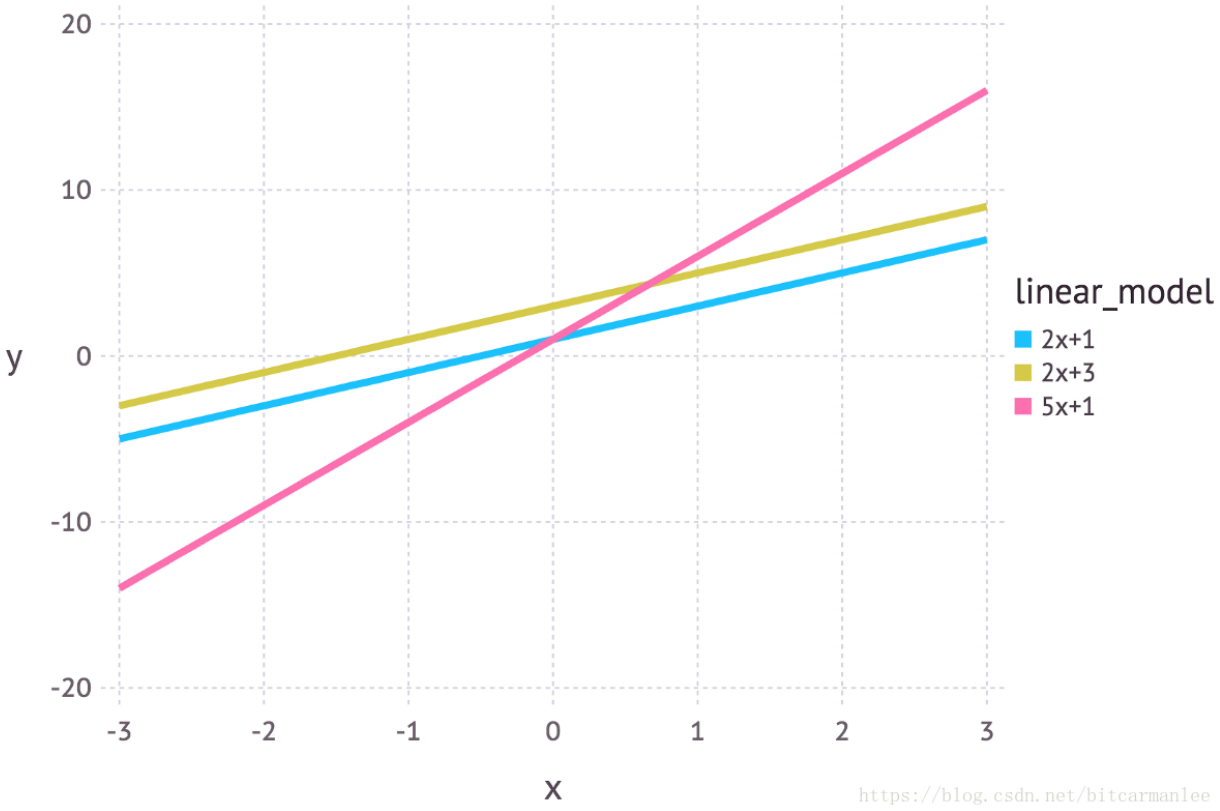
1.什么是参数

在机器学习中，我们经常使用一个模型来描述生成观察数据的过程。例如，我们可以使用一个随机森林模型来分类客户是否会取消订阅服务（称为流失建模），或者我们可以用线性模型根据公司的广告支出来预测公司的收入（这是一个线性回归的例子）。每个模型都包含自己的一组参数，这些参数最终定义了模型本身。

我们可以把线性模型写成 y = mx + c 的形式。在广告预测收入的例子中，x 可以表示广告支出，y 是产生的收入。m 和 c 则是这个模型的参数。这些参数的不同值将在坐标平面上给出不同的直线（见下图）。



2.参数估计的方法

就是根据样本统计量的数值对总体参数进行估计的过程。根据参数估计的性质不同，可以分成两种类型：点估计和区间估计。

点估计就是用样本统计量的某一具体数值直接推断未知的总体参数。例如，在进行有关小学生身高的研究中，随机抽取1000名小学生并计算出他们的平均身高

为1.45米。如果直接用这个1.45米代表所有小学生的平均身高，那么这种估计方法就是点估计。

而对总体参数进行点估计常用的方法有两种：矩估计与最大似然估计，其中最大似然估计就是我们实际中使用非常广泛的一种方法。

按这两种方法对总体参数进行点估计，能够得到相对准确的结果。如用样本均值X估计总体均值，或者用样本标准差S估计总体标准差σ。

但是，点估计有一个不足之处，即这种估计方法不能提供估计参数的估计误差大小。对于一个总体来说，它的总体参数是一个常数值，而它的样本统计量却是随机变量。当用随机变量去估计常数值时，误差是不可避免的，只用一个样本数值去估计总体参数是要冒很大风险的。因为这种误差风险的存在，并且风险的大小还未知，所以，点估计主要为许多定性研究提供一定的参考数据，或在对总体参数要求不精确时使用，而在需要用精确总体参数的数据进行决策时则很少使用。

区间估计就是在推断总体参数时，还要根据统计量的抽样分布特征，估计出总体参数的一个区间，而不是一个数值，并同时给出总体参数落在这一区间的可能性大小，概率的保证。还是举小学生身高的例子，如果用区间估计的方法推断小学生身高，则会给出以下的表达：根据样本数据，估计小学生的平均身高在1.4~1.5米之间，置信程度为95%，这种估计就属于区间估计。

3.概率与统计的区别

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。

4.最大似然估计(maximum likelihood estimates，MLE)

前文提到，最大似然估计(maximum likelihood estimates，MLE)是实际中使用非常广泛的一种方法，用我们老师的一句最简单的话来总结最大似然估计，就是“谁大像谁”。

说到最大似然估计与最大后验估计，最好的例子自然就是抛硬币了。本文也不免俗，同样以抛硬币作为例子。

于是我们拿这枚硬币抛了10次，得到的数据X是：反正正正正反正正正反。我们想求的正面概率θ是模型参数，而抛硬币模型我们可以假设是二项分布。

在概率论和统计学中，二项分布（Binomial distribution）是n个独立的是/非试验中成功的次数的离散概率分布，其中每次试验的成功概率为p。这样的单次成功/失败试验又称为伯努利试验。实际上，当n = 1时，二项分布就是伯努利分布。

伯努利分布（Bernoulli distribution，又名两点分布或者0-1分布，是一个离散型概率分布，为纪念瑞士科学家雅各布·伯努利而命名。)若伯努利试验成功，则伯努利随机变量取值为0。记其成功概率为 $p(0\leq p\leq 1)$,失败概率为 $q=1-p$。

对于伯努利分布来说：

概率质量函数为：

$$f_X(x)=p^x(1-p)^{1-x}=\{pq\text{if } x=1,\text{if } x=0.$$

期望为：

$$E[X]=\sum_{i=01}x_i f_X(x)=0+p=p$$

方差为:

$$var[X]=\sum_{i=01}(x_i-E[X])^2 f_X(x)=(0-p)^2(1-p)+(1-p)^2 p=p(1-p)=pq$$

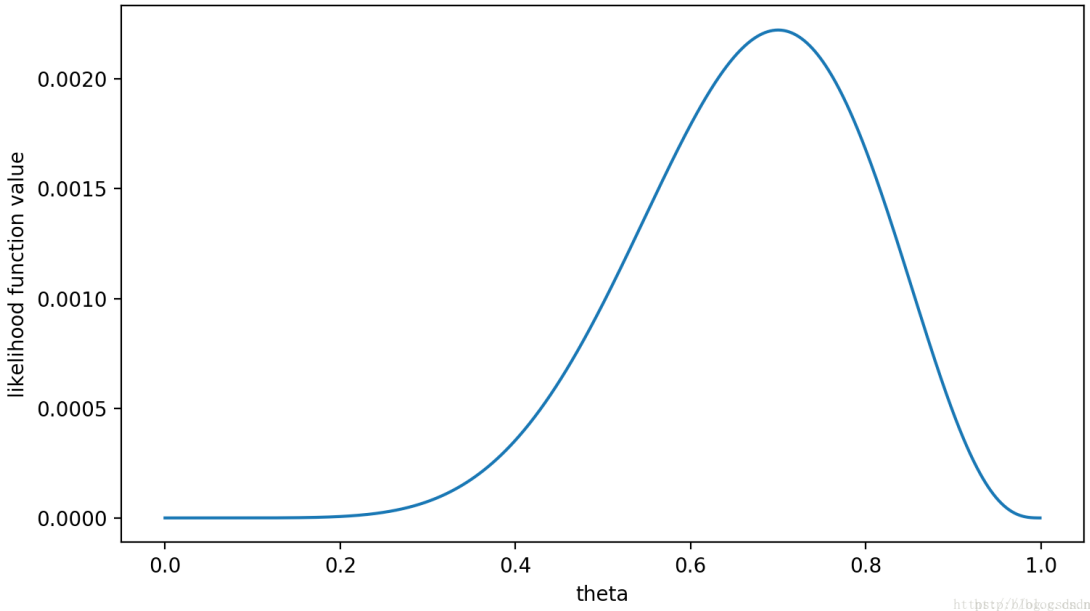
而如果X ~ B(n, p)（也就是说，X是服从二项分布的随机变量）
一般的二项分布是n次独立的伯努利试验的和。它的期望值和方差分别等于每次单独试验的期望值和方差的和：

$$\mu_n=\sum_{k=1n}\mu=np,\sigma^2_n=\sum_{k=1n}\sigma^2=np(1-p).$$

回到抛硬币的例子，出现实验结果X的似然函数是什么呢?

$$f(X,\theta)=\theta^7(1-\theta)^3$$

需要注意的是，上面只是个关于θ的函数。而最大似然估计，很明显是要最大化这个函数。
可以看一下这个函数的图像：



容易得出，在$\theta=0.7$时，似然函数能取到最大值。
当然实际中我们一般不会画图，而是通过更为简洁的数学手段来处理。
首先我们取对数似然函数，这样更方便后续的数学运算：

$$\ln(f(X,\theta))=\ln(\theta^7(1-\theta)^3)=7\ln(\theta)+3\ln(1-\theta)$$

对对数似然函数求导：

$$\ln'(f(X,\theta))=7\theta-31-\theta$$

令导数为0:

$$7(1-\theta)-3\theta=0$$

最终求得：

$$\theta=0.7$$

这样，我们已经完成了对

的最大似然估计。即，抛10次硬币，发现7次硬币正面向上，最大似然估计认为正面向上的概率是0.7。是不是非常直接，非常简单粗暴？没错，就是这样，谁大像谁！

说到这里为止，可能很多同学不以为然：你这不坑爹嘛？只要硬币一枚正常硬币，不存在作弊情况，正面朝上的概率必然为0.5么，你这怎么就忽悠我们是0.7呢。OK，如果你这么想，恭喜你，那你就天然包含了贝叶斯学派的思想！我们所谓的正常硬币向上的概率为0.5，就是贝叶斯里的先验概率。

5.最大后验估计(maximum a posteriori estimation)

上面的最大似然估计MLE其实就是求一组能够使似然函数最大的参数，即

$$\theta\hat{}_{ML}(x)=\text{argmax}_\theta f(x|\theta)$$

如果我们把问题稍微弄复杂一点，如果这个参数θ有一个先验概率呢？比如上面的例子中，实际生活经验告诉我们，硬币一般都是均匀的，也就是$\theta=0.5$的概率最大，那么这个参数该怎么估计？
这个时候就用到了我们的最大后验概率MAP。MAP的基础是贝叶斯公式：

$$p(\theta|x)=p(x|\theta)\times p(\theta)P(x)$$

其中，$p(x|\theta)$就是之前讲的似然函数，$p(\theta)$是先验概率，是指在没有任何实验数据的时候对参数 θ的经验判断，对于一个硬币，大概率认为他是正常的，正面的概率为0.5的可能性最大。

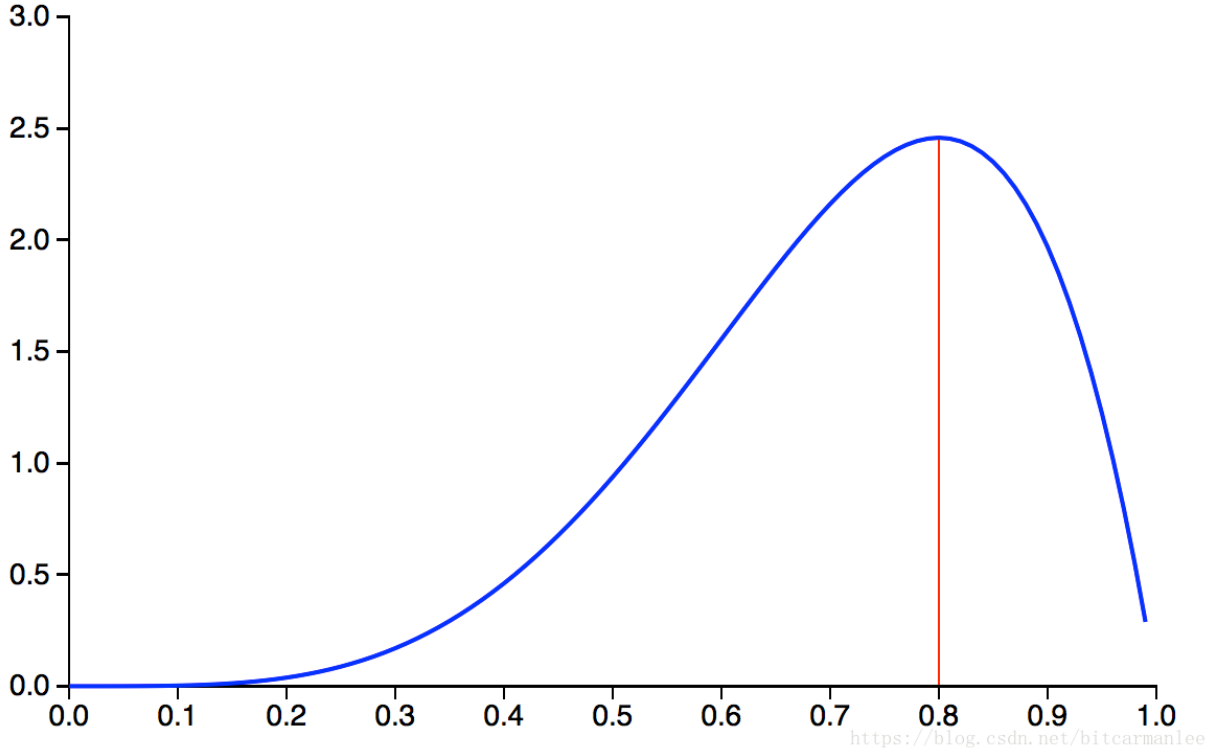
MAP优化的就是一个后验概率，即给定了观测值以后使后验概率最大：

$$\theta\hat{}_{MAP}=\text{argmax}_\theta p(\theta|x)=\text{argmax}_\theta p(x|\theta)\times p(\theta)P(x)=\text{argmax}_\theta p(x|\theta)\times p(\theta)$$

从上面公式可以看出，$p(x|\theta)$是似然函数，而$p(\theta)$就是先验概率。对其取对数：

$$\text{argmax}_\theta p(x|\theta)\cdot p(\theta)=\text{argmax}_\theta \log\prod_{i=0}^{n} p(x_i|\theta)p(\theta)=\text{argmax}_\theta \sum_i \log(p(x_i|\theta)p(\theta))=\text{argmax}_\theta \sum_i \log(p(x_i|\theta)+\log(p(\theta))$$

通过MAP最终的式子不难看出，MAP就是多个作为因子的先验概率$P(\theta)$。这个$p(\theta)$可以是任何的概率分布，比如高斯分布，比如也可以是β分布。比如$\beta(5,2)$的概率分布图如下：



如果将这个概率分布作为$p(\theta)$，那么我们在还未抛硬币前，便认为θ很可能接近于0.8，而不大可能是个很小的值或是一个很大的值。换言之，我们在抛硬币前，便估计这枚硬币更可能有0.8的概率抛出正面。

那么问题就来了，为什么我们要用β分布来描述先验概率呢？

首先一点，通过调节 Beta 分布中的a和b,你可以让这个概率分布变成各种你想要的形状！Beta 分布已经很足够表达我们事先对θ的估计了。

更重要的一点是，如果使用Beta 分布，会让之后的计算更加方便。因为有如下结论：

$p(\theta)$是个Beta分布，那么在观测到”X = 抛10次硬币出现7次正面”的事件后，$p(\theta|X)$仍然是个Beta分布，只不过此时概率分布的形状因为有了观测事件

而发生了变化！此时有

$$p(\theta|X)=Beta(\theta|a+3,b+2)$$

换句话说，数据观测前后，对θ的估计的概率分布均为 Beta 分布，这就是为什么使用 Beta 分布方便我们计算的原因。当我们得知$p(\theta|X)=Beta(\theta|a+3,b+2)$后，只要根据 Beta 分布的特性，得出θ最有可能等于多少了。即θ等于多少时，观测后得到的 Beta 分布有最大的概率密度）。

到此为止，我们可以得到“共轭性”的真正含义了！后验概率分布（正比于先验和似然函数的乘积）拥有与先验分布相同的函数形式。这个性质被叫做共轭性（Conjugacy）。共轭先验（conjugate prior）有着很重要的作用。它使得后验概率分布的函数形式与先验概率相同，因此使得贝叶斯分析得到了极大的简化。例如，二项分布的参数之共轭先验就是我们前面介绍的 Beta 分布。多项式分布的参数之共轭先验则是 Dirichlet 分布，而高斯分布的均值之共轭先验是另一个高斯分布。

总的来说，对于给定的概率分布$p(X|\theta)$，我们可以寻求一个与该似然函数$p(X|\theta)$共轭的先验分布$p(\theta)$，如此一来后验分布$p(\theta|X)$就会同先验分布具有相同的函数形式。而且对于任何指数族成员来说，都存在有一个共轭先验。

6.贝叶斯估计

贝叶斯估计是在MAP上做进一步拓展，此时不直接估计参数的值，而是允许参数服从一定概率分布。回忆下贝叶斯公式：

$$p(\theta|x)=p(x|\theta)\times p(\theta)P(x)$$

现在我们不要求后验概率最大，这个时候就需要求$p(X)$，即观察到的X的概率。一般来说，用全概率公式可以求$p(X)$

$$p(X)=\int p(X|\theta)p(\theta)d\theta$$

那么如何用贝叶斯估计来预测呢？如果我们想求一个值 x’ 的概率，可以用下面的方法

$$p(\hat{x}|X)=\int_{\theta\in\Theta}p(\hat{x}|\theta)p(\theta|X)d\theta=\int_{\theta\in\Theta}p(\hat{x}|\theta)\frac{p(X|\theta)p(\theta)}{p(X)}d\theta$$

7.什么时候 MAP 估计与最大似然估计相等?

当先验分布均匀之时，MAP 估计与 MLE 相等。直观讲，它表征了最有可能值的任何先验知识的匮乏。在这一情况中，所有权重分配到似然函数，因此当我们把先验与似然相乘，由此得到的后验极其类似于似然。因此，最大似然方法可被看作一种特殊的 MAP。

如果先验认为这个硬币是概率是均匀分布的，被称为无信息先验(non-informative prior)，通俗的说就是“让数据自己说话”，此时贝叶斯方法等同于频率方法。

随着数据的增加，先验的作用越来越弱，数据的作用越来越强，参数的分布会向着最大似然估计靠拢。而且可以证明，最大后验估计的结果是先验和最大似然估计的凸组合。

参考文献：

1.https://blog.csdn.net/baimafujinji/article/details/51374202

2.https://blog.csdn.net/yt71656/article/details/42585873

3.https://www.jiqizhixin.com/articles/2018-01-09-6

4.https://zh.wikipedia.org/zh-hans/二項分佈

5.https://zh.wikipedia.org/wiki/伯努利分布

6.Pattern Recognition And Machine Learning