先验概率，条件概率与后验概率

先验概率是基于背景常识或者历史数据的统计得出的预判概率，一般只包含一个变量，例如 $P(X)$， $P(Y)$ 。

条件概率是表示一个事件发生后另一个事件发生的概率，例如 $P(Y|X)$ 代表 X 事件发生后 Y 事件发生的概率。

后验概率是由果求因，也就是在知道结果的情况下求原因的概率，例如Y事件是X引起的，那么$P(X|Y)$ 就是后 验概率，也可以说它是事件发生后的反向条件概率。

似然函数

在数理统计学中，似然函数是一种关于统计模型中的参数的函数，表示模型参数中的似然性。似然函数可以理解为条件概率的逆反。

在已知某个参数 α 时，事件 A 会发生的条件概率可以写作 $P(A;\alpha)$ ，也就是 $P(A|\alpha)$ 。我们也可以构造似然性的方法来表示事件 A 发生后估计参数 α 的可能性，也就表示为 $L(\alpha|A)$ ，其中 $L(\alpha|A)=P(A|\alpha)$ 。

这里 Wikipedia 的解释比较全面详细，可以参见似然函数。

最大似然估计（MLE ）与最大后验概率（MAP ）

最大似然估计是似然函数最初也是最自然的应用。似然函数取得最大值表示相应的参数能够使得统计模型最为合理。从这样一个想法出发，最大似然估计的做法是：首先选取似然函数（一般是概率密度函数或概率质量函数），整理之后求最大值。实际应用中一般会取似然函数的对数作为求最大值的函数，这样求出的最大值和直接求最大值得到的结果是相同的。似然函数的最大值不一定唯一，也不一定存在。

这里简单的说一下最大后验概率（ MAP），如下面的公式

$$P(\alpha|X)=\frac{P(X|\alpha)P(\alpha)}{P(X)}$$

其中等式左边 $P(\alpha|X)$ 表示的就是后验概率，优化目标即为 $argmax_{\alpha}P(\alpha|X)$ ，即给定了观测值 X 以后使模型参数 α 出现的概率最大。等式右边的分子式 $P(X|\alpha)$ 即为似然函数 $L(\alpha|X)$ ，MAP 考虑了模型参数 α 出现的先验概率 $P(\alpha)$ 。即就算似然概率 $P(X|\alpha)$ 很大，但是 α 出现的可能性很小，也更倾向于不考虑模型参数为 α 。

生成式模型与判别式模型

最后简单说一下生成式模型与判别式模型。

判别式模型学习的目标是条件概率 $P(Y|X)$ 或者是决策函数 $Y=f(X)$ ，其实这两者本质上是相同的。例如 KNN， Decision Tree ，SVM ， CRF 等模型都是判别式模型。

生成式模型学习的是联合概率分布 $P(X,Y)$ ，从而求得条件概率分布 $P(Y|X)$ 。例如 NB，HMM 等模型都是生成式模型。

-
-
-

最大似然估计(Maximum likelihood estimation)

最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"。简单而言，假设我们要统计全国人口的身高，首先假设这个身高服从服从正态分布，但是该分布的均值与方差未知。我们没有人力与物力去统计全国每个人的身高，但是可以通过采样，获取部分人的身高，然后通过最大似然估计来获取上述假设中的正态分布的均值与方差。

最大似然估计中采样需满足一个很重要的假设，就是所有的采样都是独立同分布的。下面我们具体描述一下最大似然估计：

首先，假设$x_1, x_2, \ldots\ldots x_n$为独立同分布的采样，θ为模型参数，f为我们所使用的模型，遵循我们上述的独立同分布假设。参数为θ的模型f产生上述采样可表示为

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \ldots, f(x_n \mid \theta)$$

回到上面的"模型已定，参数未知"的说法，此时，我们已知的为$x_1, x_2, \ldots\ldots x_n$，未知为θ，故似然定义为：

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

在实际应用中常用的是两边取对数，得到公式如下：

$$\ln L(\theta \mid x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) \qquad \hat{\ell} = \frac{1}{n} \ln L$$

其中$\ln L(\theta \mid x_1, \ldots, x_n)$称为对数似然，而$\hat{\ell}$称为平均对数似然。而我们平时所称的最大似然为最大的对数平均似然，即：

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{\arg\max}\, \hat{\ell}(\theta \mid x_1, \ldots, x_n)$$

举个别人博客中的例子，假如有一个罐子，里面有黑白两种颜色的球，数目多少不知，两种颜色的比例也不知。我 们想知道罐中白球和黑球的比例，但我们不能把罐中的球全部拿出来数。现在我们可以每次任意从已经摇匀的罐中拿一个球出来，记录球的颜色，然后把拿出来的球 再放回罐中。这个过程可以重复，我们可以用记录的球的颜色来估计罐中黑白球的比例。假如在前面的一百次重复记录中，有七十次是白球，请问罐中白球所占的比例最有可能是多少？很多人马上就有答案了：70%。而其后的理论支撑是什么呢？

我们假设罐中白球的比例是p，那么黑球的比例就是1-p。因为每抽一个球出来，在记录颜色之后，我们把抽出的球放回了罐中并摇匀，所以每次抽出来的球的颜 色服从同一独立分布。这里我们把一次抽出来球的颜色称为一次抽样。题目中在一百次抽样中，七十次是白球的概率是P(Data | M)，这里Data是所有的数据，M是所给出的模型，表示每次抽出来的球是白色的概率为p。如果第一抽样的结果记为x1，第二抽样的结果记为x2... 那么Data = (x1,x2,…,x100)。这样，

P(Data | M)

= P(x1,x2,…,x100|M)

= P(x1|M)P(x2|M)…P(x100|M)

= p^70(1-p)^30.

那么p在取什么值的时候，P(Data |M)的值最大呢？将p^70(1-p)^30对p求导，并其等于零。

70p^69(1-p)^30-p^70*30(1-p)^29=0。

解方程可以得到p=0.7。

在边界点p=0,1，P(Data|M)=0。所以当p=0.7时，P(Data|M)的值最大。这和我们常识中按抽样中的比例来计算的结果是一样的。

假如我们有一组连续变量的采样值（x1,x2,…,xn），我们知道这组数据服从正态分布，标准差已知。请问这个正态分布的期望值为多少时，产生这个已有数据的概率最大？

P(Data | M) = ?

根据公式

$$L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

可得：

$$L(\theta \mid x_1, \ldots, x_n) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

对μ求导可得 $$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\frac{(\sum_{i=1}^{n}x_i-n\mu)}{\sigma^2}$$,则最大似然估计的结果为μ=(x1+x2+…+xn)/n

由上可知最大似然估计的一般求解过程：

（1） 写出似然函数；

（2） 对似然函数取对数，并整理；

（3） 求导数 ；

（4） 解似然方程

注意：最大似然估计只考虑某个模型能产生某个给定观察序列的概率。而未考虑该模型本身的概率。这点与贝叶斯估计区别。