Bayesian Statistics Stuff

贝叶斯公式：

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)} \tag{1}$$

1 先验概率 Prior Probability

基于背景常识、或历史数据的统计得出的预判概率。

2 条件概率 Conditional Probability

表示一个事件发生后另一个事件发生的概率。

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{2}$$

3 后验概率 Posterior Probability

一个随机事件（或一个不确定事件）的后验概率：是在考虑和给出相关数据后所得到的条件概率。

3.1 Example:

假设某学校里有：

- 男生：６０％
- 女生：４０％
- ∀ 男生，穿裤子。
- ５０％ 女生，穿裤子；另外５０％女生，穿裙子。

问题：<u>现知道</u> 某学生穿裤子，问其是女生的概率是多少？

Let

- event A: 看到的是女生
- event B: 看到的是穿裤子的学生

To answer the question, we need to have $\mathbf{P(A|B)}$, meaning: given 穿裤子的学生 (B), the probability of 看到的是女生 (A).

$$\begin{aligned} P(A) &= 40\% \\ P(B|A) &= 50\% \\ P(B) &= 60\% \times 100\% + 40\% \times 50\% = 80\% \end{aligned} \tag{3}$$

使用贝叶斯定理， 求得后验概率：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8} = 0.25$$

4 似然函数 Likelihood Function

关于统计模型中参数的函数, 表示模型参数的拟然性。
在已知某参数 α 时，事件 A 会发生的条件概率 为 $P(A;\alpha)$, same as $P(A|\alpha)$. 我们可以用构造拟然性的方法来表示 事件 A 发生后估计参数 α 的可能性， 表示为 $L(\alpha|A)$。

4.1 拟然与概率的区别：

概率 用于在已知参数的情况下，预测接下来的观测所得到的结果。
而似然 性则是用于在已知某些观测所得到的结果时，对有关事物的性质的参数 进行估计。

4.2 似然函数可以理解为条件概率的逆反:

已知参数 α, 事件 A 的发生概率:

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

使用 Bayes Rule:

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

形式上，似然函数也是一种条件概率函数，但我们关注的变量改变了：$L(\alpha|A) = \lambda P(A|\alpha)$

4.3 Example:

Todo: give example, or refer to the coin-tossing example given in Wikipedia.

5 Review: 最大拟然估计 Maximum Likelihood Estimation (MLE) - Frequentist Statistics

From the point of view of Bayesian Inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters.

模型已定，参数未知。

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。

前提假设：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。

Warning: 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。

Example:

We have i.i.d. samples $x_1, x_2, \cdots, x_n$. The assumed model is f and the parameter for the model is θ. Thus we have:

$$f(x_1, x_2, \cdots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot f(x_3|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Having *log* on both sides, we have:

$$\ln L(\theta|x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} f(x_i|\theta)$$

Thus,

$$\hat{\theta}_{mle} = \arg\max L(\theta|x_1, x_2, \cdots, x_n)$$

Usually, we apply derivatives to get the maximum value of θ

For example, if assumed model follows Gaussian Distribution, then: We have Gaussian i.i.d. samples $x_1, x_2, \cdots, x_n$. Gaussian distribution parameters are μ, σ^2

$$\begin{aligned} L(\mu, \sigma^2|x_1, x_2, \cdots, x_n) &= f(x_1|\mu, \sigma^2) \cdot f(x_2|\mu, \sigma^2) \cdots f(x_n|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n exp\Big(-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \mu)^2\Big) \end{aligned}$$

$$\frac{d}{d\mu} L(\mu, \sigma^2|x_1, x_2, \cdots, x_n) = 0$$

$$\Longrightarrow \mu = \frac{x_1 + x_2 + \cdots x_n}{n}$$

6 最大后验概率 Maximum A Posteriori Estimation

$$\underbrace{P(\theta|X)}_{Posterior Probability} = \frac{P(X|\theta)P(\theta)}{P(X)}$$

We try to have $\hat{\theta}$ from $\arg\max P(\theta|X)$. The $P(X|\theta)$ is similar to the one in MLE, as $P(x_1|\theta) \cdot P(x_2|\theta) \cdots P(x_n|\theta)$. MAP take prior distribution into consideration, that's $P(\theta)$.

Reference: Maximum a Posteriori Probability Estimation (MAP) - Ch2 Estimating Probabilities

7 共轭先验 Conjugate Prior