

PRML (Pattern Recognition And Machine Learning) 读书会

## 第二章 Probability Distributions

主讲人 网络上的尼采

(新浪微博:@Nietzsche\_复杂网络机器学习)

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码



网络上的尼采(813394698) 9:11:56

开始吧，先不要发言了，先讲 PRML 第二章 Probability Distributions。今天的内容比较多，还是边思考边打字，会比较慢，大家不要着急，上午讲不完下午会接着讲。

顾名思义，PRML 第二章 Probability Distributions 的主要内容有：伯努利分布、二项式 - beta 共轭分布、多项式分布 - 狄利克雷共轭分布、高斯分布、频率派和贝叶斯派的联系、指数族等。

先看最简单的伯努利分布：

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad (2.2)$$

which is known as the *Bernoulli* distribution. It is easily verified that this distribution is normalized and that it has mean and variance given by

$$\mathbb{E}[x] = \mu \quad (2.3)$$

$$\text{var}[x] = \mu(1-\mu). \quad (2.4)$$

最简单的例子就是抛硬币，正反面的概率。

再看二项式分布：

written

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

where

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

抛 N 次有 m 次是正面或反面的概率，所以伯努利分布是二项式分布的特例。

向大家推荐一本好书，陈希孺的《数理统计简史》，对数理统计的一些基本东西的来龙去脉介绍的很详细，这样有助于理解。先 818 二项式分布，正态分布被发现前，二项式分布是大家研究的主要内容。

由二项式分布可以推出其他很多分布形式，比如泊松定理：

$$\lim_{n \rightarrow \infty} \binom{n}{x} p_n^x (1-p_n)^{n-x} = \frac{\lambda^x}{x!} e^{-\lambda}$$

泊松分布是二项式分布的极限形式，这个估计大家都推导过。由二项式分布也能推出正态分布。

贝叶斯思想也是当时对二项式分布做估计产生的，后来沉寂了一百多年。

数据少时用最大似然方法估计参数会过拟合，而贝叶斯方法认为模型参数有一个先验分布，因此共轭分布在贝叶斯方法中很重要，现在看二项式分布的共轭分布 beta 分布：

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \quad (2.13)$$

where  $\Gamma(x)$  is the gamma function defined by (1.141), and the coefficient in (2.13) ensures that the beta distribution is normalized, so that

$$\int_0^1 \text{Beta}(\mu|a, b) d\mu = 1. \quad (2.14)$$

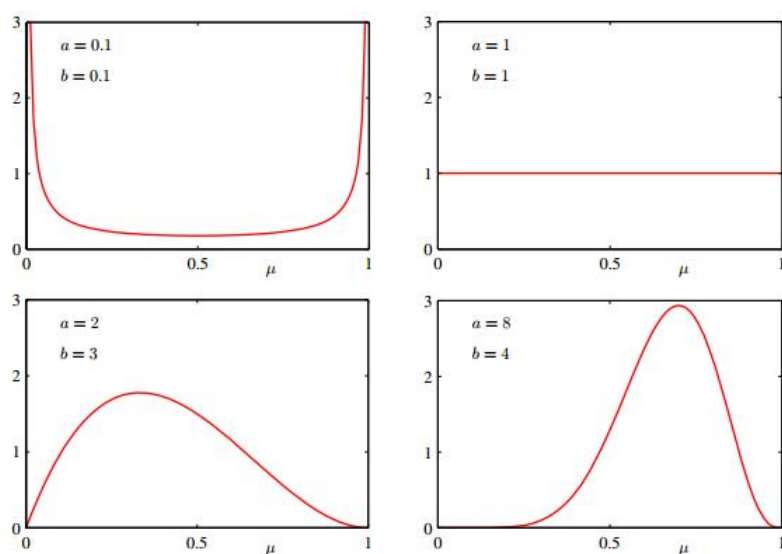
The mean and variance of the beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \quad (2.15)$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}. \quad (2.16)$$

结合上面的二项式分布的形式，不难看出 beta 分布和二项式分布的似然函数有着相同的形式，这样用 beta 分布做二项式分布参数的先验分布，乘似然函数以后得到的后验分布依然是 beta 分布。

a b 是超参，大家可以看到 beta 分布的形式非常灵活：



**Figure 2.2** Plots of the beta distribution  $\text{Beta}(\mu|a, b)$  given by (2.13) as a function of  $\mu$  for various values of the hyperparameters  $a$  and  $b$ .

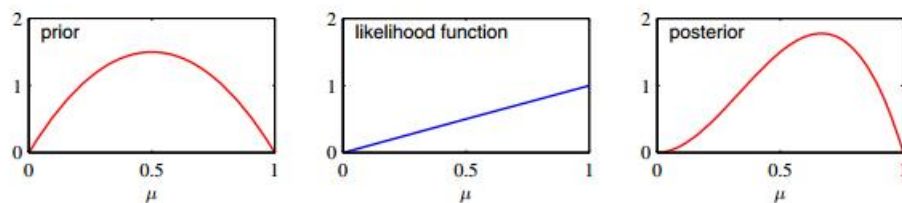
假设抛硬币  $N$  次， $l$  和  $m$  分别为正反面的记数，那么参数的后验分布便是：

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a-1} (1-\mu)^{l+b-1}. \quad (2)$$

不难看出，后验分布是先验和数据共同作用的结果。

这种数据矫正先验的形式可以通过序列的形式进行，非常适合在线学习。

单拿一步来说明问题：



**Figure 2.3** Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters  $a = 2$ ,  $b = 2$ , and the likelihood function, given by (2.9) with  $N = m = 1$ , corresponds to a single observation of  $x = 1$ , so that the posterior is given by a beta distribution with parameters  $a = 3$ ,  $b = 2$ .

可以看出， $a$  的记数增加了 1。

书上通过序列数据流的形式来矫正先验的描述，每次可以用一个观测数据也可以用 small batches，很适合实时的学习：

We see that this *sequential* approach to learning arises naturally when we adopt a Bayesian viewpoint. It is independent of the choice of prior and of the likelihood function and depends only on the assumption of i.i.d. data. Sequential methods make use of observations one at a time, or in small batches, and then discard them before the next observations are used. They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets. Maximum likelihood methods can also be cast into a sequential framework.

回到上面的二项式-beta 共轭，随着数据的增加， $m, l$  趋于无穷大时，这时参数的后验分布就等于最大似然解。

有些先验分布可以证明，随着数据的增加方差越来越小，分布越来越陡，最后坍缩成狄拉克函数，这时贝叶斯方法和频率派方法是等价的。举个第三章的贝叶斯线性回归的例子，对于下图中间参数  $W$  的高斯先验分布，随着数据不断增加，参数后验分布的不确定性逐渐减少，朝一个点坍缩：

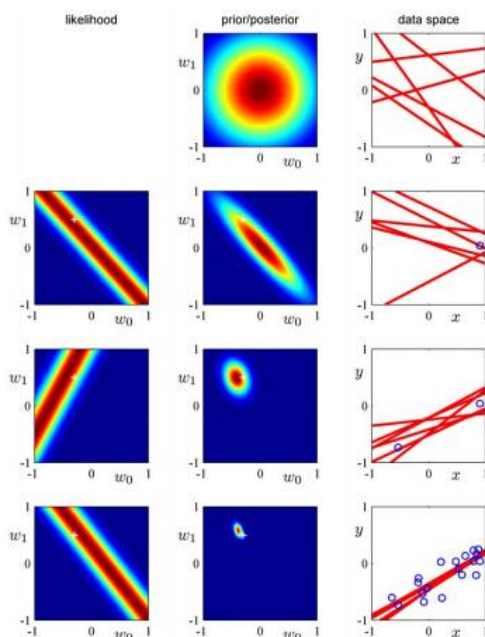


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form  $y(x, \mathbf{w}) = w_0 + w_1 x$ . A detailed description of this figure is given in the text.

接着看多项式分布：把抛硬币换成了掷骰子

We can consider the joint distribution of the quantities  $m_1, \dots, m_K$ , conditioned on the parameters  $\mu$  and on the total number  $N$  of observations. From (2.29) this takes the form

$$\text{Mult}(m_1, m_2, \dots, m_K | \mu, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k} \quad (2.34)$$

which is known as the *multinomial* distribution. The normalization coefficient is the number of ways of partitioning  $N$  objects into  $K$  groups of size  $m_1, \dots, m_K$  and is given by

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! m_2! \dots m_K!}. \quad (2.35)$$

Note that the variables  $m_k$  are subject to the constraint

$$\sum_{k=1}^K m_k = N. \quad (2.36)$$

同样它的共轭分布狄利克雷分布也得和似然函数保持相同的形式。

狄利克雷分布：

The normalized form for this distribution is by

$$\text{Dir}(\mu | \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1} \quad (2.38)$$

which is called the *Dirichlet* distribution. Here  $\Gamma(x)$  is the gamma function defined by (1.141) while

$$\alpha_0 = \sum_{k=1}^K \alpha_k. \quad (2.39)$$

后验形式：

Multiplying the prior (2.38) by the likelihood function (2.34), we obtain the posterior distribution for the parameters  $\{\mu_k\}$  in the form

$$p(\mu|\mathcal{D}, \alpha) \propto p(\mathcal{D}|\mu)p(\mu|\alpha) \propto \prod_{k=1}^K \mu_k^{\alpha_k+m_k-1}. \quad (2.40)$$

We see that the posterior distribution again takes the form of a Dirichlet distribution, confirming that the Dirichlet is indeed a conjugate prior for the multinomial. This allows us to determine the normalization coefficient by comparison with (2.38) so that

$$\begin{aligned} p(\mu|\mathcal{D}, \alpha) &= \text{Dir}(\mu|\alpha + \mathbf{m}) \\ &= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1) \cdots \Gamma(\alpha_K + m_K)} \prod_{k=1}^K \mu_k^{\alpha_k+m_k-1} \end{aligned} \quad (2.41)$$

大家依然能看到记数。

下面讲高斯分布，大家看高斯分布的形式：

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

多元高斯分布的形式：

multivariate Gaussian distribution takes the form

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

高斯分布有着优良的性质，便于推导，很多时候会得到解析解。一元高斯分布是个钟形的曲线，大部分都集中在均值附近，朝两边的概率呈指数衰减，这个可以用契比雪夫不等式来说明，偏离均值超过 3 个标准差的概率就非常低了：

The standard deviation is a simple but valuable measure of how far values of  $x$  are likely to depart from the mean. Its very name suggests that it is the standard or typical amount one should expect a randomly drawn value for  $x$  to deviate or differ from  $\mu$ . *Chebyshev's inequality* (or Bienaymé-Chebyshev inequality) provides a mathematical relation between the standard deviation and  $|x - \mu|$ :

$$\Pr\{|x - \mu| > n\sigma\} \leq \frac{1}{n^2}. \quad (45)$$

This inequality is not a tight bound (and it is useless for  $n < 1$ ); a more practical rule of thumb, which strictly speaking is true only for the normal distribution, is that 68% of the values will lie within one, 95% within two, and 99.7% within three standard deviations of the mean (Fig. A.1). Nevertheless, Chebyshev's inequality shows the

正态分布是如何发现的，在《数理统计简史》有详细的介绍，当时已经有很多人包括拉普拉斯在找随机误差的分布形式，都没有找到，高斯是出于一个假设找到的，也就是随机误差分布的最大似然解是算数平均值，只有正态分布这个函数满足这个要求。

然后高斯进一步将随机误差的正态分布假设和最小二乘联系到了一块，两者是等价的：

使用这个误差分布，就容易对最小二乘法给出一种解释。回到四章的方程(3)，其中  $(x_0, \dots, x_n)$ ,  $i=1, \dots, n$ , 是观测数据。记

$$e_i = x_{0i} + x_1\theta_1 + \dots + x_n\theta_n, 1 \leq i \leq n.$$

理论它们应为 0, 但因有测量误差存在, 实际不必为 0, 故  $e_1, \dots$ ,



$e_n$  可视为误差. 按高斯的第一个原则(极大似然), 结合误差密度 (11),  $(e_1, \dots, e_n)$  的概率为

$$(\sqrt{2\pi}h)^{-n} \exp\left\{-\frac{1}{2h^2} \sum_{i=1}^n (x_{0i} + x_{1i}\theta_1 + \dots + x_{ki}\theta_k)^2\right\}.$$

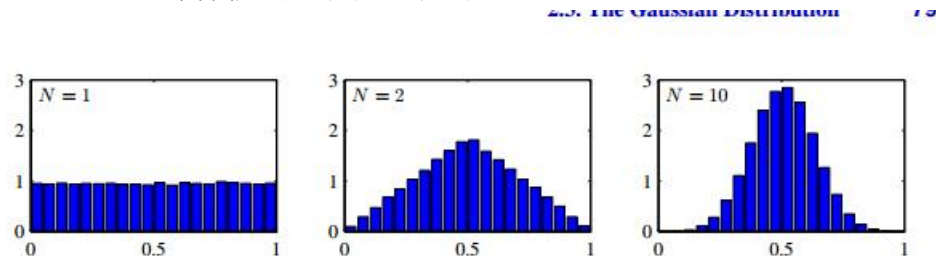
要此式达到最大, 必须取  $\theta_1, \dots, \theta_k$  之值, 使表达式  $\sum_{i=1}^n (x_{0i} + x_{1i}\theta_1 + \dots + x_{ki}\theta_k)^2$  达到最小, 于是得到  $\theta_1, \dots, \theta_k$  的最小二乘估计. 要注

后来就是拉普拉斯迅速跟进, 提出了中心极限定理, 大量随机变量的和呈正态分布, 这样解释了随机误差是正态分布的原因. 中心极限定理的公式:

Now consider  $N$  random variables with pdf's (not necessarily Gaussian)  $p(x_i)$ , each with mean  $\mu$  and variance  $\sigma^2$ . We assume each variable is **independent and identically distributed** or **iid** for short. Let  $S_N = \sum_{i=1}^N X_i$  be the sum of the rv's. This is a simple but widely used transformation of rv's. One can show that, as  $N$  increases, the distribution of this sum approaches

$$p(S_N = s) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right) \quad (2.96)$$

大家看 PRML 上的图, 很形象的说明高斯分布是怎么生长出来的:



**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.

从[0,1]随机取  $N$  个变量, 然后算它们的算术平均, 随着  $N$  的增大, 均值的分布逐渐呈现出高斯分布, 可以比较直观的了解中心极限定理。

接着看高斯分布的几何形式:

先给出样本到均值的马氏距离  $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$

把协方差矩阵的逆  $\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$  带入上式

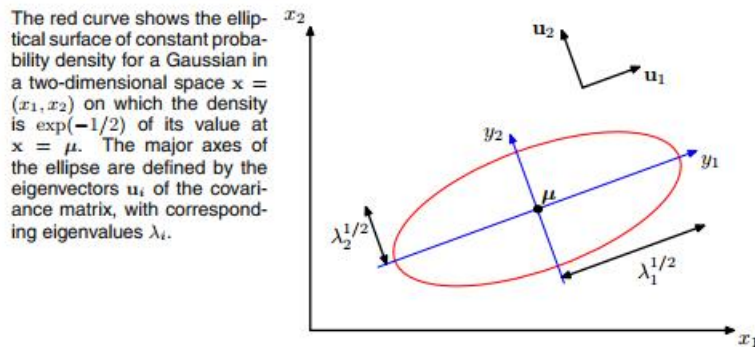
$$\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i}$$

会得到以协方差矩阵的特征值平方根为轴长的标准椭圆方程

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

$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu})$ , 也就是原来的坐标系经过平移和旋转, 由协方差矩阵特征向量组成的矩阵  $\mathbf{U}$  负责旋转坐标轴。

看下面张图就很明白了:



接着是条件高斯分布和边缘高斯分布，这两个分布由高斯分布组成，自身也是高斯分布。

条件高斯分布的推导过程略过，大家记住这个结论：

From these we obtain the following expressions for the mean and covariance of the conditional distribution  $p(\mathbf{x}_a|\mathbf{x}_b)$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \quad (2.81)$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \quad (2.82)$$

上面是条件高斯分布的均值和方差，以后的 Gaussian Processes 在最后预测时会用到均值。

另一个是线性高斯模型  $p(y|x)$  均值是  $x$  的线性函数，协方差与  $x$  独立，也会经常用到。

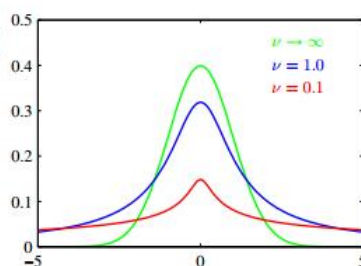
接下来是关于高斯分布的贝叶斯方法，方差已知均值未知，先验用高斯分布；均值已知方差未知用 Gamma 分布；都不知道用 Gaussian-Gamma distribution。这方面的推导略过，大家用到时翻书查看就行了：

2. **Conjugate prior:** lead to posterior distribution having the same functional form as the prior

Distribution	Conjugate Prior
Bernoulli	Beta distribution
Multinomial	Dirichlet distribution
Gaussian, Given variance, mean unknown	Gaussian distribution
Gaussian, Given mean, variance unknown	Gamma distribution
Gaussian, both mean and variance unknown	Gaussian-Gamma distribution

接下来看 Student t-distribution，Student 是笔名，此人在数理统计史上是非常 nb 的人物。

**Figure 2.15** Plot of Student's t-distribution (2.159) for  $\mu = 0$  and  $\lambda = 1$  for various values of  $\nu$ . The limit  $\nu \rightarrow \infty$  corresponds to a Gaussian distribution with mean  $\mu$  and precision  $\lambda$ .



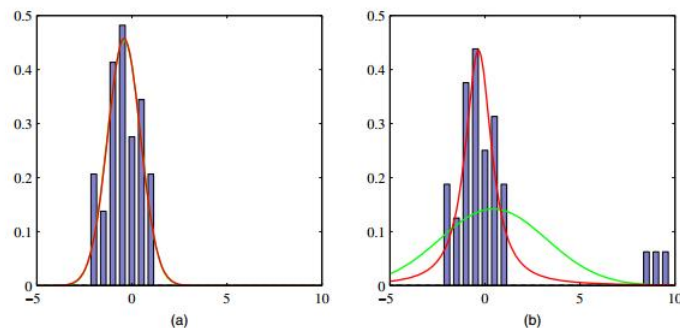
$$\begin{aligned} p(x|\mu, a, b) &= \int_0^\infty \mathcal{N}(x|\mu, \tau^{-1}) \text{Gam}(\tau|a, b) d\tau \quad (2.158) \\ &= \int_0^\infty \frac{b^a e^{(-b\tau)} \tau^{a-1}}{\Gamma(a)} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(x-\mu)^2\right\} d\tau \\ &= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left[b + \frac{(x-\mu)^2}{2}\right]^{-a-1/2} \Gamma(a+1/2) \end{aligned}$$

where we have made the change of variable  $z = \tau[b + (x-\mu)^2/2]$ . By convention we define new parameters given by  $\nu = 2a$  and  $\lambda = a/b$ , in terms of which the distribution  $p(x|\mu, a, b)$  takes the form

$$\text{St}(x|\mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \quad (2.159)$$

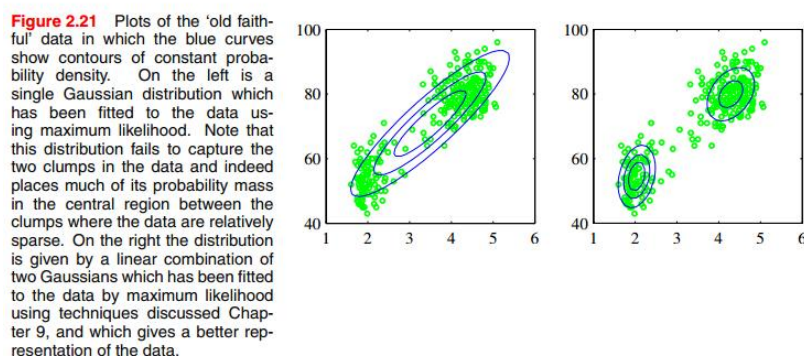
上面是 t 分布的形式，具体如何发现的可以参看《数理统计简史》，大家看上面的积分形式，t 分布其实是无限个均值一样，方差不同的高斯分布混合而成，高斯分布是它的特例，相比较高斯分布，t 分布对 outliers 干扰的鲁棒性要强很多。

从这个图就可以看出，高斯分布对右边孤立点的干扰很敏感，t 分布基本上没有变化：



**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.

接着讲混合高斯分布：看下图里的例子，单个高斯分布表达能力有限，无法捕捉到两个簇结构：



我们可以多个高斯分布的线性组合来逼近复杂的分布，并且对非指数族的分布也一样有效。

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$$

混合高斯分布的形式：

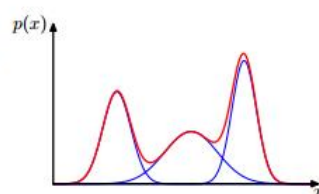
RIVERS(773600590) 11:01:09

可不可以使用非线性的组合呢？

网络上的尼采(813394698) 11:01:48

那就太复杂了

**e 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.



这个图是三个高斯分布混合逼近一个

复杂分布的例子。

混合高斯模型里面有一个隐变量，也就是数据点属于哪个高斯分布。



这个就是隐变量的期望：

$$\begin{aligned}\gamma_k(\mathbf{x}) &\equiv p(k|\mathbf{x}) \\ &= \frac{p(k)p(\mathbf{x}|k)}{\sum_l p(l)p(\mathbf{x}|l)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}.\end{aligned}$$

这个是我们的最大似然目标函数：

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

可以用 EM 算法，一边是隐变量，一边是模型的参数，迭代着来回倒腾，收敛到局部最优。混合高斯我在第九章详细讲了，感兴趣的可以看下原来的记录。

xunyu(2118773) 11:09:18

隐变量和最大似然函数的联系在哪里

落英缤纷(348609341) 11:10:16

不设置隐变量直接用 ML 不好解

网络上的尼采(813394698) 11:10:31

下面讲指数族，很多分布包括我们上面提到的二项式分布、beta 分布、多项式分布、狄利克雷分布、高斯分布都可以转换成这种指数族的形式：

The exponential family of distributions over  $\mathbf{x}$ , given parameters  $\boldsymbol{\eta}$ , is defined to be the set of distributions of the form

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp \left\{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}) \right\} \quad (2.194)$$

其中 $\boldsymbol{\eta}$ 是参数， $g(\boldsymbol{\eta})$ 是归一化因子， $\mathbf{u}(\mathbf{x})$ 是  $\mathbf{x}$  的函数。

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left( \prod_{n=1}^N h(\mathbf{x}_n) \right) g(\boldsymbol{\eta})^N \exp \left\{ \boldsymbol{\eta}^T \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) \right\}.$$

指数族的似然函数：










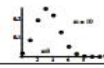
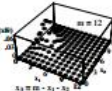
对  $\ln p(\mathbf{X}|\boldsymbol{\eta})$  关于  $\boldsymbol{\eta}$  求导，令其等于 0，会得到最大似然解的形式：

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

很显然， $\sum_{n=1}^N \mathbf{u}(\mathbf{x}_n)$  是充分统计量。充分统计量其实很好理解，拿最简单的二项式分布来说，抛硬币我们只需要记住正反面出现的次数就行，原来的数据就可以丢弃了。

DUDA 是指数族专家，这是从他书上截的图，大家可以看下表中的指数族：

Table 3.1: Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		$\mathbf{s}$	$[g(\mathbf{s}, \boldsymbol{\theta})]^{1/n}$
Normal	$p(x \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n} \sum_{k=1}^n x_k \\ \frac{1}{n} \sum_{k=1}^n x_k^2 \end{bmatrix}$	$\sqrt{\theta_2} e^{-\frac{1}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multi-variate Normal	$p(\mathbf{x} \boldsymbol{\theta}) = \frac{ \boldsymbol{\Theta}_2 ^{1/2}}{(2\pi)^{d/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\theta}_1)^T \boldsymbol{\Theta}_2 (\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\Theta}_2$ positive definite		$\begin{bmatrix} \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \\ \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T \end{bmatrix}$	$ \boldsymbol{\Theta}_2 ^{1/2} e^{-\frac{1}{2}[\text{tr} \boldsymbol{\Theta}_2 s_2 - 2\boldsymbol{\theta}_1^T \boldsymbol{\Theta}_2 s_1 + \boldsymbol{\theta}_1^T \boldsymbol{\Theta}_2 \boldsymbol{\theta}_1]}$
Exponential	$p(x \theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta e^{-\theta s}$
Rayleigh	$p(x \theta) = \begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n} \sum_{k=1}^n x_k^2$	$\theta e^{-\theta s}$
Maxwell	$p(x \theta) = \begin{cases} \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n} \sum_{k=1}^n x_k^2$	$\theta^{3/2} e^{-\theta s}$
Gamma	$p(x \boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\begin{bmatrix} \left( \prod_{k=1}^n x_k \right)^{1/n} \\ \frac{1}{n} \sum_{k=1}^n x_k \end{bmatrix}$	$\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} s_1^{\theta_1} e^{-\theta_2 s_2}$
Beta	$p(x \boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1} (1-x)^{\theta_2} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > -1$		$\begin{bmatrix} \left( \prod_{k=1}^n x_k \right)^{1/n} \\ \left( \prod_{k=1}^n (1-x_k) \right)^{1/n} \end{bmatrix}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$
Poisson	$P(x \theta) = \frac{\theta^x}{x!} e^{-\theta} \quad x = 0, 1, 2, \dots$	$\theta > 0$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta^s e^{-\theta}$
Bernoulli	$P(x \theta) = \theta^x (1-\theta)^{1-x} \quad x = 0, 1$	$0 < \theta < 1$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta^s (1-\theta)^{1-s}$
Binomial	$P(x \theta) = \frac{m!}{x!(m-x)!} \theta^x (1-\theta)^{m-x} \quad x = 0, 1, \dots, m$	$0 < \theta < 1$		$\frac{1}{n} \sum_{k=1}^n x_k$	$\theta^s (1-\theta)^{m-s}$
Multinomial	$P(\mathbf{x} \boldsymbol{\theta}) = \frac{m!}{\prod_{i=1}^d x_i!} \prod_{i=1}^d \theta_i^{x_i} \quad \begin{matrix} x_i = 0, 1, \dots, m \\ \sum_{i=1}^d x_i = m \end{matrix}$	$0 < \theta_i < 1$ $\sum_{i=1}^d \theta_i = 1$		$\frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$	$\prod_{i=1}^d \theta_i^{s_i}$

指数族的共轭先验形式： $p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^\nu \exp \{ \nu \boldsymbol{\eta}^T \boldsymbol{\chi} \}$

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp \left\{ \boldsymbol{\eta}^T \left( \sum_{n=1}^N \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi} \right) \right\}.$$

后验形式：