Linear Model

Fish

2019 年 7 月 9 日

Content

1 What

define the example x described by d property, $x = (x_1; x_2; ...; x_d)$, among the variable x, x_i is decided by ith property of x.

linear model showed as follow:

$$f(x) = w_1x_1 + w_2x_2 + ... + w_dx_d + b \tag{1}$$

whether x has d property

vector as well:

$$f(x) = w^Tx + b \tag{2}$$

2 one sample has one property

to simplify the situation, we assump that the input x propert has only one.

2.1 background

we have dataset: $D = (x_1, y_1)), (x_2, y_2), ..., (x_m, y_m)$, to study a model predict tab of output more correctly.

2.2 analysis

to study model: $f(x) = w^Tx + b$, make the predict result $f(x_i) \cong y_i$, the w, b need to be known.

mean square error is correspondinng to Euclidean distance. The way solve model based on mean square error named least square method. according to linear regression, least square method try to find a line that make the Euclidean distance sum of sample to line be least.

2.3 processing

find least value of equation 3 and 4, get the w^*, b^*

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (f(x_i) - y_i)^2 \tag{3}$$

$$= \arg\min_{(w,b)} \sum_{i=1}^{m} (wx_i + b - y_i)^2 \tag{4}$$

the progress solving the above equation named least square parameter estimation. there differenetiate w and b, could get following part:

$$\frac{\partial E(w,b)}{\partial w} = \sum_{i=1}^{m} 2(-y_i + b + wx_i)x_i = 2\sum_{i=1}^{m}[wx_i^2 - (y_i - b)x_i] \tag{5}$$

$$\frac{\partial E(w,b)}{\partial b} = \sum_{i=1}^{m} 2(-y_i + b + wx_i) = 2[mb - \sum_{i=1}^{m}(y_i - wx_i)] \tag{6}$$

then optimizer solution fo closed-form is shown:

$$w = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}(\sum_{i=1}^{m} x_i)^2} \tag{7}$$

$$b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i) \tag{8}$$

3 one sample has multiply property

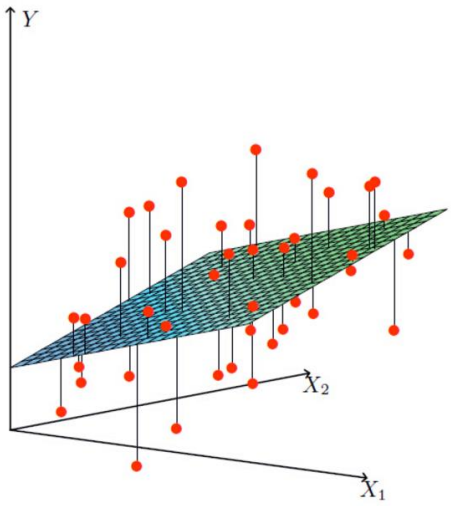


Fig 1: fitting plane in sample space

the common condition is as section one show, dataset D has d property. at present, it should try to study

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b \tag{9}$$

let

$$f(\boldsymbol{x}_i) \cong y_i \tag{10}$$

this condition named multivariate linear regression.

3.1 analysis

similarly, estimate w and b by least square method. conveniently,

$$\hat{\boldsymbol{w}} = (\boldsymbol{w}; b) \tag{11}$$

dataset D show as a matrix $\boldsymbol{X}$, shape $m \times (d+1)$

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{pmatrix} = \begin{pmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{pmatrix} \tag{12}$$

$$\text{tag } \boldsymbol{y} = (y_1; y_2; \cdots; y_m) \tag{13}$$

so fitting plane

$$h_w(x) = w_0 + w_1 x + w_2 x + ... + w_d x, w_0 \text{ is bias} \tag{14}$$

in summary,

$$h_w(x) = \sum_{i=0}^{n} w_i x_i = w^T x \tag{15}$$

3.2 processing

$\boldsymbol{\varepsilon}$ error, represent the difference between true and predict value.

for every sample,

$$y^i = w^T x^i + \varepsilon^i \tag{16}$$

whether ε obey the independent and identical distribution, and observe mean 0 and var θ^2 distribution.

ε obey gause distribution,

$$p(\varepsilon^i) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\varepsilon^i)^2}{2\sigma^2}) \tag{17}$$

substitution equation 16 into equation 17, as following:

$$p(y^i | x^i; w) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{18}$$

注意, 这里分号后面的参数 w 仅表示 p 依赖于 w 的值, w 并不是 p 的前置条件, 而只是这个概率分布的一个参数而已, 也可以省略分号后的内容

likelihood function:

$$L(w) = \prod_{i=1}^{m} p(y^i|x^i; w) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{19}$$

logarithmic likelihood function:

$$\log L(w) = \log \prod_{i=1}^{m} p(y^i|x^i; w) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \tag{20}$$

expand equation 20th:

$$\begin{aligned} L(w) &= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^i - w^T x^i)^2}{2\sigma^2}) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^i - w^T x^i)^2 \end{aligned} \tag{21}$$

target: $L(w)$ more big and more better

$$J(w) = \frac{1}{2} \sum_{i=1}^{m} (y^i - w^T x^i)^2 \quad \text{(least square method)} \tag{22}$$

target function:

$$\begin{aligned} J(w) &= \frac{1}{2} \sum_{i=1}^{m} (h_w(x^i) - y^i)^2 \\ &= \frac{1}{2} (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})^T (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y}) \end{aligned} \tag{23}$$

solve partial derivation:

$$\begin{aligned} \bigtriangledown_w J(w) &= \bigtriangledown_w (\frac{1}{2} (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})^T (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})) \\ &= \bigtriangledown_w (\frac{1}{2} (\hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} - \boldsymbol{y}^{\boldsymbol{T}}) (\boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{y})) \\ &= \bigtriangledown_w ((\frac{1}{2} (\hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\boldsymbol{w} - \hat{\boldsymbol{w}}^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} - \boldsymbol{y}\boldsymbol{X}\hat{\boldsymbol{w}} + \boldsymbol{y}^{\boldsymbol{T}} \boldsymbol{y})) \\ &= \frac{1}{2} (2\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y}) \\ &= \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X}\hat{\boldsymbol{w}} - \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} \end{aligned} \tag{24}$$

let partial derivation be equal to 0:

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-\boldsymbol{1}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y} \tag{25}$$

estimate method: original estimate part

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{m} (y_i - \bar{y}_i)^2} \tag{26}$$

when R^2 value is very close approximation to 1, the model fits dataset D well.

3.3 deep analysis

if X^TX is irreversible or avoid over fitting, increase λ destabilization

$$\theta = (X^TX + \lambda I)^{-1}X^Ty \tag{27}$$

to remember conclusion simply,

$$X\theta = y \Rightarrow X^TX\theta = X^Ty \Rightarrow \theta = (X^TX)^{-1}X^Ty \tag{28}$$

after add λ destabilization:

- X^TX is positive semidefinite matrix: for any non-zero vector u

$$\theta X^TX\theta = (X\theta)^TX\theta \overset{define\ v=X\theta}{\longrightarrow} v^Tv \geq 0 \tag{29}$$

- for any real $\lambda > 0$, $X^TX + \lambda I$ is positive definite matrix, and is reversible. be sure that regerssion equation must be significative.

$$\theta = (X^TX + \lambda I)^{-1}X^Ty \tag{30}$$

3.4 penalty factor

- the target function of linear regerssion:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \tag{31}$$

- increase square sum loss:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 + \lambda\sum_{j=1}^{m}\theta_j^2 \tag{32}$$

- actually, assume that parameter θ obey Gaussian distribution

3.5 |w|and $|w|^2$

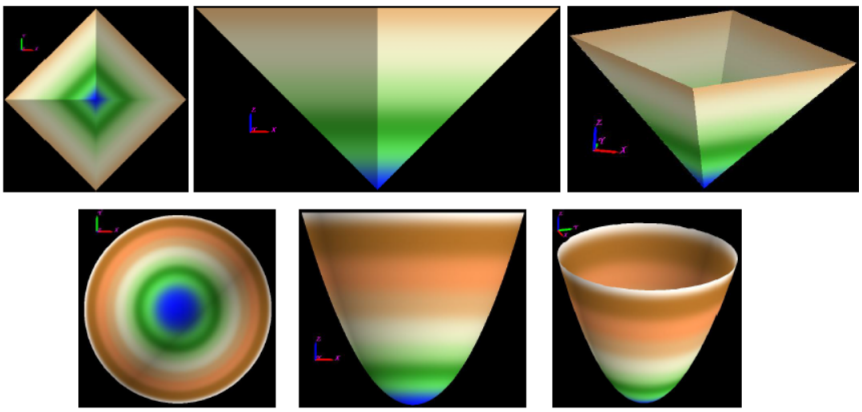


Fig 2: |w| and $|w|^2$

4 machine learning and data analysis

Fig 3: machine learning and data analysis

- cross validation:

  example: 10 folk cross validation

5 conclusion

5.1 probability and statistics

概率（probabilty）和统计（statistics）看似两个相近的概念，其实研究的问题刚好相反。

概率研究的问题是，已知一个模型和参数，怎么去预测这个模型产生的结果的特性（例如均值，方差，协方差等等）。举个例子，我想研究怎么养猪（模型是猪），我选好了想养的品种、喂养方式、猪棚的设计等等（选择参数），我想知道我养出来的猪大概能有多肥，肉质怎么样（预测结果）。

统计研究的问题则相反。统计是，有一堆数据，要利用这堆数据去预测模型和参数。仍以猪为例。现在我买到了一堆肉，通过观察和判断，我确定这是猪肉（这就确定了模型。在实际研究中，也是通过观察数据推测模型是 / 像高斯分布的、指数分布的、拉普拉斯分布的等等），然后，可以进一步研究，判定这猪的品种、这是圈养猪还是跑山猪还是网易猪，等等（推测模型参数）。

一句话总结：概率是已知模型和参数，推数据。统计是已知数据，推模型和参数。

显然，对于最大似然估计，最大后验估计，贝叶斯估计来说，都属于统计的范畴。

5.2theory of MLE

目的: 利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值

原理：极大似然估计是建立在极大似然原理的基础上的一个统计方法，是概率论在统计学中的应用。极大似然估计提供了一种给定观察数据来评估模型参

数的方法，即：“模型已定，参数未知”。通过若干次试验，观察其结果，利用试验结果得到某个参数值能够使样本出现的概率为最大，则称为极大似然估计。

5.3 likelihood and probability

在非正式场合似然和概率 Probability 几乎是一对同义词，但是在统计学中似然和概率却是两个不同的概念。

概率是在特定环境下某件事情发生的可能性，也就是结果没有产生之前依据环境所对应的参数来预测某件事情发生的可能性，比如抛硬币，抛之前我们不知道最后是哪一面朝上，但是根据硬币的性质我们可以推测任何一面朝上的可能性均为 50%，这个概率只有在抛硬币之前才是有意义的，抛完硬币后的结果便是确定的；

而似然刚好相反，是在确定的结果下去推测产生这个结果的可能环境（参数），

还是抛硬币的例子，假设我们随机抛掷一枚硬币 1,000 次，结果 500 次人头朝上，500 次数字朝上（实际情况一般不会这么理想，这里只是举个例子），我们很容易判断这是一枚标准的硬币，两面朝上的概率均为 50%，这个过程就是我们根据结果来判断这个事情本身的性质（参数），也就是似然。

5.4 representation

结果和参数相互对应的时候，似然和概率在数值上是相等的，如果用 θ 表示环境对应的参数，x 表示结果，那么概率可以表示为：

$$P(x|\theta)$$

是条件概率的表示方法，θ 是前置条件，理解为在 θ 的前提下，事件 x 发生的概率，相对应的似然可以表示为：

$$L(\theta|x)$$

理解为已知结果为 x ，参数为 θ (似然函数里 θ 是变量，这里说的参数是相对与概率而言的) 对应的概率，即：

$$L(\theta|x) = P(x|\theta)$$

需要说明的是两者在数值上相等，但是意义并不相同，

$$L(\theta|x)$$

是关于 θ 的函数，而 P 则是关于 x 的函数，两者从不同的角度描述一件事情。

6 least square method vs maximum likelihood estimate

6.1 difference

对于最小二乘法，当从模型总体随机抽取 n 组样本观测值后，最合理的参数估计量应该使模型最好地拟合样本数据，也就是使估计值和观测值之差的平方和最小。而对于最大似然法，当从模型中随机抽取 n 组样本观测值之后，最合理的参数估计量应该使得从模型中抽取的该 n 组样本观测值的概率最大。显然，这是从不同的原理出发的两种参数估计方法。

6.2 the similarity

在最大似然法中，通过选择参数，使已知数据在某种意义下最有可能出现，而某种意义通常指似然函数最大，而似然函数又往往指数据的概率分布函数。与最小二乘法不同的是，最大似然法需要已知这个概率分布函数，这在实践中是很困难的。一般假设其满足正态分布函数的特性，在这种情况下，最大似然估计和最小二乘估计相同. 最小二乘法以估计值与观测值的差的平方和作为损失函数，极大似然法则是以最大化目标值的似然概率函数为目标函数，从概率统计的角度处理线性回归并在似然概率函数为高斯函数的假设下同最小二乘建立了的联系

7 MAP

7.1 concept

最大似然估计是求参数 θ, 使似然函数 $P(x_0|\theta)$ 最大。最大后验概率估计则是想求 θ 使 $P(x_0|\theta)P(\theta)$ 最大。求得的 θ 不单单让似然函数大，θ 自己出现的先验概率也得大。(这有点像正则化里加惩罚项的思想，不过正则化里是利用加法，而 MAP 里是利用乘法)

7.2 theory

MAP 其实是在最大化 $P(\theta|x_0) = \frac{P(x_0\theta)P(\theta)}{P(x_0)}$ ，不过因为 x_0 是确定的（即投出的“反正正正正反正正正反”），$P(x_0)$ 是一个已知值，所以去掉了分母 $P(x_0)$ (假设“投 10 次硬币”是一次实验，实验做了 1000 次，“反正正正正反正正正反”出现了 n 次，则 $P(x_0) = n/1000$。总之，这是一个可以由数据集得到的值)。最大化 $P(\theta|x_0)$ 的意义也很明确，x_0 已经出现了，要求 θ 取什么值使 $P(\theta|x_0)$ 最大。顺带一提，$P(\theta|x_0)$ 即后验概率，这就是“最大后验概率估计”名字的由来。

7.3 $P(\theta)$ value

$P(\theta)$ 实际中是用的 θ 的概率密度函数描述的。因为在式子里面它使用的在 θ 这个点处的概率，但是由于 θ 可以看作是一个连续随机变量，所以某个点处的

概率可能直接无法得到一个函数来描述。但是利用概率密度函数可以达到同样的效果，因为每个点出的概率可以看作是概率函数每个点对应的小矩形条的面积，所以其实和乘上每个 θ 值处的概率效果是一样

7.4 analysis

对于投硬币的例子来看，我们认为 (先验地知道) θ 取 0.5 的概率很大，取其他值的概率小一些。我们用一个高斯分布来具体描述我们掌握的这个先验知识，例如假设 $P(\theta)$ 为均值 0.5，方差 0.1 的高斯函数，如下图：

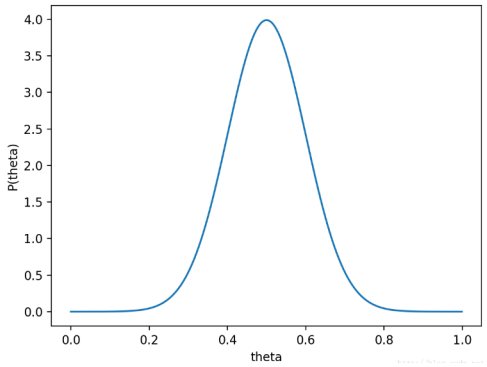


Fig 4: gussian distribution

则 $P(x_0|\theta)P(\theta)$ 的函数图像为 $((1-\theta)^3 * \theta^7 * e^{(-100(\theta-0.5)^2)})$

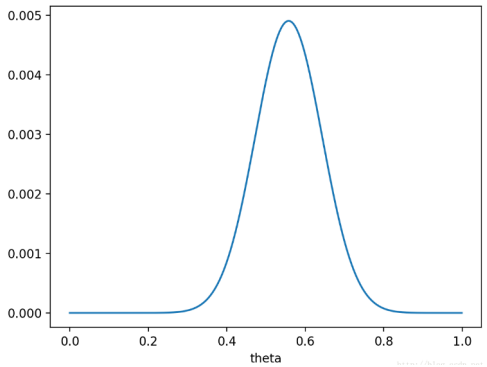


Fig 5: MAP gussian distribution

注意，此时函数取最大值时，θ 取值已向左偏移，不再是 0.7。实际上，在 $\theta = 0.558$ 时函数取得了最大值。即，用最大后验概率估计，得到 $\theta = 0.558$
最后，那要怎样才能说服一个贝叶斯派相信 $\theta = 0.7$ 呢？你得多做点实验。
如果做了 1000 次实验，其中 700 次都是正面向上，这时似然函数为:

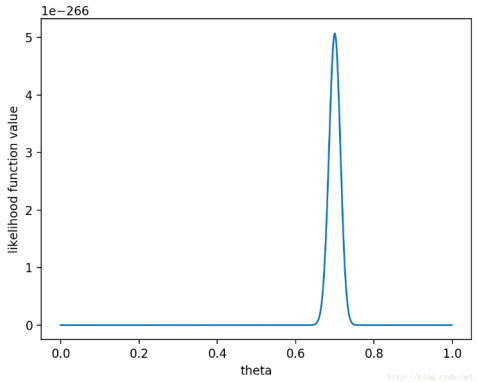


Fig 6: likelihood function gussian distribution

如果仍然假设 $P(\theta)$ 为均值 0.5，方差 0.1 的高斯函数，$P(x_0|\theta)P(\theta)$ 的函数图像为：

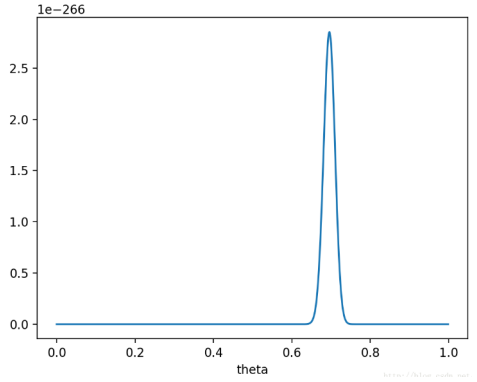


Fig 7: big sample MAP gussian distribution

在 $\theta = 0.696$ 处，$P(x_0|\theta)P(\theta)$ 取得最大值。

这样，就算一个考虑了先验概率的贝叶斯派，也不得不承认得把 θ 估计在 0.7 附近了。

PS. 要是遇上了顽固的贝叶斯派，认为 $P(\theta = 0.5) = 1$ ，那就没得玩了。。无论怎么做实验，使用 MAP 估计出来都是 $\theta = 0.5$ 。这也说明，一个合理的先验概率假设是很重要的。(通常，先验概率能从数据中直接分析得到)

8 the difference between MLE and MAP

- MAP 就是多了一个作为因子的先验概率 $P(\theta)$
- MLE 可以看做 MAP 的一个特例，它把先验概率分布看做均匀分布，所以 $P(\theta)$ 是常数，因此 $P(\theta)$ 就不用出现在目标函数里了。
- 可以认为 $P(\theta)$ 是个均匀分布，处处相等。均匀分布的物理意义是，我们对 θ 的分布没有任何先验的知识

9 理清 MLE 和 Bayes 中 $L(\theta|D), P(D|\theta), P(\theta|D), P(D), P(\theta)$

1. 似然函数: $L(\theta|D)$
2. 其定义表示根据给定数据，找到一个概率最大 (即使数据发生可能最大) 的参数 $L(\theta|D) = P(D|\theta) = P(x_1, x_2, x_3..., x_n|\theta) = \prod_{k=1}^{n}(x_k|\theta)$ (此处假设样本之间相互独立)
3. $P(\theta|D)$ 表示后验概率，指掌握了一定量的数据后我们的参数分布是怎么样的
4. $P(\theta)$ 表示先验概率，指在没有掌握数据后我们的参数怎么分布
5. $P(D)$ 为数据分布
6. $P(\theta)$ 为先验概率

10 the blob about MLE and Bayed

https://blog.csdn.net/liu1194397014/article/details/52766760

序言

本序言是对整体思想进行的一个概括。若没有任何了解，可以先跳过，最后回来看看；若已有了解，可以作为指导思想。
极大似然估计与贝叶斯估计是统计中两种对模型的参数确定的方法，两种参数估计方法使用不同的思想。前者来自于频率派，认为参数是固定的，我们要做的事情就是根据已经掌握的数据来估计这个参数；而后者属于贝叶斯派，认为参数也是服从某种概率分布的，已有的数据只是在这种参数的分布下产生的。所以，直观理解上，极大似然估计就是假设一个参数 θ，然后根据数据来求出这个θ. 而贝叶斯估计的难点在于$p(\theta)$ 需要人为设定，之后再考虑结合MAP （maximum a posterior）方法来求一个具体的θ.
所以极大似然估计与贝叶斯估计最大的不同就在于是否考虑了先验，而两者适用范围也变成了：极大似然估计适用于数据大量，估计的参数能够较好的反映实际情况；而贝叶斯估计则在数据量较少或者比较稀疏的情况下，考虑先验来提升准确率。

预知识

为了更好的讨论，本节会先给出我们要解决的问题，然后给出一个实际的案例。这节不会具体涉及到极大似然估计和贝叶斯估计的细节，但是会提出问题和实例，便于后续方法理解。

问题前提

首先，我们有一堆数据$D=\{x_1,x_2,\ldots,x_n\}$，当然这些数据肯定不是随便产生的，我们就假设这些数据是以含有未知参数θ 某种概率形式（如Bernoulli分布即0-1分布）分布的。我们的任务就是通过已有的数据，来估计这个未知参数θ。估计这个参数的好处就在于，我们可以对外来的数据进行预测。

问题实例

假设一个抛硬币实验，我们之前不知道这些硬币是不是正反均匀的，也许硬币正反不等，假设正面向上设为1的概率为ρ，反面向上设为0为（$1-\rho$）. 我们进行了3次实验，得到两次正面，一次反面，即序列为$'110'$。这里，$D=(1,1,0)$，$\theta=\rho$。

符号说明

这里给出一些符号表示。可看到不理解时过来查看。

符号	含义
D	已有的数据(data)
θ	要估计的参数(parameter)
$p(\theta)$	先验概率(prior)
$p(\theta\|D)$	后验概率(posterior)
$p(D)$	数据分布(evidence)
$p(D\|\theta)$	似然函数(likelihood of θ w.r.t. D)
$p(x,\theta\|D)$	已知数据条件下的x,θ概率

方法介绍

这一节将会详细阐明极大似然估计和贝叶斯估计，要注意到两种方法在面对未知参数θ 时采用的不同态度。

模型推导

极大似然估计法认为参数是固有的，但是可能由于一些外界噪声的干扰，使数据看起来不是完全由参数决定的。没关系，数学家们觉得，虽然有误差存在，**但只要让在这个数据给定的情况下，找到一个概率最大的参数就可以了。**那问题其实就变成了一个条件概率最大的求解，即求使得$p(\theta|D)$ 最大的参数θ，形式化表达为求解

$$\arg\max_{\theta} p(\theta|D) \tag{1}$$

而根据条件概率公式有

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{2}$$

因为我们在极大似然估计中假设 θ 是确定的，所以$p(\theta)$就是一个常数。$p(D)$ 同样是根据已有的数据得到的，也是确定的，或者我们可以把其看作是对整个概率的一个归一化因子。这时候，求解公式(1) 就变成了求解

$$\arg\max_{\theta} p(D|\theta) \tag{3}$$

的问题。
(3) 式中的$p(D|\theta)$ 就是似然函数，我们要做的就是求一个是似然最大的参数，所以称为极大似然估计。
想求解这个问题，需要假设我们的数据是相互独立的。$D = \{x_1, x_2, x_3, \ldots, x_n\}$，这时候有

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta), \tag{4}$$

一般对(4)式取对数求解对数极大似然，就可以把连乘变成求和，然后求导取极值点就是要求的参数值，不在此赘述。

实例

为了便于理解，我们以之前的抛硬币实验作为实例。
回到当时我们一开始抛硬币实验，$D = (1,1,0)$, $\theta = \rho$ 的话，我们可以得到

$$\begin{aligned} p(D|\theta) &= p(x_1|\rho)p(x_2|\rho)p(x_3|\rho) \\ &= p(1|\rho)p(1|\rho)p(0|\rho) \\ &= \rho * \rho * (1-\rho) \end{aligned} \tag{5}$$

然后使用对数极大似然估计就可以得到参数ρ 的值了。

贝叶斯估计

考虑到这节对先验概率(prior)这个概念用的次数比较多，我们首先先介绍先验与后验概率是什么，怎么得到；其次会介绍贝叶斯估计模型的推导过程；最后会举一个例子来加深理解。

先验概率、后验概率

先验概率(prior)与后验概率(posterior)简称为**先验**和**后验**。这两个概念其实是来自于贝叶斯定理，相信学过概率论的一定有所了解。在此试作简单介绍。
之前提到的先验概率到底是什么呢？，毫无疑问必须得与放在一起来介绍。一个先一个后，我们肯定是针对同一个事物才有先后之分，如果针对两个事物，先后不是没有意义了么？那这个共同的对象，就是我们的参数θ。后验概率是指掌握了一定量的数据后我们的参数分布是怎么样的，表示为$p(\theta|D)$；那先验就是在没有掌握数据后我们的参数怎么分布。

看到这里，你可能会问：如果连数据都没有，我怎么知道我的参数是怎么分布的？你提出这个问题，就说明你是一个赤裸裸的频率派学家，你需要通过数据来得到你的参数！而这并不是贝叶斯派的考虑，贝叶斯估计最重要的就是那个先验的获得。虽然你这次的一组数据，比如说扔三次硬币产生的序列是（110）这样分布的，但是其实我根据我历史的经验来看，一枚硬币正反面其实很有可能是按照均匀分布来的，只不过可能因为你**抛得次数少**了所以产生了不是均匀分布的效果。所以我要考虑我以往的经验在里面。

你可能又会问：那你这个均匀分布不就是完全猜来的嘛，你怎么知道我这次是不是一样的硬币呢？没错！就是**“猜来的”**。先验在很多时候完全是假设，然后去验证有的数据是否吻合先验猜想，所以这里的猜很重要。还要注意，先验一定是与数据无关的，你不能看到了数据再做这些猜想，一定是**没有任何数据之前**你就猜了一个参数的先验概率。

有个这部分知识，我们可以开始推导贝叶斯估计模型了。

模型推导

还是继续上面的模型，注意公式(2)其实是一个很概括的模型，既没有对概率形式以及概率参数进行定义，也没有运用到参数固定与否的思想，所以公式(2)同样适用于贝叶斯模型，我们仍然想对该式进行处理得出我们的贝叶斯估计方法。照抄下来(2)式为

$$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{p(D)}.$$

此时，这里面除了分母可以看作是一个归一化因子外，其余均是概率分布的函数。也就是说，无法再像极大似然估计那样将先验概率($p(\theta)$)看作一个常量。这时候就需要考虑用到我们的先验概率了。我们这次把分母也展开来看看，根据全概率公式[1]得到

$$p(D)=\int_\theta p(D|\theta)p(\theta)d\theta. \tag{6}$$

我们来把这个式子(4)

$$p(D|\theta)=\prod_{i=1}^n p(x_i|\theta)$$

和式子(6)一起带入(2)式，得到

$$p(\theta|D)=\frac{(\prod_{i=1}^n p(x_i|\theta))p(\theta)}{\int_\theta(\prod_{i=1}^n p(x_i|\theta))p(\theta)d\theta} \tag{7}$$

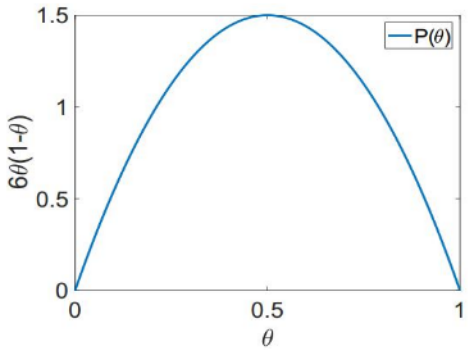
至此，我们就完成了对贝叶斯估计模型的推到过程。有人会问，怎么就完成了？还有那么长一段公式，我们怎么计算啊？其实仔细看看(7)式，其实这些符号我们都是知道的，我们就通过下面的实例来详述。

实例

式(7)中的符号有先验，根据之前对先验的介绍，这是在没有数据之前我们就已经知道的函数了。知道是什么意思？不妨还是在那个抛硬币试验中，我们假设这个$\theta(\rho)$的先验概率是服从

$$f_\rho(\rho) = 6\rho(1-\rho) \tag{8}$$

概率分布的。如图



然后$(\prod_{i=1}^{n} p(x_i|\theta))$也已经知道是$\rho * \rho * (1-\rho)$了。这时要的事情，其实就是把所有已知的全都一股脑带进去就可以了。有人问，已知概率分布怎么知道概率，我想这个问题，可以去概率论的书上找找。

但是，其实做到这一步，我们会发现虽然解决了问题，但是又会带来新的问题，因为在解决这一类贝叶斯估计的问题的时候，我们让参数以某种概率密度函数分布，就会导致在计算过程中不可避免的高复杂度，人们为了计算上的方便，就提出不再是把所有的后验概率$p(\theta|D)$都找出来，而是仍然采用类似于极大似然估计的思想，来**极大后验概率(Maximum A Posterior)**，得到这种简单有效的叫做**MAP**（前面英文的首字母）的算法。下面我们再一步步介绍一下**MAP**。

极大后验概率(MAP)

虽然本节独自成为一节，但是其实是隶属于贝叶斯估计的，属于贝叶斯估计里面的一个trick，放弃一点的准确性，来极大提升算法性能。所以，这个部分不能算是模型，只能算是算法。
MAP（Maximum A Posterior）的理论依据是绝大部分情况下，参数值最有可能出现在概率最大点附近。为了说清楚MAP的来龙去脉，本节将首先介绍如何利用贝叶斯估计的参数进行预测，然后分析直接使用之前得到的后验概率有什么不好，最后介绍MAP算法做的工作。

使用贝叶斯估计的参数做预测

前一节中，我们通过贝叶斯估计得到了后验概率$p(\theta|D)$。那么这个后验概率能用来做什么呢？当然，就比如我们一直在说的那个例子，得到了数据$D=(110)$，还想预测第四次得到的结果什么是什么怎么办？我们当然就需要计算$p(1|D)$和$p(0|D)$看看谁大谁小，哪个更有可能发生。这里，为了泛化，我们将问题再次形式化一下为

已知数据$D=(x_1,x_2,\dots,x_n)$，预测新的数据x的值。

这个问题还有很多细节，比如先验概率，后验概率，数据分布等一些细节，因为前面已经介绍过了，这里为了突出重点，不再重复。在此需要关注的是，所谓预测新的数据的值，其实就是能够在**已知数据D的情况下，找到数据的数学期望**[2]。即求

$$E(x|D)=\int_x xp(x|D)dx. \tag{9}$$

也就是我们需要求$p(x|D)$，这该怎么办？其实这个式子比较迷惑人的点就在于，它内藏了一个参数，也就是x的分布其实与参数是有关的，但是又参数θ是服从某种概率分布的，要对参数所有可能的情况都考虑就得到了

$$p(x|D)=\int_\theta p(x,\theta|D)d\theta \tag{10}$$

这一式子。
接下来还是运用基本的条件概率公式

$$p(x,\theta|D)=p(x|\theta,D)p(\theta|D). \tag{11}$$

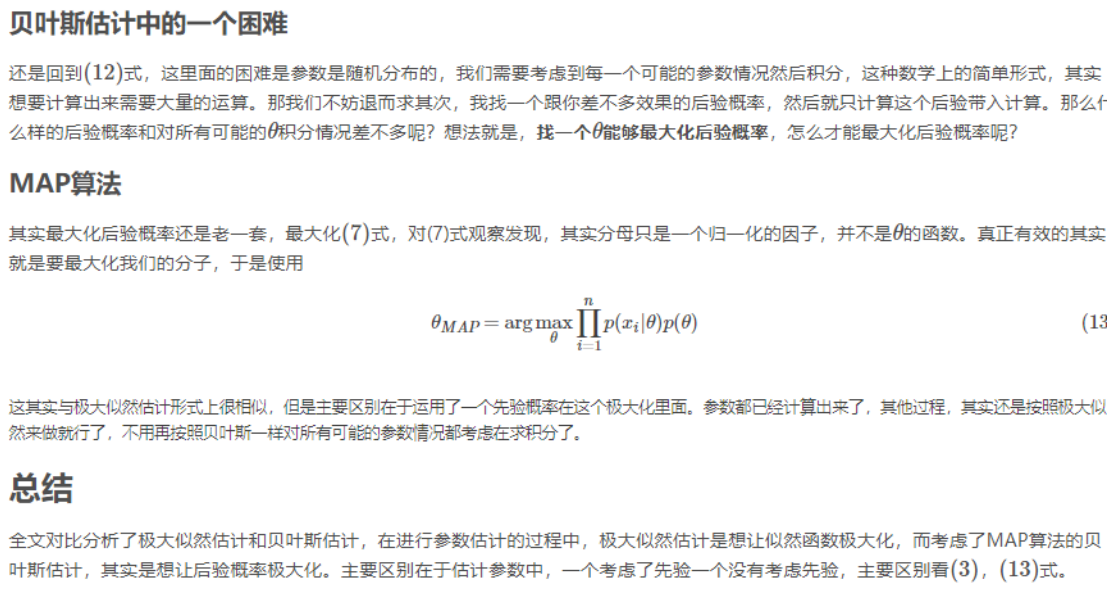
对这一句公式的解释就是，x和θ在已知数据D的条件下的概率，等于x在已知θ和数据D的条件下的概率乘θ在已知数据D的条件下的概率。为什么我要费这个心来说这个，一方面是我为了方便大家理解这个多维条件概率符号的含义，另一方面更重要的是右边式子的第一项$p(x|\theta,D)$可这样

$$p(x|\theta,D)=p(x|\theta)$$

化简。为什么？因为我们从数据里面得到的东西对一个新的数据来说，其实只是那些参数，所以对x而言，θ就是D，两者是同一条件。
那么(10)式就变成了[3]

$$p(x|D)=\int_\theta p(x,\theta|D)d\theta=\int_\theta p(x|\theta)p(\theta|D)d\theta. \tag{12}$$

$p(x|\theta)$是已知的(*例如在我们的问题里面可以是$p(1|\rho)$或者$p(0|\rho)$*)；$p(\theta|D)$也是已知的，我们在贝叶斯估计中已经通过(7)式求出来了。所以这个式子完全就是一个只含有x的函数，带入(9)式完全可以计算出来数学期望。但是！这里面我忽略了一个事实，这里面存在什么困难呢？下面会帮助大家分析。



贝叶斯估计中的一个困难

还是回到(12)式，这里面的困难是参数是随机分布的，我们需要考虑到每一个可能的参数情况然后积分，这种数学上的简单形式，其实想要计算出来需要大量的运算。那我们不妨退而求其次，我找一个跟你差不多效果的后验概率，然后就只计算这个后验带入计算。那么什么样的后验概率和对所有可能的θ积分情况差不多呢？想法就是，**找一个θ能够最大化后验概率**，怎么才能最大化后验概率呢？

MAP算法

其实最大化后验概率还是老一套，最大化(7)式，对(7)式观察发现，其实分母只是一个归一化的因子，并不是θ的函数。真正有效的其实就是要最大化我们的分子，于是使用

$$\theta_{MAP} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)p(\theta) \tag{13}$$

这其实与极大似然估计形式上很相似，但是主要区别在于运用了一个先验概率在这个极大化里面。参数都已经计算出来了，其他过程，其实还是按照极大似然来做就行了，不用再按照贝叶斯一样对所有可能的参数情况都考虑在求积分了。

总结

全文对比分析了极大似然估计和贝叶斯估计，在进行参数估计的过程中，极大似然估计是想让似然函数极大化，而考虑了MAP算法的贝叶斯估计，其实是想让后验概率极大化。主要区别在于估计参数中，一个考虑了先验一个没有考虑先验，主要区别看(3)，(13)式。

Fig 8: 极大似然估计与贝叶斯估计

11 reference

11.1 the relationship and difference between MLE、MAP、Bayse estimate

https://blog.csdn.net/bitcarmanlee/article/details/81417151

11.2 detail of MLE and MAP

https://www.cnblogs.com/sylvanas2012/p/5058065.html

11.3 least square and gradient descend and maximum likelihood

https://blog.csdn.net/FrankieHello/article/details/81432769

11.4 detail of MLE

https://blog.csdn.net/u014182497/article/details/82252456

11.5 understand deeply MLE and MAP

https://blog.csdn.net/u011508640/article/details/72815981

11.6 difference between Bayes and MLE

https://blog.csdn.net/feilong_csdn/article/details/61633180