1.
三要素：模型、策略、方法。
模型：伯努利模型，定义是取值在 0 与 1 的随机变量上的概率分布。
策略：对数损失函数，不过贝叶斯估计使用的是结构风险最小化。
算法：极大似然估计使用的是取经验风险函数的极值，贝叶斯估计使用的算法是求取参数的后验分布，然后计算其期望。

设 P（A=1）=X
极大似然估计：
对于 n 个同分布的随机变量，A1，A2，…，An 取经验风险函数的极值点：

$L(P)=-\sum_{i=1}^{n}\log P(A_i)=-k\log X-(n-k)\log(1-X)$，k 为 Ai=1 的个数

求 L（P）的极小值点，对 X 求导：

L（P）的导数 $=-k/X-(n-k)/(1-X)=0$

解得：$X = k/n$

贝叶斯估计：
假设其先验分布为均匀分布：

先验分布：

$f(\theta)=1$

后验分布：

$f(\theta|A_1,\ldots,A_n)=f(A_1,\ldots,A_n|\theta)\cdot f(\theta)/\int f(A_1,\ldots,A_n)\cdot f(\theta)d\theta$

因为 $f(\theta)=1$，因为我们要取使 θ 出现概率最大的值，即得期望的值，

分母不变，所以只用看分子，

$$\theta = arg\max_{\theta} P(A_1, A_2, \ldots A_n|\theta)P(\theta)$$

$$= arg\max_{\theta} \prod_{i=1}^{n} P(A_i|\theta)P(\theta)$$

$$= arg\max_{\theta} \theta^k (1-\theta)^{n-k}$$

上式中分母与θ无关，所以可忽略，即：

$$f(\theta|A_1,\ldots,A_n) \propto \theta^k(1-\theta)^{(n-k)} = \theta^{k+1-1}(1-\theta)^{n-k+1-1}$$

注意，参数为a,b的Beta分布的概率密度函数如下

$$f(p;a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

因此可以看出，θ的后验分布服从参数为$k+1$和$n-k+1$的Beta分布，即：

$$f(\theta|A_1,\ldots,A_n) = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \theta^{k+1-1}(1-\theta)^{n-k+1-1}$$

因此，上式的期望（即θ的估计值）为：

$$E(\theta) = \frac{k+1}{n+2}$$

Beta 分布求期望为 =

- beta分布的均值是 $\frac{\alpha}{\alpha+\beta}$

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

先验分布 f（X）= 1 的原因:

- 3.4.1 贝叶斯假设
- 所谓参数θ 的无信息先验分布是指除参数θ 的取值范围Θ 和θ 在总体分布中的地位之外，再也不包含θ 的任何信息的先验分布。有人把“不包含θ 的任何信息”这句话理解为对θ 的任何可能值，他都没有偏爱，都是同等无知的。因此很自然地把θ 的取值范围上的“均匀”分布看作θ 的先验分布，即

$$\pi(\theta)=\begin{cases}c,\theta\in\Theta\\0,\theta\notin\Theta\end{cases}$$

www.docin.com

- 其中Θ是θ 的取值范围，c是一个容易确定的常数。这一看法通常被称为贝叶斯假设，又称拉普拉斯(Laplace)先验。

- 使用贝叶斯假设也会遇到一些麻烦，主要是以下二个：
- (1)当θ 为无限区间时，如为$(0,\infty)$ 或$(-\infty,\infty)$ 时，在Θ 上无法定义一个正常的均匀分布。
- (2)贝叶斯假设不满足变换下的不变性。

- 定义3.4.1
  设总体$X\sim p(x|\theta),\theta\in\Theta$ 。若θ 的先验分布$\pi(\theta)$ 满足下列条件：

www.docin.com

- 1. $\pi(\theta)\geq 0$，且 $\int_{\Theta}\pi(\theta)d\theta=\infty$
- 2. 由此决定的后验密度$\pi(\theta|x)$是正常的密度函数，则称$\pi(\theta)$为θ 的广义先验密度，或广义先验分布。

贝叶斯定理:

假设Ⅰ 随机变量$\mathbf{X}$有一个密度函数$\mathbf{p}(\mathbf{x};\ \boldsymbol{\theta})$，其中$\boldsymbol{\theta}$是一个参数，不同的$\boldsymbol{\theta}$对应不同的密度函数，故从贝叶斯观点看，$\mathbf{p}(\mathbf{x};\ \boldsymbol{\theta})$在给定$\boldsymbol{\theta}$后是个条件密度函数，因此记为$\mathbf{p}(\mathbf{x}|\boldsymbol{\theta})$更恰当一些。这个条件密度能提供我们的有关的$\boldsymbol{\theta}$信息就是总体信息。

假设Ⅱ 当给定$\boldsymbol{\theta}$后，从总体$\mathbf{p}(\mathbf{x}|\boldsymbol{\theta})$中随机抽取一个样本$\mathbf{X_1}$，…，$\mathbf{X_n}$，该样本中含有$\boldsymbol{\theta}$的有关信息。这种信息就是样本信息。

www.docin.com

假设Ⅲ 我们对参数$\boldsymbol{\theta}$已经积累了很多资料，经过分析、整理和加工，可以获得一些有关$\boldsymbol{\theta}$的有用信息，这种信息就是先验信息。参数$\boldsymbol{\theta}$不是永远固定在一个值上，而是一个事先不能确定的量。

在贝叶斯统计学中，把以上的三种信息归纳起来的最好形式是在总体分布基础上获得的样本$\mathbf{X_1}$，…，$\mathbf{X_n}$，和参数的联合密度函数

$$p(x_1,\cdots,x_n,\theta)=p(x_1,\cdots,x_n|\theta)\pi(\theta)$$

在这个联合密度函数中。当样本 $X_1,\cdots,X_n$ 给定之后未知的仅是参数θ了，我们关心的是样本给定后，θ的条件密度函数，依据密度的计算公式，容易获得这个条件密度函数

$$\pi(\theta|x_1,\cdots,x_n)=\frac{p(x_1,\cdots,x_n,\theta)}{p(x_1,\cdots,x_n)}$$

$$=\frac{p(x_1,\cdots,x_n|\theta)\pi(\theta)}{\int p(x_1,\cdots,x_n|\theta)\pi(\theta)d\theta}$$

www.docin.com

这就是贝叶斯公式的密度函数形式，$\pi(\theta|x_1,\cdots,x_n)$ 称为Θ 的后验密度函数，或后验分布。而

$$p(x_1,\cdots,x_n)=\int p(x_1,\cdots,x_n|\theta)\pi(\theta)d\theta$$

6

是样本的边际分布，或称样本 $X_1,\cdots,X_n$ 的无条件分布，它的积分区域就是参数θ的取值范围，随具体情况而定。

2.

模型是条件概率分布：$P_\theta(Y|X)$,

损失函数是对数损失函数：$L(Y,P(Y|X)=-\log P(Y|X)$,

经验风险为：

$$
\begin{aligned}
R_{emp}(f) &= \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \\
&= \frac{1}{N}\sum_{i=1}^{N} -logP(y_i|x_i) \\
&= -\frac{1}{N}\sum_{i=1}^{N} logP(y_i|x_i)
\end{aligned}
$$

最小化经验风险，也就是最大化

$\sum_{i=1}^{N} logP(y_i|x_i)$，也就是最大化$\prod_{i=1}^{N} P(y_i|x_i)$（对数的连加就是这个的连乘）这个就是极大似然估计。