1. 条件概率 / 联合概率

联合概率: A和B共同发生的概率

$$P(A,B) = P(A|B)P(B)$$

条件概率： B的概率下，A发生的概率

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

联合概率与条件概率的区别

条件概率和联合概率有着相似之处，它们的结果都需要**由多个条件决定**，但条件概率的条件之间并不是平行关系，而是**层层包含**的关系。

求联合概率的时候我们会说“既满足是女生又满足成绩在 90 分以上的学生概率"

而求条件概率的时候我们会说“满足成绩在 90 分以上的学生在女生中的概率"

链式法则 chain rule

$$P(A,B,C,D) = P(A|B,C,D)P(B,C,D)$$
$$= P(A|B,C,D)P(B|C,D)P(C,D)$$
$$= P(A|B,C,D)P(B|C,D)P(C|D)P(D)$$

2. 全概率公式

> B发生的概率 = 所有条件下B发生概率的和

$$P(B) = P(B|A)P(A) + P(B|C)P(C) + \ldots$$

3. 贝叶斯公式：先验概率 / 后验概率

先验概率：

是指根据以往经验和分析得到的概率。意思是说我们人有一个常识,比如骰子,我们都知道概率是1/6。

后验概率：

事情已经发生，要求这件事情发生的原因是由某个因素引起的可能性的大小。

例子：

某城市发生了一起汽车撞人逃跑事件，该城市只有两种颜色的车，蓝色15%，绿色85%，事发时有一个人在现场看见了，他指证是蓝车。但是根据专家在现场分析,当时那种条件能看正确的可能性是80%。那么,肇事的车是蓝车的概率到底是多少?

先验概率： P(蓝) = 0.15 P(绿) = 0.85

条件概率：P(肇事｜蓝色) = 0.8 P(肇事｜绿色) = 0.2

贝叶斯公式求后验概率（用到了条件概率+全概率转换）：

$$P(\text{蓝色} \mid \text{肇事})=\frac{P(\text{蓝色},\text{肇事})}{P(\text{肇事})}=\frac{P(\text{肇事}|\text{蓝色})*P(\text{蓝色})}{P(\text{肇事} \mid \text{蓝色})*P(\text{蓝色})+P(\text{肇事}|\text{绿色})*P(\text{绿色})}$$

$$=\frac{0.8*0.15}{0.8*0.15+0.2*0.85}=0.41$$

4. 条件独立假设

如果X和Y相对于Z是条件独立的，也就是说，X是否发生与Y是否发生之间毫无关系，则:

$$P(X,Y|Z)=P(X|Z)P(Y|Z)$$

$$P(X|Y,Z)=P(X|Z)$$

5. 概率和似然区别

概率probability P(x|θ)： 是在特定环境下某件事情发生的可能性

- 已知一枚硬币是均匀的，连续10次正面朝上的概率是多少?
- 对应着机器学习的**预测阶段**: 就是已知参数θ, 来估计该分布下, x应该是什么

似然likelihood L(θ|x)： 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）

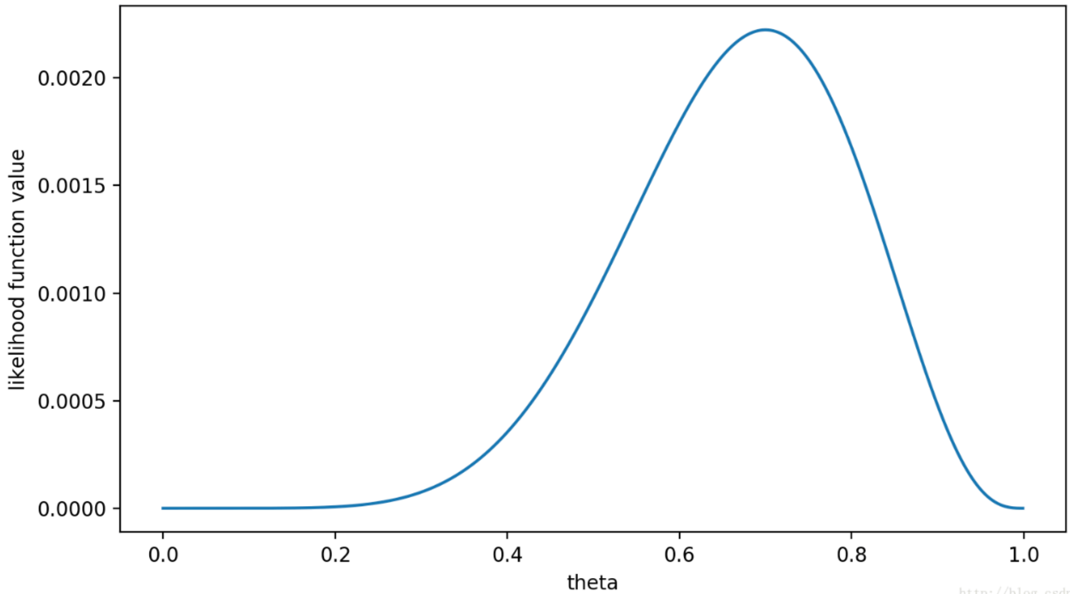
- 一枚硬币连续抛10次正面朝上，这枚硬币是均匀的可能性是多少?
- 对应着机器学习的**训练阶段**：想要机器根据已有的数据(相当于X)学到相应的分布(即θ)

注：在机器学习中，概率函数与似然函数常常都用$P(x|\theta)$表示，区别在于

- 若θ是未知，而x是已知，则称为似然函数
- 若x是变量，而θ是已知，则称为概率函数

6. 最大似然估计和最大后验概率的区别?

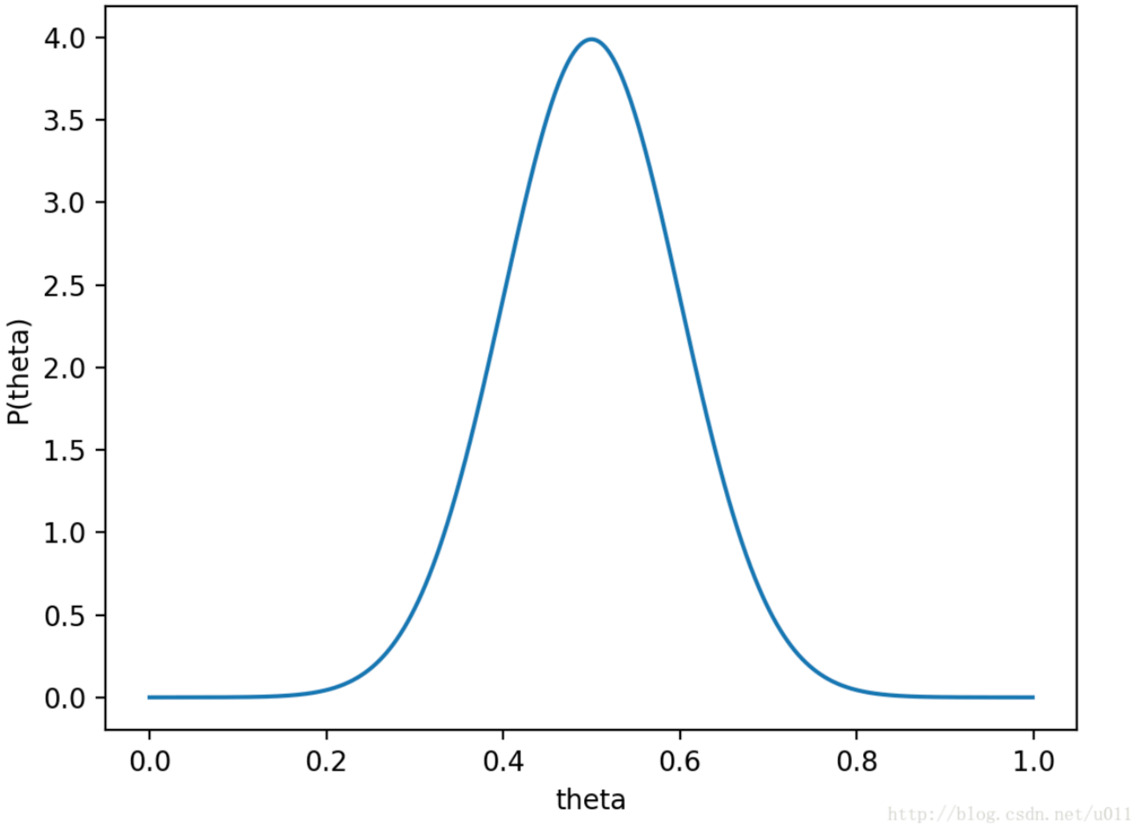
> 两者目标都是求得未知的θ, 主要区别是，最大后验概率假设θ符合某种先验分布。
>
> 对于概率看法不同的两大派别**频率学派**与**贝叶斯派**。他们看待世界的视角不同，导致他们对于产生数据的模型参数的理解也不同。
>
> **频率学派**
>
> - 频率学派认为世界是确定的，通过每个人(X)对于世界的看法(Y)，得到世界的真实环境($\hat{\theta}$)
> - 参数估计方法-**最大似然估计（MLE）**
> - 例子：抛10次硬币$X=(x_1,x_2,...x_{10})$，$Y=(\text{反正正正正反正正正反})$，所以抛硬币x_i得到正面的概率θ 是多少?
>   - θ是当作变量，x是已知定值，所以似然函数：$f(\theta)=P(X|\theta)=(1-\theta)^3\theta^7$

- 

- 所以当 $\theta = 0.7$时候，似然函数最大，即最大似然估计认为，正面向上的概率是0.7
- 显然，当实验次数足够多（数据集足够大），得到的结果越准确

贝叶斯派

- 而贝叶斯学派认为世界是不确定的，世界的环境变化符合一个规律$(P(\theta))$，通过每个人(X)对世界的看法(Y)，不断调整这一规律，我们的目标是要找到最优的描述这个世界的当前环境($\hat{\theta}$)的概率分布($P(\theta|X)$)。
- 估计参数的常用方法-**最大后验概率估计（MAP）**
- 例子：抛10次硬币$X = (x_1, x_2, \ldots x_{10})$，$Y = (\text{反 正 正 正 正 反 正 正 正 反})$，所以抛硬币$x_i$得到正面的概率$\theta$ 是多少?
  - 后验概率：$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$，其中$P(X)$可以忽略
  - 假设θ符合均值0.5，方差0.1的高斯分布，如图

- 

  - $P(X|\theta) = (1-\theta)^3\theta^7$, 所以$P(X|\theta)P(\theta)$，结果如下图
  - 所以当 $\theta = 0.558$的时候，后验概率最大，即最大化后验概率认为，正面向上的概率为0.558，更加符合实际
- 特点：
  - 随着数据量的增加，参数分布会越来越向数据靠拢，先验的影响力会越来越小
  - 如果先验是均匀分布，则贝叶斯方法等价于频率方法。因为直观上来讲，先验是均匀分布本质上表示对事物没有任何预判。

最大似然估计 MLE

$\hat{\theta}_{MLE} = argmax P(X|\theta)$（存在某个$\theta$，使结果尽量拟合训练数据）

$= argmax P(x_1|\theta)P(x_2|\theta)\ldots P(x_n|\theta)$（条件分布假设）

$= argmax log \prod_{i=1}^{n} P(x_i|\theta)$（log化：乘法转为加法，防止上下溢出）

$= argmax \sum_{i=1}^{n} logP(x_i|\theta)$

$= argmin - \sum_{i=1}^{n} logP(x_i|\theta)$

最大后验概率 MAP

$\hat{\theta}_{MAP} = argmax P(\theta|X)$（假设$\theta$是个受到$X$影响的分布）

$= argmin \frac{P(X|\theta)P(\theta)}{P(X)}$ (贝叶斯公式)

$= argminP(X|\theta)P(\theta)$ (忽略分母)

$= argmaxP(x_1|\theta)P(x_2|\theta)...P(x_n|\theta)P(\theta)$ （条件分布假设）

$= argmaxlogP(x_1|\theta)P(x_2|\theta)...P(x_n|\theta)P(\theta)$ （log化：乘法转为加法，防止上下溢出）

$= argmax\sum_{i=1}^{n}logP(x_i|\theta)+logP(\theta)$ （log化：乘法转为加法，防止上下溢出）

$= argmin-\sum_{i=1}^{n}logP(x_i|\theta)-logP(\theta)$

结论

- MLE 与 MAP优化时的不同在于 先验分布项 $-logP(\theta)$
  $MAP(\theta)=MLE(\theta)+P(\theta)$
- 若$P(\theta)$为均匀分布，则二者相同
- 随着数据量的增加，参数分布会越来越向数据靠拢，先验的影响力会越来越小
- 若$P(\theta)$符合正态分布， 所以$P(\theta)=ce^{-\frac{\theta^2}{2\sigma^2}}$，所以$-logP(\theta)=c+\frac{\theta^2}{2\sigma^2}$
  所以，在MAP中使用一个高斯分布的先验等价于在MLE中采用L2的regularizaton！
  在MAP中使用一个拉普拉斯分布的先验等海域在MLE中采用L1的regularizaton