这一系列（机器学习的数学基础）主要包括目前学习过程中回过头复习的基础数学知识的总结。

基础知识：conjugate priors共轭先验

共轭先验是指这样一种概率密度：它使得后验概率的密度函数与先验概率的密度函数具有相同的函数形式。它极大地简化了贝叶斯分析。

如何解释这句话。由于

P(u|D) = p(D|u)p(u)/p(D) （1.0式）

其中D是给定的一个样本集合，因此对其来说p(D)是一个确定的值，可以理解为一个常数。P(u|D)是后验概率----即观察到一系列样本数据后模型参数服从的概率，p(D|u)是似然概率----在给定的模型参数u下样本数据服从这一概率模型的相似程度，p(u)是u的先验概率---- 在我们一无所知的情况下u的概率分布。P(u|D)的函数形式完全由p(D|u)和p(u)的乘积决定。如果p(u)的取值使p(u|D)和p(D|u) 相同的表达形式（关于u的表达形式），就称p(u)为共轭先验。一个最简单的p(u)的取值就是常数1，只不过1是p(u)的一种取值。

在了解了共轭先验的概念后，我们主要针对二项分布和多项分布找到他们的共轭先验呈现出什么样的形式，从而引出Dirichlet分布的概念。

二项分布和Beta分布：

如果随机变量x的取值只能取0或1，则称x为服从二项分布的随机变量：

$$\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \quad (1.1式)$$

其中u为p(x=1)。注意上式中x只能为0或1，因此当x为0时p=1-u，当x为1时p=u。写成上面的乘积形式是为了数学描述的方便。

如果对此二值实验重复进行N次，出现的结果将会有m次1和N-m次0。出现m次1和N-m次0的概率为

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \quad (1.2式)$$

又把它称为伯努利实验。给定一个数据集D={x1,x2,x3……xN}，其似然函数可以写为：

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \quad (1.3式)$$

现在来看这个概率模型（这是一个似然概率）。式1.2前面的括号项可以理解为一个概率的归一化系数，它与u无关。我们考虑与u有关的这个部分。为了使后验概率具有相同的数学结构，我们引入beta函数

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \quad (1.4式)$$

这样得到的后验概率就具有以下形式：

$$p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1} \quad \text{(1.5式)}$$

其中l=N-m。可以看出1.5式和1.2式具有相同的形式，都是u和1-u的指数的乘积。因此beta分布就是二项分布的共轭先验分布，其中a和b就是beta分布的参数。

在进入到Dirichlet分布之前，我们再观察一下Beta分布。1.4式与1.2式的形式是一样的（除去前面的归一化系数不管）。而不一样的地方在于 1.2式中要求N和m都为整数，而Beta函数中的a和b可以是任意实数（其中当a为整数时r(a)=(a-1)!）。换句牛逼的总结，Beta函数将伯努利实验的概率从整数扩展到了所有实数。

先验概率取为conjugate prior的好处在于做贝叶斯推断。以 二项分布为例，如果我们只有一个观测样本（假设样本观测值为1），那么后验概率仍然是1.4式的形式，只不过a的值更新为a+1。往后如果再有新的观测数 据，就把上一次的后验概率作为先验，乘以新数据的似然函数，就能更新到新的后验概率（传统的做法则是用先验概率乘以所有数据的似然函数得到后验概率）。这一sequential method与传统做法得到的后验概率结果是完全一致的（注意仅仅在我们讨论的这些例子中是这样，如果先验概率不选择为conjugate prior一般不会有这样的等效(2013.5.9note:不会有这样的等效么？再不确定)）。而sequential method的优点则在于每用于更新一次后验概率后观测样本可以不用记录下来----这对于大规模的数据下做模型训练是非常有用的。

而Dirichlet分布就是多项分布的共轭先验分布。因此要理解Dirichlet分布，先看看多项分布。

多项分布和Dirichlet分布

如果x的取值有K种情况，就称x服从多项分布。往往用维数为K的矢量来描述。矢量中仅可能一个xk取值为1，其他都为0，用来描述x取第k个值。这样其概率分布可以描述为：

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

其中 $\mu_k \geqslant 0$ 且 $\sum_k \mu_k = 1$。当对多项分布的事件进行多次，取值为1至K项的事件分别发生mk次的概率则为：

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

与beta分布之于二项分布一样，我们找寻多项分布的共轭先验，其共轭先验应该具有这样的形式：

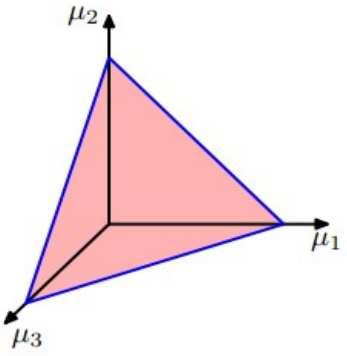
$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

归一化后的表达形式为：

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

这个分布就叫做Dirichlet分布，其中**α**是dirichlet分布的参数，**μ**是变量。

由于限制$\sum_k \mu_k = 1$且0≤uk≤1，因此u1,u2……uk被限制在单纯形中（下图以k=3为例展示了这个单纯形，注意这个单纯形是一个平面，而不是那个三角体。因为$\sum_k \mu_k = 1$使得u1,u2,u3虽然有三个参数但实际自由度为2，换句话说可以投影到u1-u2的平面上成为一个平面三角形）。



在上面这个介绍的例子中，可以将Dirichlet分布理解为概率的概率。因为u表示的是多项分布的概率，而Dir(u)表达的是u取某种值情况下的概 率，所以可以理解为概率的概率。举个经典的例子，扔骨子。很显然这是一个多项分布，骨子的呈现只可能是1-6中的一种情况。如果我们将这个事件重复 10000次，其中出现1-6的次数分别为2000,2000,2000,1500，1500,1000，那么u的取值就是（0.2,0.2,0.2,0.15,0.15,0.1）。那么Dirichlet概率描述的就是u取值为（0.2,0.2,0.2,0.15,0.15,0.1）的概率。