，只是对照似然找到这么一个方便计算的μ的分布罢了）

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}$$

超参数

$$\alpha_0 = \sum_{k=1}^{K}\alpha_k.$$

第三步，μ先验分布乘以似然函数得到μ的后验分布，同2.1，认为后验分布也是狄利克雷分布

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1}.$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k+m_k-1} \end{aligned}$$

αk的变化代表观测到xk对分布的影响，是观测的影响数。

2.3 高斯分布

从这里开始数学要求提高，要懂得多元微积分，对矩阵求导，加上线性代数忘了很多，矩阵性质不熟悉，所以遇到很大困难。这一小节卡了12天左右，之前没有做题，补做题目的时间花了很多，本来是边看边做题，想要做题目辅助发现很多数学知识缺失，一下发现很多不懂或者忘记的数学概念，所以就一直卡着。

介绍多维高斯分布

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

x是D维向量，Σ是D×D协方差矩阵，维度之间的协方差的矩阵，μ是D维均值向量（第一个不懂又没有介绍的数学概念）。

高斯分布有很多很好的性质，在很多地方出现。一个性质是对于一组给定数据集，高斯分布是使之熵最大的分布，从熵的角度来理解，这样的分布保留的信息最多。

第二个性质，当考虑随机变量的和时，由中心极限定理，随机变量的数目越大，和的分布越接近高斯分布，在实际应用中这个接近很快，从图上看随机变量到十个时就很接近了。

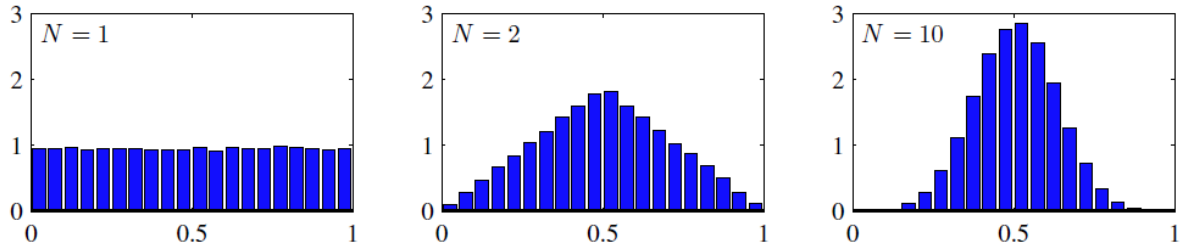


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

现在回到多维高斯分布，引入马氏距离概念（第二个没有学过的数学概念），μ到x的马氏距离为

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

当Σ为单位矩阵时马氏距离退化成欧几里得距离（常用距离），Δ决定多维高斯分布的变化。

通过一系列线性代数的变换（还不会，过三天会了），可以得到

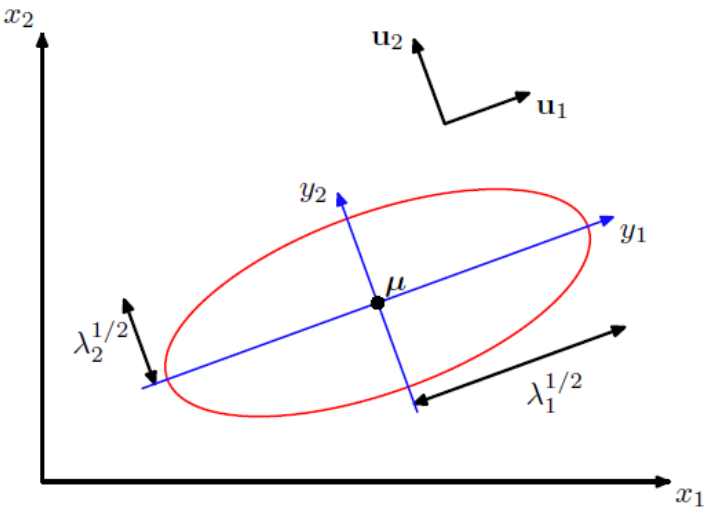
$$\Delta^2 = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}).$$

从式子上，可以把{yi}解读为基于正交向量ui的新的坐标系，通过把原坐标系xi旋转到μ方向得到。

（这个x的坐标系有什么物理意义，有什么源头，作用？感觉像是凭空冒出来一样，对于一个高斯分布来说怎么体现在这样坐标系上面呢，对于一个多维的高斯分布，这种坐标系确实能很方便的表示出点的排列，这个是用这种坐标系一个原因，比如多维的高斯分布就可以看成是分布在一个超椭圆体表面的数据点的集合）

Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



（这个椭圆体的表面代表什么，和高斯分布的概率密度的关系是什么，这也没搞清楚，现在看貌似是这个高斯分布的点构成了这个椭圆，所以才使用这种坐标系，这种坐标系体现了数据点的分布的关系，和概率的关系要怎么体现呢？就是我知道了数据点分布在椭圆上，那数据点的概率也不懂怎么求）

在x坐标系到y坐标系的变换得到雅各宾矩阵（第三个没头没尾的数学概念），

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = U_{ji}$$

Uji是UT的元素，UT的列为向量ui，ui是协方差矩阵的**单位**特征向量。
通过一系列变换（不懂，现在貌似懂了），得到y坐标系下高斯分布的形式

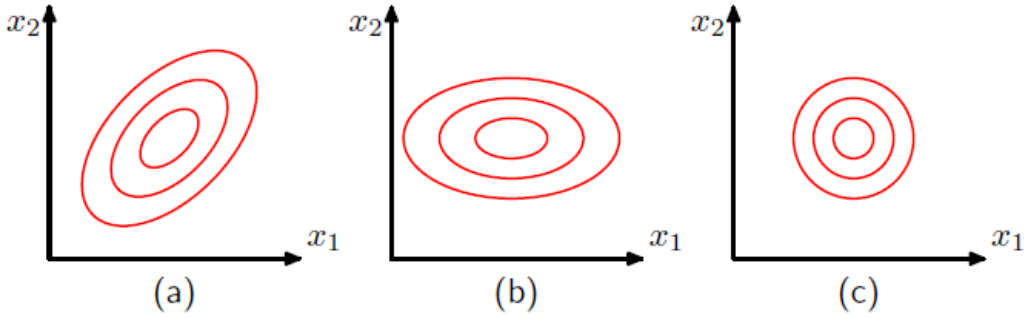
$$p(\mathbf{y}) = p(\mathbf{x})|\mathbf{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\}$$

分布的均值和协方差分别为

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

$$\mathrm{cov}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

高斯分布的缺点：一个高斯分布的独立参数，算上协方差矩阵的D(D + 1)/2，还有μ的D个独立参数，总共有D(D + 3)/2个独立参数，数量太多，对于大的数据集计算量就会很大。一个方法是限制协方差矩阵，让他只能是对角矩阵或者单位矩阵。这种方法限制了模型的形式，削减了高斯分布最大熵的优良性能。如图



另一个缺点，高斯分布是单峰的分布，即只有一个最大值，如果一个分布有多个最大值就很难使用高斯分布模拟。在后文中，引入离散隐变量，进而引入混合高斯分布解决这个问题。类似方法引入连续隐变量，减少参数大小同时模型模拟能力不减弱。

两种方法都有很大的应用，例如马尔科夫随机场和线性动力系统。可以用概率图模型来表示这种方法的结构。

2.3.1 条件高斯分布

多元高斯分布有一个重要性质，如果两组变量的联合分布是高斯分布，那么他的条件分布和边缘分布也是高斯分布。

比如一个D维高斯分布，N(x|μ,Σ)，把x分为

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}.$$

xa占M维，xb占D-M维。均值和协方差可以分为

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

协方差矩阵是对称矩阵，这里再引入精度，即协方差矩阵的逆矩阵

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

现在求xa在xb条件下的分布，一系列变换得到

$$\begin{aligned} \boldsymbol{\mu}_{a|b} &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \end{aligned}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1}.$$

均值μa|b是xb的线性方程，与xa无关，协方差矩阵也和xa无关，所以这叫线性高斯模型。

2.3.2 边缘高斯分布

单求xa的边缘高斯分布，利用下式

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b)\, \mathrm{d}\mathbf{x}_b$$

具体计算不写，思路和之前一样，展开马氏距离的式子，在所求变量的二次项xTa（ ）xa，的括号中间，这个二次项的系数是协方差矩阵，如果是一次项，系数是均值，常数不管。得到

$$\begin{aligned} \mathbb{E}[\mathbf{x}_a] &= \boldsymbol{\mu}_a \\ \mathrm{cov}[\mathbf{x}_a] &= \boldsymbol{\Sigma}_{aa}. \end{aligned}$$

分布可以写为

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}).$$

2.3.3 高斯分布的变量的贝叶斯定理

给定边缘高斯分布p(x)和条件高斯分布p(y|x)，使用之前的结果可以求得p(x|y)，p(y)

> Marginal and Conditional Gaussians
>
> Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form
>
> $$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$
> $$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$
>
> the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by
>
> $$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$
> $$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$
>
> where
>
> $$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$

2.3.4 高斯分布最大似然

给定数据集X = (x1, . . . , xN)T，求其似然函数。类似2.2的步骤，条件概率取对数

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}). \tag{2.118}$$

对μ求导得到期望的似然，用一个比较复杂没看懂的方法，求协方差矩阵，最后得到

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$
$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

发现估计协方差矩阵的期望和真实的协方差矩阵有偏差，修正以后得到

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

2.3.5 逐次估计

通过最大似然求高斯分布的参数，可以看成一个逐步估计的形式。考虑最大似然均值，摘出第N项数据的贡献，可以得到

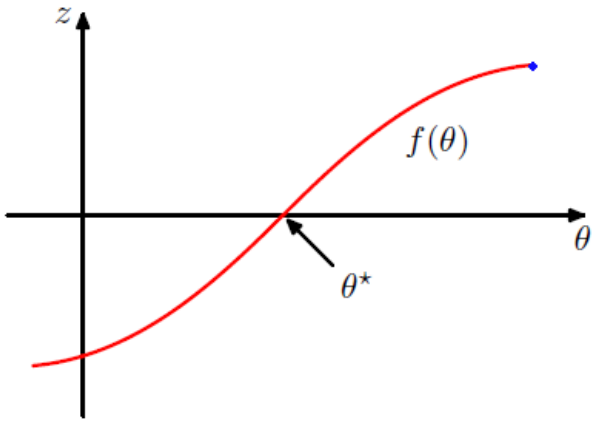
$$\begin{aligned}\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\ &= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\ &= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}).\end{aligned}$$

可以看出，在观测第xn数据之后，把之前由N-1个数据估计的均值，往偏差(xN −μ(N−1)ML)的方向移动一点，得到N个数据估计的均值。

接下来提供一个在线学习更普遍的步骤。给出一个参数θ和z构成的联合分布，z在θ下的条件期望可以构成一个确定性函数（不理解什么叫确定性函数）

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)\,\mathrm{d}z$$

目标是找到使函数等于0的根（为什么）



如果有很多数据，可以直接给这个函数建模，但是在线学习，需要逐步给出数据。提到了Robbins-Monro给的解决办法，定义

$$\theta^{(N)} = \theta^{(N-1)} + a_{N-1}z(\theta^{(N-1)})$$

$$\begin{aligned}\lim_{N\to\infty} a_N &= 0 \\ \sum_{N=1}^{\infty} a_N &= \infty \\ \sum_{N=1}^{\infty} a_N^2 &< \infty.\end{aligned}$$

上标N是第N次观测，aN是自己定义的数字，满足三个限制条件。

接下来回到高斯分布，由似然的定义

$$\frac{\partial}{\partial\theta}\left\{\frac{1}{N}\sum_{n=1}^{N}\ln p(\mathbf{x}_n|\theta)\right\}\Bigg|_{\theta_{\mathrm{ML}}}=0.$$

经过交换，对N取极限，可以得到

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta)=\mathbb{E}_x\left[\frac{\partial}{\partial\theta}\ln p(x|\theta)\right]$$

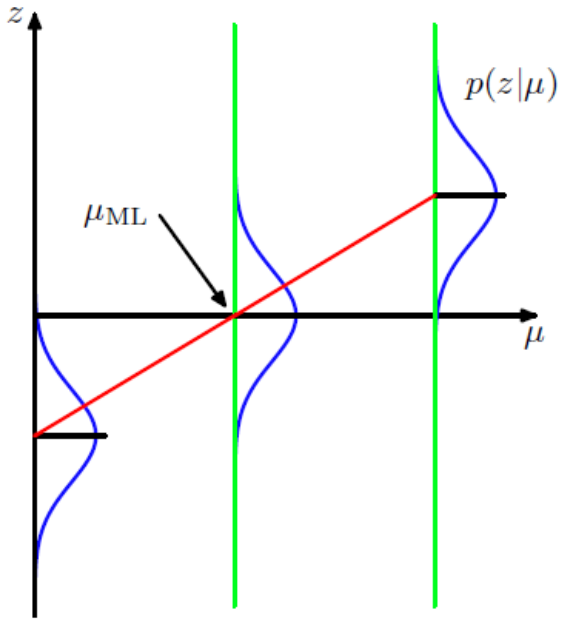
上面这步没毛病，然后就可以看到这个期望也是θ的一个函数，然后找θ的最大似然，就是找这个期望等于0的θ的根，所以可以应用Robbins-Monro的定义得到

$$\theta^{(N)}=\theta^{(N-1)}+a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}\ln p(x_N|\theta^{(N-1)}).$$

回到之前均值最大似然估计的问题，如果把第N次θ当成是第N次最大似然估计的均值，这就是这个方法的一个特例，可以得到

$$z=\frac{\partial}{\partial\mu_{\mathrm{ML}}}\ln p(x|\mu_{\mathrm{ML}},\sigma^2)=\frac{1}{\sigma^2}(x-\mu_{\mathrm{ML}}).$$

z本身也是一个高斯分布，如下图。



2.3.5 高斯分布的贝叶斯推断

之前通过介绍了通过最大似然方法推测高斯分布的均值和协方差，接下来贝叶斯方法推断，先考虑一维高斯分布求方差和均值。

主要分为以下几种

1. 给定方差和数据集，推断均值μ。
2. 给定均值μ和数据集，推断方差。
3. 给定数据集，均值和精度未知，求两个参数的共轭先验(为什么求共轭先验？不接着求了？），得到的共轭先验名叫正态伽马分布

步骤都一样。第一步求数据集在要推断的参数的似然，第二步根据前一步找共轭的参数的先验分布，第三步找在数据集条件下参数的分布。

多维高斯分布，已知均值不知精度，求精度的共轭先验，得到分布命名为Wishart分布。求方差的共轭先验，分布名为逆Wishart分布。均值精度都未知，求它们联合分布的共轭先验，得到分布名为正态Wishart分布。

2.3.7 t分布

如果给定一个单峰高斯分布和一个精度的伽马分布相乘，对精度积分，可以得到x的边缘分布。经过一系列转换得到

$$\mathrm{St}(x|\mu,\lambda,\nu)=\frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2}$$

v是自由度，当v=1是t分布变为柯西分布，v趋向于无穷t分布变为高斯分布

从下图中可以看出t分布的优点，在于t分布受异常数据点的影响更小，高斯分布则比较敏感，这种性质翻译为鲁棒性。字面上理解就是t分布比较坚硬，虽然都是材料可以塑造成某个形态，但是t分布比较坚硬，塑造主要形态以后改变形态比较难。

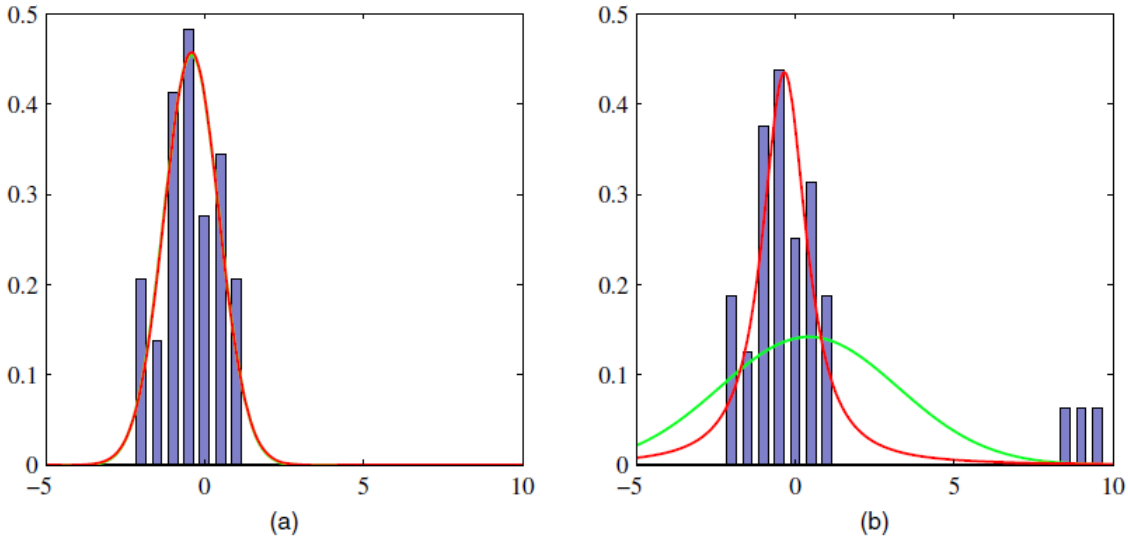


Figure 2.16 Illustration of the robustness of Student's t-distribution compared to a Gaussian. (a) Histogram distribution of 30 data points drawn from a Gaussian distribution, together with the maximum likelihood fit obtained from a t-distribution (red curve) and a Gaussian (green curve, largely hidden by the red curve). Because the t-distribution contains the Gaussian as a special case it gives almost the same solution as the Gaussian. (b) The same data set but with three additional outlying data points showing how the Gaussian (green curve) is strongly distorted by the outliers, whereas the t-distribution (red curve) is relatively unaffected.

鲁棒性在回归问题里也很重要，因为最小二乘法本质上也是求高斯分布的最大似然，所以使用t分布更好。

考试终于考完，开始接着更新，刚好趁着考完把第二章都看完，可以愉快的进行总结了。

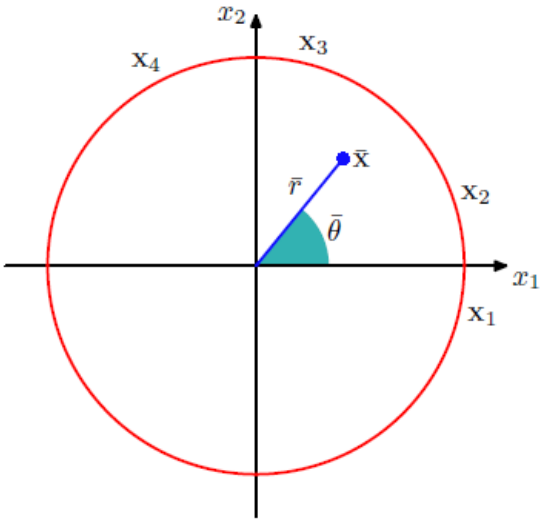
2.3.8 周期变量

高斯分布不适合作为周期分布的密度估计模型。例如在风向问题中使用高斯分布，原点的选择会影响分布的参数。比如有两个数据θ1 = 1°，θ2 = 359°（单位°），选择0°为原点，样本均值为180°，标准差为179°；选择180°为原点，样本均值为0°，标准差为1°，（这种影响对于数据来说是不公平的，选某个角度作为原点对于远离这个原点的角度来说它的值就大了，就变成相对大小的度量，这种相对大小的度量和要观测的目的没有关系，是一个无关的因素，重要的是角度的分布，角度应该是统一的，是没有大小分别的，它本身只是代表方向而已，所以对于建立的预测模型会有失误？会更加麻烦？不知道这种失误，麻烦

具体怎么体现出来）

接下来提供了一个方法，把原来角度的数据集D = {θ1, . . . , θN}变为单位圆上面点代表的单位向量xn的数据集（一个二维数据集，一维是正弦一维是余弦）

Illustration of the representation of values θ_n of a periodic variable as two-dimensional vectors $\mathbf{x}_n$ living on the unit circle. Also shown is the average $\bar{\mathbf{x}}$ of those vectors.



$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

这个均值向量就把相对大小的度量给去掉了，或者说平均化了，数据集中的点这样表示：xn = (cosθn, sin θn)，

通过求解得到均值角度，这个角度就是只代表方向，没有相对大小

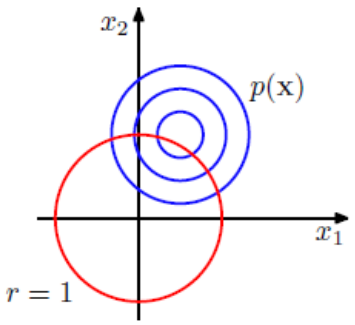
$$\bar{\theta} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}.$$

接下来介绍高斯分布的周期泛化，就是把一个高斯分布转换成适用于周期性变量的变种高斯分布，叫von Mises分布。 根据传统，考虑有2π周期的分布p(θ)，他需要满足三个要求

$$\begin{aligned} p(\theta) &\geqslant 0 \\ \int_0^{2\pi} p(\theta)\,\mathrm{d}\theta &= 1 \\ p(\theta + 2\pi) &= p(\theta). \end{aligned}$$

给出一个二维高斯分布，把原来很多个半径的圆的分布变成只有一个固定半径圆的分布，（不知道为什么，可能是出于消除半径只要方向的动机，也不知道为什么要消除半径）

The von Mises distribution can be derived by considering a two-dimensional Gaussian of the form (2.173), whose density contours are shown in blue and conditioning on the unit circle shown in red.



$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2}\right\}.$$

$$x_1 = r\cos\theta, \qquad x_2 = r\sin\theta.$$

$$\mu_1 = r_0\cos\theta_0, \qquad \mu_2 = r_0\sin\theta_0.$$

利用三角函数知识和一系列转化，定义m = r0/σ2，得到

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)}\exp\{m\cos(\theta-\theta_0)\}$$

这个就叫von Mises分布，也叫圆型正态，θ0是均值，m叫浓度？类似于精度

正则化系数（这是第一类零阶Bessel函数，不知道是什么）

$$I_0(m) = \frac{1}{2\pi}\int_0^{2\pi}\exp\{m\cos\theta\}\,\mathrm{d}\theta.$$

求这个分布的似然，得到

$$\theta_0^{\mathrm{ML}} = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}$$

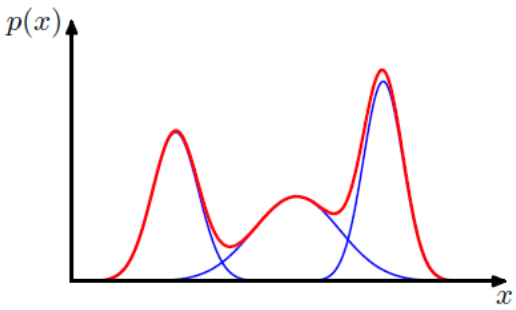
还得到一个像是之前说的浓度的似然，不知道有什么用

$$A(m) = \frac{1}{N}\sum_{n=1}^{N}\cos(\theta_n - \theta_0^{\mathrm{ML}})$$

2.3.9 混合高斯

高斯分布的一个问题是单峰分布，对于有多个峰值的数据点比较好模拟。一个混合的方式就是把几个高斯分布线性组合，叠加在一起，成为一个新的高斯分布模型

Figure 2.22 Example of a Gaussian mixture distribution in one dimension showing three Gaussians (each scaled by a coefficient) in blue and their sum in red.



数学表示

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

$$\sum_{k=1}^{K}\pi_k = 1.$$

$$0 \leqslant \pi_k \leqslant 1.$$

它的似然

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \right\}$$

看起来就很复杂，这里给出了两个方式最大化似然函数，一个叫迭代数值优化，一个叫期望似然。

2.4 指数族

本章讨论的分布，本质上都可以归类为指数族。形式可以这样定义

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

η是参数，g(η)可以理解为确保分布积分为1的系数，u(x)是x的函数

接下来给出了如何把之前给出的三个分布归类到这个形式的推导。直接跳过。

然后求最大似然和充分统计数。

对下式两边求梯度（对于梯度的性质了解不多，不明白求梯度的原因，之前都是对分布求导）

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1$$

得到

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})].$$

通过这个式子，加上后面练习的证明和类比，说只要一个指数族分布是正则化的，就永远可以通过微分找到它的矩（没学过），上网查了查，矩有数学期望是一阶原点矩（表示分布重心）、方差是二阶中心距（表示离散程度）、偏态是三阶中心矩（表示分布偏离对称的程度）、峰态是四阶中心距（描述分布的尖峰程度，例如正态分布峰态系数=0）

通过对似然函数的对数求梯度，得到

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

得到所有u(xn)的加和就是这个分布的充分统计数，如果要求似然函数，充分统计数是需要的数据，其他不需要。

下面直接给出了对应的先验共轭分布，看起来很复杂

$$p(\boldsymbol{\eta}|\boldsymbol{\chi},\nu) = f(\boldsymbol{\chi},\nu)g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right\}$$

2.4.3 无信息先验

在模型推断的实例当中我们如果有背景知识，可以通过调整先验分布应用这些背景知识。有时候要是不知道分布的形式是什么，就要找一种不会影响后验的先验分布，让已有的数据集来决定一个参数的分布是什么，这个先验最好是没有倾向的，对于所有情况都是公平的。

它的思路是把先验当成一个常数，这样对所有情况都公平了。比如一个分布p(x|λ)，被参数λ控制，如果λ是有K个状态的离散变量，很自然他的先验分布就是每个状态概率各占1/K。

到了连续的λ会有问题，它的定义域如果无界，它的先验分布就不能收敛。当然这种先验还是可以用，只要有一个数据点，它的后验就会收敛。

第二个问题在需要对λ非线性变换出现，变换以后的先验分布就不再是一个常数了。

然后给了两个无信息先验的例子，分别为

$$p(x|\mu) = f(x-\mu)$$

$$p(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

第一类具有平移不变性，第二类具有缩放不变性。

2.5 无参方法

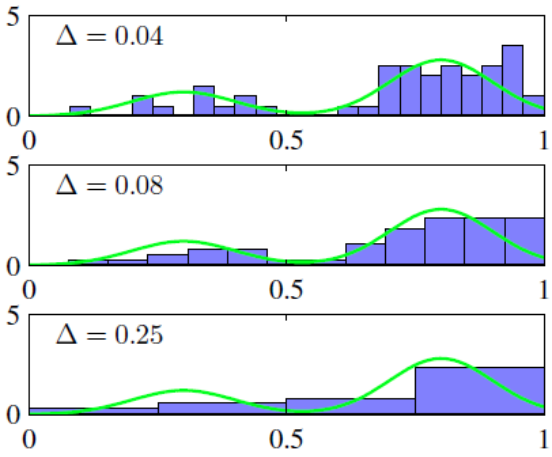
之前介绍的分布，都是被少量参数控制而分布，这种密度估计的方法叫做参数化方法，问题在于如果模型选错了，后面预测的效果就很差，例如多峰分布的数据，使用一个单峰的高斯分布去模拟。所以给几个密度估计的方法，对分布的形式尽量少做假设，这里给的方法是简单的频率学家方法。

第一个方法叫直方图方法，就是把连续变量x的域按一定宽度分成很多份，每个宽度里面x的个数代表这个密度的大小，也是这个区间的概率，宽度用Δi表示，得到密度的式子

$$p_i = \frac{n_i}{N\Delta_i}$$

还有图像

Figure 2.24 An illustration of the histogram approach to density estimation, in which a data set of 50 data points is generated from the distribution shown by the green curve. Histogram density estimates, based on (2.241), with a common bin width Δ are shown for various values of Δ.



从图像可以看出Δ选择太大，图像相比原分布太尖，太小相比原分布太平滑。选择要适中。直方图方法的一个优点，就是把直方做出来后相应的数据就不用保存了，可以节省资源。缺点是直方是有边缘的，对于连续型变量的分布模拟不好。而且直方数量会随着维数倍增，如果是一个D维输入空间，把一个变量的分布分为M个直方，直方的总数量有M的D次方。所以直方一般只是用于一维或者二维。

接下来两个方法从一个式子来

$$p(\mathbf{x}) = \frac{K}{NV}.$$

P是密度p(x)在密度空间R空间的概率，V是R空间的体积，K是给定N个数据点，点在R空间的数量。这里用了两个近似，因为E[K/N] = P，当N很大的时候，这个分布图形很尖锐，K/N处分布的数值就刚好等于P了（就像直方宽度无限小的时候），所以K约等于NP；当R很小的时候，p(x)可以视为一个常数，可以得到概率P约等于p(x)V。这两个假设有些矛盾，如果N足够大的话，R就不会太小，R很小，前面的近似也没法成立。

引发两个方向，第一个是把V当成常数，为K构建函数，叫核方法。最后给出式子

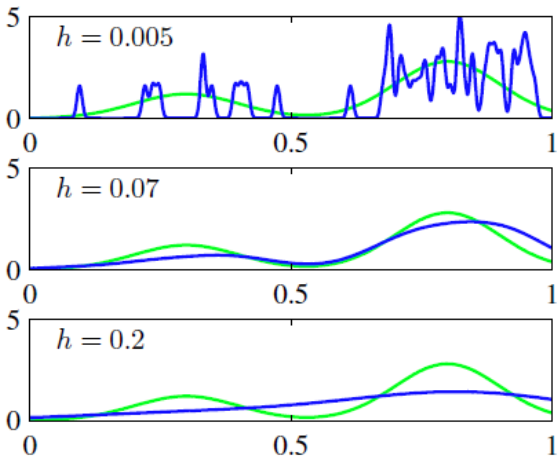
$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}k\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right)$$

下面为核函数是高斯分布，跳跃太大，不理解这个式子的意义

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{1/2}}\exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

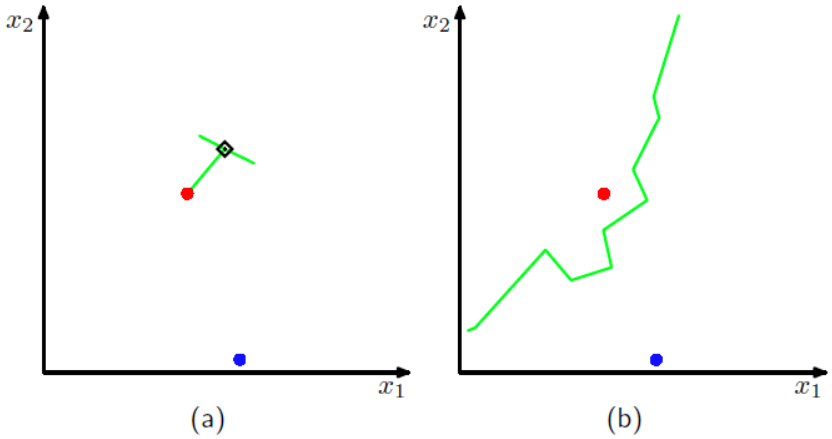
这个图也没办法理解

Figure 2.25 Illustration of the kernel density model (2.250) applied to the same data set used to demonstrate the histogram approach in Figure 2.24. We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model, whereas if it is set too large (bottom panel), then the bimodal nature of the underlying distribution from which the data is generated (shown by the green curve) is washed out. The best density model is obtained for some intermediate value of h (middle panel).



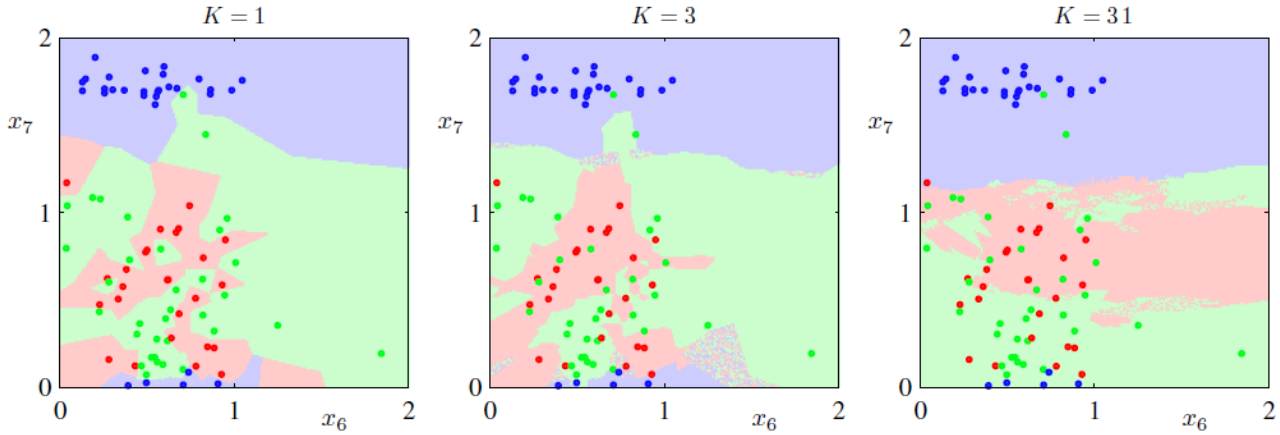
第二个方向是近邻方法，把K当成常数，由距离一个数据点最近的K个点，决定V的大小，然后得到p(x)。它给出了在分类问题的一个分类方法，直观的看就是

Figure 2.27 (a) In the K-nearest-neighbour classifier, a new point, shown by the black diamond, is classified according to the majority class membership of the K closest training data points, in this case $K = 3$. (b) In the nearest-neighbour ($K = 1$) approach to classification, the resulting decision boundary is composed of hyperplanes that form perpendicular bisectors of pairs of points from different classes.



靠近红点的新数据点可以分类为红类，靠近蓝点的可以分类为蓝类

高级一点是这样，K=1对偏离的点也考虑，比较敏感，K=31就是看一块里面谁最多，但是次一点的类别就被忽略了。



无参分布缺点在于需要储存所有数据点，计算花费太大，有很多限制。