2.1 二元变量

伯努利分布

- 分布形式:

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

- 期望:

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

- 方差:

$$\mathbf{var}[\mathbf{x}] = \mu(1-\mu)$$

- 似然:

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

二项分布

- 分布形式:

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

- 期望:

$$\mathbf{E}[\mathbf{x}] = N\mu$$

- 方差:

$$\mathbf{var}[\mathbf{x}] = N\mu(1-\mu)$$

2.1.1 beta 分布

- 分布形式:

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

,$\Gamma(x) = \int_0^\infty u^x e^{-u} du$.可知，$\int_0^1 Beta(\mu|a,b)\mathrm{d}u = 1$

- 期望:

$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

- 方差:

$$\frac{ab}{(a+b)^2(a+b+1)}$$

将Beta先验与二项似然函数$Bim(m|N,\mu)$相乘之后归一化，只保留与μ相关的变量，则得到后验分布有以下形式:

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

我们可以发现后验概率同样是一个Beta分布，得到归一化系数之后，我们可以得到:

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

所以在接受贝叶斯观点的基础上，我们可以很自然地得到学习过程的顺序方法。其先验与似然函数的选择是无关的，只取决于数据同分布的假设。所以顺序学习很适用于实时学习的场景。
根据概率的加和与成绩规则，我们可以得到

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

.
所以我们继而可以得到:

$$p(x=1|D) = \frac{m+a}{m+a+l+b}$$

当数据集无限大的时候，这个结果与似然得到的结果是一样的。
随着观测数据的不断增多，后验概率的不确定性会不断下降。当数据有限时，后验概率的结果会处于先验与最大似然估计得到的结果之间。
对于一个参数为θ的贝叶斯推断问题而言，我们观测到一个数据集$\mathcal{D}$,有联合概率密度$p(\theta, D)$:

$$\mathbb{E}_\theta[\theta] = \int p(\theta)\theta\mathrm{d}\theta$$

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\theta}[\theta|\mathcal{D}]] = \iint \theta p(\theta|\mathcal{D})\mathrm{d}\theta p(\mathcal{D})\mathrm{d}\mathcal{D}$$

可以得到：

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

所以θ的后验均值，在产生数据集的整个分布上做平均等于θ的先验均值。

$$\mathbf{var}_{\theta}[\theta] = \mathbb{E}_{\mathcal{D}}[\mathbf{var}_{\theta}[\theta|\mathcal{D}]] + \mathbf{var}_{\mathcal{D}}[\mathbf{E}_{\theta}[\theta|\mathcal{D}]]$$

所以θ的先验方差等于其平均后验方差加上其后验均值的方差。这表明平均来看，θ的后验方差小于先验方差。但是这只在平均情况下成立，在某些情况下依然会出现后验方差大于先验方差的情况。

多项式变量

- 形式

$$p(x|\mu) = \prod_{k=1}^{K} \mu^{x_k}, \sum_{k=1}^{K} x_k = 1$$

- 期望

$$\mathbb{E}[x|\mu] = \sum_{x} p(x|\mu)x = (\mu_1, \cdots, \mu_K)^T = \mu$$

考虑在$m_1, \ldots, m_K$在参数μ和观测总数N的条件下的联合分布为

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

其中$\sum_{k=1}^{K} m_k = N$

2.2.1 狄利克雷分布

先验：$p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \sum_{k=1}^{K} \mu_k = 1.$

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

其中$\alpha_0 = \sum_{k=0}^{K}\alpha_k$.
将先验与似然结合可以得到后验：

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

经过归一化之后可以得到：

$$p(\mu|D,\alpha) = \mathbf{Dir}(\mu|\alpha+m) = \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\dots\Gamma(\alpha_K+m_K)}$$

所以二变量可以表示为二元变量然后使用二项分布来建模，也可以看成one of two, 然后用多项分布来建模。

2.3.9 混合高斯模型

对基本的概率分布进行线性组合被形式化为概率模型。用足够多的高斯分布并且调节其均值与方差以及线性组合的系数，极具所有连续概率密度都可以以任意精度近似。
下面对K个高斯概率密度叠加：

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

每一个高斯概率密度被称为混合密度的一个成分。其中$0 \le \pi_k \le 1$,$\sum_{k=1}^{K}\pi_k = 1$.
其中$\pi_k = p(k)$可以看作是第k个成分的先验概率。
为了确定这些参数的值，我们可以使用最大似然法。对数似然函数为$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}(x_n|\mu_k,\Sigma_k)\right)$,其中$X = \{x_1,\dots,x_N\}$.
因为对数中存在求和式，所以不存在闭式解，所以可以用迭代数值优化的方式来求解，或者可以使用 EM 。

2.4 指数族分布

定义：

参数为η的指数族分布：

$$p(x|\eta) = h(x)g(\eta)e^{\eta^T u(x)}$$

,x可以是标量也可以是向量，可以离散也可以连续。η是概率分布的自然参数，$u(x)$是x的某一个函数。$g(\eta)$可以看成系数，确保概率可以归一化。

$$g(\eta)\int h(x)e^{\eta^T u(x)}\mathrm{d}x = 1$$

实例：

伯努利分布：

$$p(x|\mu) = Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

我们可以将其化为：

$$p(x|\mu) = \exp(x\ln\mu + (1-x)\ln(1-\mu)) = e^{x\ln(\mu)-(x-1)\ln(1-\mu)} = (1-\mu)\exp(\ln(\frac{\mu}{1-\mu})x)$$

其中$\eta = \ln(\frac{\mu}{1-\mu}), \mu = \sigma(\eta)$,由于$1-\sigma(x) = \sigma(-x)$
所以可以得到

$$p(x|\mu) = \sigma(-\eta)\exp(\eta x)$$

所以可以得到：

$$u(x) = 1, h(x=1), g(\eta) = \sigma(-\eta)$$

单一观测x的多项式分布：

$$p(x|\mu) = \prod_{k=1}^{M}\mu_k^{x_k} = \exp(\sum_{k=1}^{M}x_k\ln\mu_k)$$

其中$x=(x_1,\dots,x_M)$.所以可以得到:

$$p(x|\mu)=\exp(\eta^T x)$$

其中$\mu_k=\ln\mu_k$,$\eta=(\eta_1,\dots,\eta_M)$
可以得到:

$$u(x)=x,h(x)=1,g(\eta)=1$$

此时参数之间并不是独立的，并且有如下限制：$\sum_{k=1}^{M}\mu_k=1$
当给定了M-1个参数的时候，剩下的参数就固定了，当去掉这个限制，用M-1个参数来表示这个分布的时候，有$0\le\mu_k\le 1,\sum_{k=1}^{M-1}\mu_k\le 1$
此时多项式分布变成

$$\begin{aligned}\exp\left(\sum_{k=1}^{M}x_k\ln\mu_k\right)&=\exp\left(\sum_{k=1}^{M-1}x_k\ln\mu_k+\left(1-\sum_{k=1}^{M-1}x_k\right)\ln\left(1-\sum_{k=1}^{M-1}\mu_k\right)\right)\\&=\exp\left(\sum_{k=1}^{M-1}x_k\ln\frac{\mu_k}{1-\sum_{j=1}^{M-1}\mu_j}+\ln\left(1-\sum_{k=1}^{M-1}\mu_k\right)\right)\end{aligned}$$

此时令

$$\ln\frac{\mu_k}{1-\sum_j\mu_j}=\eta_k$$

对于两侧对k进行加和

$$\mu_k=\frac{\exp(\eta_k)}{1+\sum_j\exp(\eta_j)}$$

这被称为softmax函数。
因此多项式分布

$$p(x|\eta)=\left(1-\sum_{k=1}^{M-1}\exp(\eta_k)\right)^{-1}\exp(\eta^T x)$$

令$\eta = (\eta_1, \ldots, \eta_{M-1})^T$可以得到：

$$u(x) = x, h(x) = 1, g(\eta) = (1 + \sum_{k=1}^{M-1} \exp(\eta_k))^{-1}$$

高斯分布

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2) \end{aligned}$$

其中

$$\begin{aligned} \eta &= \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \\ u(x) &= \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ h(x) &= (2\pi)^{-\frac{1}{2}} \\ g(\eta) &= (-2\eta - 2)^{\frac{1}{2}} \exp(\frac{\eta_1^2}{4\eta_2}) \end{aligned}$$

2.4.1 最大似然估计和充分统计量

对

$$g(\eta) \int h(x) e^{\eta^T u(x)} \mathrm{d}x = 1$$

两侧进行求导，可以得到：

$$\nabla g(\eta) \int h(x) \exp(\eta^T \mu(x)) \mathrm{d}x + g(\eta) \int h(x) \exp(\eta^T \mu(x)) u(x) \mathrm{d}x = 0$$

可以得到：

$$-\frac{\nabla g(\eta)}{g(\eta)} = g(\eta)\int h(x)\exp(\eta^T\mu(x))u(x)\mathrm{d}x = \mathbb{E}[u(x)]$$

最后可以得到:

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

对数据集$X = (x_1, \ldots, x_N)$,似然函数为:

$$p(X|\eta) = \prod_{i=1}^{N}(h(x_i))g(\eta)^N \exp(\eta^T\sum_{n=1}^{N}u(x_n))$$

,令关于η的导数等于零，可以得到:

$$-\nabla \ln g(\eta_{ML}) = \frac{1}{N}\sum_{i=1}^{N}u(x_n)$$

当N趋于无穷时，右侧变为$\mathbb{E}[u(x)]$,可以看到在极限情况下，η_{ML}与真实的η相等

2.4.2 共轭先验

对于指数族分布来说，都有共轭先验:

$$p(\eta|\mathcal{X},\nu) = f(\mathcal{X},\nu)g(\eta)^\nu \exp(\nu\eta^T\mathcal{X})$$

其中$f(\mathcal{X},\nu)$是归一化系数，$g(\eta)$与指数族分布定义中的相同。
将先验与似然相乘，得到忽略归一化系数的后验概率:

$$p(\eta|X,\mathcal{X},\nu) \propto g(\eta)^{\nu+N}\exp(\eta^T(\sum_{n=1}^{N}u(x_n)+\nu\mathcal{X}))$$

可以看到，先验与后验形式相同，证明了共轭性。所以ν可以看作是先验分布中假想观测的有效观测数目。

2.5 非参数化方法

参数化方法用于概率密度建模有其局限性，比如对于一个多峰分布，就不可能被单峰的高斯分布描述出来。因此引入非参数化方法。

直方图法：

标准的直方图法将x划分到不同的宽度为Δ_i的箱子内，然后对每一个箱子内部的样本数n_i进行统计，然后每一个宽度为Δ_i的箱子所占的概率值为

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

所以可以很容易的得到：

$$\int p(x)\mathrm{d}x = 1$$

通常情况下每一个Δ_i的大小是一致的，这样每一个宽度内的概率值都是恒定的。
所以一旦直方图计算出来，数据就没用了，这对于大型数据集来说很有用，并且很适用于不断到来的数据。
在实际应用上，直方图很适用于一维，二维数据的可视化，但是在大多数密度估计问题并不适用。一是因为生成的概率密度并不连续，二是因为 维度灾难 ，在多维空间中所需的数据量太多。
可以提供的启发：首先，进行密度估计的时候，应关注数据点的邻域。其次，为获得好的结果，平滑参数的值不能太小也不能太大。

核估计法

对于一个在$\mathcal{D}$维空间中未知的概率密度$p(x)$,我们想进行密度估计，则我们考虑一个包含x的小的区域$\mathcal{R}$，该区域的概率质量为

$$P = \int_{\mathcal{R}} p(x)\mathrm{d}x$$

从$p(x)$中得到的N个观测值，每个数据点落到$\mathcal{R}$区域的概率是P,$\mathcal{R}$内落有K个点的概率分布为

$$\mathbf{Bin}(K|N,P) = \frac{N!}{K!(N-K)!}P^K(1-P)^{1-K}$$

可以得到：

$$\mathbb{E}[\frac{K}{N}] = P$$
$$\mathbf{var}[\frac{K}{N}] = \frac{P(1-P)}{N}$$

对于大的N，可以得到$K \simeq NP$,如果V是区域R的体积，那么$P \simeq p(x)V$，可以得到概率密度估计的形式为

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

现在我们固定V,从数据中确定K，那么就得到核方法，为了确定K的数目，我们使用一个核函数

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leq \frac{1}{2} \\ 0, & \mathbf{otherwise} \end{cases}$$

所以使用$k(\frac{x-x_n}{h})$就可以判断在以x为中心h范围内数据点的数量

$$K = \sum_{n=1}^{N} k(\frac{x-x_n}{h})$$

于是根据上式，令$V = h^D$可以得到:

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k(\frac{x-x_n}{h})$$

但是用这个核函数k会导致概率密度估计不连续，我们可以用高斯核函数进行代替就得到:

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{\frac{D}{2}}} \exp(-\frac{||x-x_h||^2}{2h^2})$$

其中h表示高斯分布的标准差。
选择核函数的标准:

$$k(\mathbf{u}) \geq 0$$
$$\int k(\mathbf{u}) \mathrm{d}u = 1$$

最近邻法

与核方法相对应的是最近邻法，在核方法中我们固定V,由数据来确定K,在最近邻法中，我们选择固定K。
我们对于一个以x为中心的球体，估计概率密度$p(x)$,我们允许球体半径自由增加，直到精确包含K个数据点。最后得到概率密度形式为

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

但是由K近邻法得到的模型并非真正的概率密度模型，因为它在整个空间中的积分是发散的。
我们使用K近邻法来进行分类。利用贝叶斯定理：
我们假设在N个数据所组成的数据集中，类别C_k所包含的数据点数目为N_k,所以我们有

$$\sum_{k=1}^{K} N_k = N$$

当我们想要判别一个新的数据点x所属的类别时，我们用一个以精确包含K个数据点的x为中心的球。假设这个球的体积是V,并且包含C_k类的数据点数目为K_k。则我们可以得到:

$$p(x|C_k) = \frac{K_k}{N_k V}$$

而我们根据前面提到的

$$p(x) = \frac{N_k}{NV}$$

以及:

$$p(C_k) = \frac{N_k}{N}$$

所以我们可以得到：

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{K_k}{K}$$

因此如果想要最小化错误率，我们将最小化$\frac{K_k}{K}$得到应该分类的C_k.
而对于K=1的最近邻法，当$N \to \infty$时，我们可以得到其错误率不会超过最优分类器（即使用真实概率分布的分类器）可以达到的最小错误率的二倍。

非参数化方法的缺陷

对于K近邻法和核密度估计方法都需要存储整个训练数据集，如果数据集很大的话，会带来很大的计算负担。