3.11最大似然估计法 -张满满.md

原理

给定一个概率分布D，假定其概率密度函数（连续分布）或概率聚集函数（离散分布）为*f***D*，以及一个分布参数θ，我们可以从这个分布中抽出一个具有*n*个值的采样X1,X2,...,Xn，通过利用*f***D*，我们就能计算出其概率：

$$P=(x_1,x_2,...,x_n)=f_D(x_1,x_2,...,x_n|\theta)$$

但是，我们可能不知道θ的值，尽管我们知道这些采样数据来自于分布D。那么我们如何才能估计出θ呢？一个自然的想法是从这个分布中抽出一个具有*n*12n

一旦我们获得，我们就能从中找到一个关于θ的估计。最大似然估计会寻找关于 θ的最可能的值（即，在所有可能的θ取值中，寻找一个值使这个采样的"可能性"最大化）。这种方法正好同一些其他的估计方法不同，如θ的非偏估计，非偏估计未必会输出一个最可能的值，而是会输出一个既不高估也不低估的θ值。

要在数学上实现最大似然估计法，我们首先要定义可能性：

$$lik(\theta)=f_D(x_1,x_2,...,x_n|\theta)$$

并且在θ的所有取值上，使这个函数最大化。这个使可能性最大的值即被称为θ的**最大似然估计**

#解决的问题：

-它是建立在极大似然原理的基础上的一个统计方法，极大似然原理的直观想法是，一个随机试验如有若干个可能的结果A，B，C，...，若在一次试验中，结果A出现了，那么可以认为实验条件对A的出现有利，也即出现的概率P(A)较大。极大似然原理的直观想法我们用下面例子说明。设甲箱中有99个白球，1个黑球；乙箱中有1个白球．99个黑球。现随机取出一箱，再从抽取的一箱中随机取出一球，结果是黑球，这一黑球从乙箱抽取的概率比从甲箱抽取的概率大得多，这时我们自然更多地相信这个黑球是取自乙箱的。一般说来，事件A发生的概率与某一未知参数 有关，
取值不同，则事件A发生的概率
也不同，当我们在一次试验中事件A发生了，则认为此时的 值应是t的一切可能取值中使
达到最大的那一个，极大似然估计法就是要选取这样的t值作为参数t的估计值，使所选取的样本在被选的总体中出现的可能性为最大。

-极大似然估计，只是一种概率论在统计学的应用，它是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。极大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。当然极大似然估计只是一种粗略的数学期望，要知道它的误差大小还要做区间估计

-在已经得到试验结果的情况下，我们应该寻找使这个结果出现的可能性最大的那个参数
 作为真 参数 的估计

#例题

- $f(x;p) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$

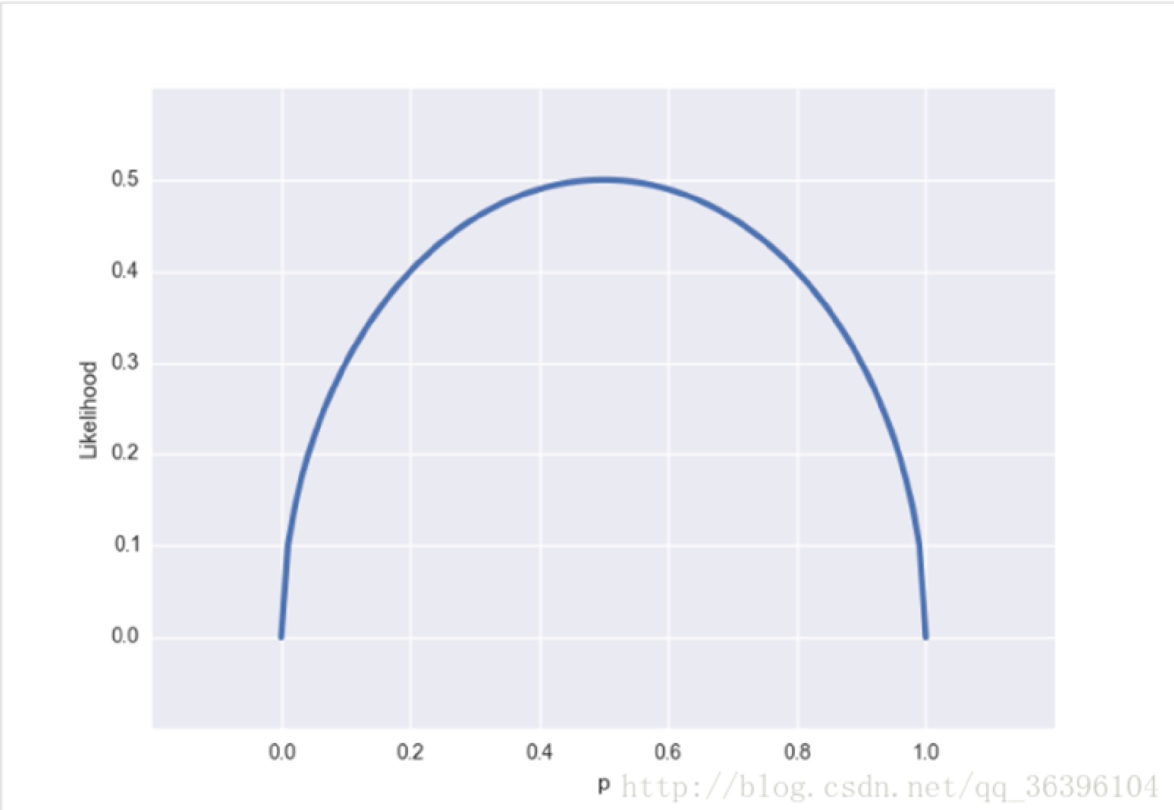
•以伯努利分布（Bernoulli distribution，又叫做两点分布或0-1分布）为例：

•也可以写成以下形式：

$$f(x;p) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0,1\}$$

•这里注意区分 f(x;p)f(x;p) 与前面的条件概率的区别，引号后的 pp 仅表示 ff 依赖于 pp 的值，pp 并不是 ff 的前置条件，而只是这个概率分布的一个参数而已，也可以省略引号后的内容：

- $$f(x) = p^x(1-p)^{1-x} \quad \text{for } x \in \{0,1\}$$



似然函数的最大值

似然函数的最大值意味着什么？让我们回到概率和似然的定义，概率描述的是在一定条件下某个事件发生的可能性，概率越大说明这件事情越可能会发生；而似然描述的是结果已知的情况下，该事件在不同条件下发生的可能性，似然函数的值越大说明该事件在对应的条件下发生的可能性越

#实例代码

```
import numpy as np
import matplotlib.pyplot as plt
```

```
fig = plt.figure()
mu = 30  # mean of distribution
sigma = 2  # standard deviation of distribution
x = mu + sigma * np.random.randn(10000)

def mle(x):
    """
    极大似然估计
    :param x:
    :return:
    """
    u = np.mean(x)
    return u, np.sqrt(np.dot(x - u, (x - u).T) / x.shape[0])

print(mle(x))
num_bins = 100
plt.hist(x, num_bins)
plt.show()
```

最大似然估计法的应用

无论是在有监督还是无监督，判别模型还是生成模型，但凡是和概率有挂钩的，最终是模型是预测概率的，都少补了最大似然估计的应用。

3.1、有监督学习

3.1.1 逻辑回归分类（判别模型==>条件概率）

- 目标：对于新来的样例，预测其属于y=1 该类的概率
- 已有数据：样例x，标签y。
- 事件：在样例X(i)=x(i)的条件下，类别是y。（这是已知的，这个事件也是服从一个由参数θ控制的分布的。）
  于是得到模型：

$$p(y \mid x; \theta) = (h_\theta(x))^y \left(1 - h_\theta(x)\right)^{1-y}$$

对于所有的样本来说，在样例取得m个值的情况下，m个类别分别是y的概率。就是这些小事件一起发生的概率。于是有极大似然函数：

$$
\begin{aligned}
L(\theta) &= p(\vec{y} \mid X; \theta) \\
&= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}
\end{aligned}
$$

image.png

于是此时，找到使L(θ)最大的参数θ，就能够使上述事件尽可能的发生， 也是最接近实际值的θ了。于是可以用来预测。

3.1.2、高斯判别模型（生成模型 ==>联合分布）

为什么是生成模型，因为这里认为，数据（样本，类别）都是在满足这些分布的情况下生成的。
 判别的时候，模型表达的意思，“先采样生成类别y，再采样生成新来样例xi”，这个事件发生的概率，那个大，就说明更符合实际情况。比如在类别是1的情况下， 采样生成新来样例的概率是0.6，在类比是2的情况下采样生成新来样例的概率 是0.8，那么新来阳历属于类比2的情况更符合实际。

- 目标：每个类别服从一个分布P(Y=y) = p(y)，确定类别以后每个样例也服从一个分布P(X=x|Y=y) ~p，学习完后，最终可以用“先采样生成一个类别标签，在已知类别标签的情况下采样生成新来样例”的概率，来判断数据哪一类 。
- 数据：样例x，标签y
- 事件：1、同时观测到（x，y），于是我们可以认为**一个事件是（X=x，Y=y）同时发生**。2、由联合分布公式可知，p(x,y)=p(x|y)p(y)。于是我们也可以认为，**一个事件（x，y）是 先采样得到y，再在y的条件下采样生成x得到的** 。
  所以此时，我们想要知道的是，y的分布（伯努利分布），以及在y确定的情况下x的分布（多值高斯分布），于是可以得到模型。

$$
\begin{aligned}
p(y) &= \phi^y (1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) \\
p(x|y=1) &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)\right)
\end{aligned}
$$

已有的m个数据对，就是取到m个（x，y）数据对的事件，它发生的概率为：

$$\begin{aligned}\ell(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma)\\ &= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi).\end{aligned}$$

找到上式中的参数，使上述事件尽可能的发生，就是要估计的参数了。

并且，参数的实际意义是 可以根据表达式理解出来的。也就是最接近似然函数的情况下，参数的理想状况。
比如对上面目标函数求导以后得到各参数的值。其中

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

这里的φ代表类别是1的概率，就等于 样本中y=1对的个数除以样本总数m。

3.2、无监督学习

- 目标，对于新来的样例，预测其属于某一类（k个类）的概率
- 已有数据： 样例x
- 事件：不同于有监督学习中，（有监督：一个事件是（X=x，Y=y）同时发生，y已经确定，所以可以直接用p(x,y)=p(x|y)p(y)来 表示此事件。）
  此时的每个事件，就是样例x发生。（但是每个样例都有k个可能的类与之对应，所以需要全概率公式。）所以得到每个事件的模型：

$$p(x^{(i)};\phi,u,\Sigma) = \sum_{z^{(i)}=1}^{k} p(x^{(i)}\,|\,z^{(i)};u,\Sigma)p(z^{(i)};\phi)$$

那么数据就是代表着，m个事件X=x同时发生的概率：

$$\begin{aligned}l(\phi,u,\Sigma) &= \prod_{i=1}^{m} p(x^{(i)};\phi,u,\Sigma)\\ &= \prod_{i=1}^{m}\sum_{z^{(i)}=1}^{k} p(x^{(i)}\,|\,z^{(i)};u,\Sigma)p(z^{(i)};\phi)\end{aligned}$$

但是，这个式子一开始并不好求，于是我们先随机为每个样例选一个相应的类别，，，接下来就是EM思想，可以看EM算法这一块。

每个类别的概率是所有样例的后验概率的平均值（参考GMM）

总之，最大化这个似然函数，最终得到的，也是我们想要的参数。

3.3 最大后验概率估计（MAP）

逻辑回归中的模型是，认为θ是一个常数，一个事件就是，在样例X=x的条件下，类别是y的概率。

而贝叶斯学派就认为，θ是一个随机变量，最大后验概率估计的模型是：

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)p(\theta).$$

也就是认为，是先采样生成θ以后，再在x和θ的情况下，类别是y的概率。

二者（逻辑回归与MAp）都是通过极大似然来找到合适的θ，为什么说贝叶斯最大后验概率估计就能跟好的克服过拟合问题呢?