Too Young Too Simple
sometimes

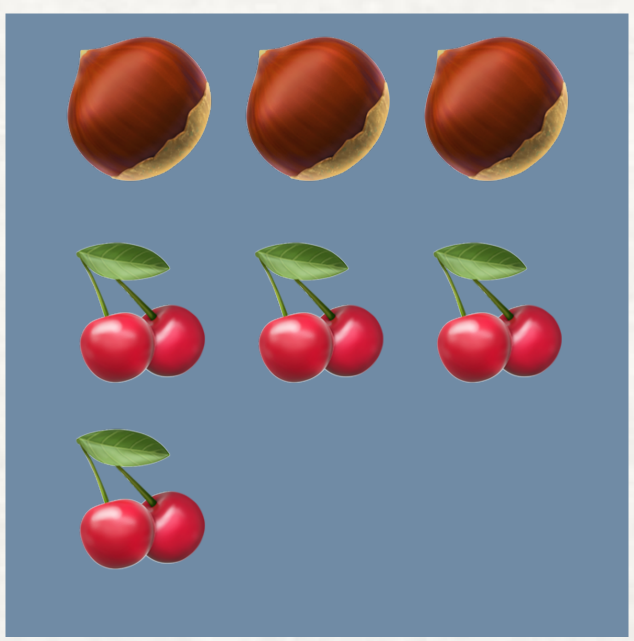
Naive Bayes

朴素贝叶斯

贝叶斯概率

- 贝叶斯概率以18世纪的一位神学家托马斯·贝叶斯命名。贝叶斯概率引入先验知识和逻辑推理来处理不确定命题。

让我们来举个栗子

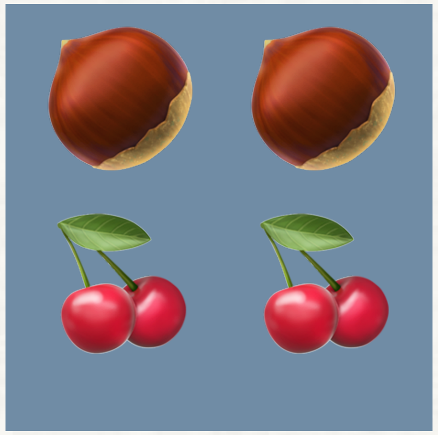


假设有个盒子，里面有三个🌰，四个🍒，那么我们随手抓一个，抓到🌰的概率是多少？抓到🍒的概率又是多少？

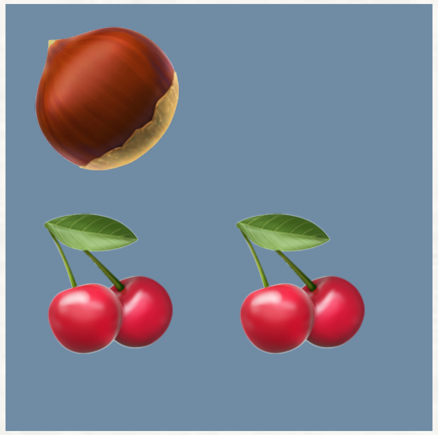
P(🌰) = 3/7　　P(🍒) = 4/7

贝叶斯概率

现在有两个盒子，
A盒里有2个🌰，2个🍒；
B盒里有1个🌰，2个🍒。
那么我们从B盒里随手抓一个，抓到🌰的概率是多少?



A盒



B盒

在“已知我们是从B盒里抽取的条件下，取出🌰的概率”。这便被称为“条件概率”。
在 事件B 发生条件下 事件A 发生的概率如下：

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

P(🌰 | B盒) = P(抽中B盒中的🌰) / P(B盒)

贝叶斯概率

在 事件B 发生条件下 事件A 发生的条件概率：

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

那么，在 事件A 发生条件下 事件B 发生的条件概率：

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

所以，我们可以这样计算：

$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

于是，我们对这个引理进行变换，两边同除 P(A) ，若 P(A) 不为零，便得到了著名的贝叶斯定理：

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

贝叶斯概率

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

其中，P(B) 被称为“先验概率”，在A事件发生之前，对B事件概率的一个判断。

P(B | A) 则被称为“后验概率”，是在A事件发生之后，对B事件的重新评估。

P(A | B) / P(A) 则被称为“可能性函数”，是一个调整因子，使得预估概率更接近真实概率。

朴素贝叶斯

- 朴素贝叶斯是基于 贝叶斯定理 与 特征条件独立假设 的分类方法。
- 对于给定的训练集，首先基于特征条件独立假设学习输入 / 输出的联合概率分布
- 然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。
- 因为我们假设每个特征条件之间是独立的，比如一个单词出现的概率和其他相邻单词没有关系，所以这个算法就被称为“天真”的贝叶斯。
- 然而虽然Naive，但朴素贝叶斯实现简单，学习和预测的效率都很高，所以应用很广泛。

邮件分类器

- 朴素贝叶斯的一个常见用途是来区分垃圾邮件。
- 我们用S表示垃圾邮件，H表示正常邮件。
- 随机抽取一封邮件，抽到垃圾邮件的概率是 P(S) ,正常邮件的概率是 P(H)
- 给出一封邮件 D，D= { W_1, W_2, …, W_n }

D是垃圾邮件的概率：

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)}$$

D是正常邮件的概率：

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

如果 P(S | D) > P(H | D) ,则邮件D是垃圾邮件，
反之，若 P(S | D) < P(H | D) ,则邮件D是正常邮件。

先验概率 P(S) & 条件概率 P(D | S) 如何求?

先验概率 P(S)： 极大似然估计

- 垃圾邮件的概率是 $P(S)$,正常邮件的概率是 $P(H)$, $P(S) + P(H) = 1$

- 从网上所有的邮件中抽取 m+n 封邮件，其中 m 封垃圾邮件，n 封正常邮件的概率为：

$$P = P(S)^m P(H)^n = P(S)^m (1 - P(S))^n$$

- 假设有100封邮件的训练样本，其中30封垃圾邮件，70封正常邮件，它是由上述的概率模型产生的，那么我们就可以依靠这个样本来估计参数 $P(S)$，这个估计基于这样的思想：我们所估计的模型参数，要使得产生这个样本集的可能性最大。

- 所以我们所求出的 $P(S)$，要使 P 最大：

$$P = P(S)^{30} (1 - P(S))^{70}$$

- 对上式求导，令其等于 0 ，得：

$$30P(S)^{29}(1 - P(S))^{70} - 70(1 - P(S))^{69} P(S)^{30} = 0$$
$$30P(S)^{29}(1 - P(S))^{70} = 70(1 - P(S))^{69} P(S)^{30}$$

$$\rightarrow \quad \frac{1 - P(S)}{P(S)} = \frac{7}{3}$$

- 所以 P(S) = 0.3

条件概率 P(D | S)：特征条件独立假设

- 给出一封邮件 D，D= { $W_1, W_2, \ldots, W_n$ }

- $P(D \mid S) = P(W_1, W_2, \ldots, W_n \mid S)$

- $P(D \mid S) = P(W_1 \mid S)\, P(W_2 \mid S, W_1)\, P(W_3 \mid S, W_1, W_2) \ldots P(W_n \mid S, W_1, \ldots, W_{n-1})$

这么麻烦

Are You Kidding?

你是柯丁吗?

- $P(D \mid S) = P(W_1 \mid S)\, P(W_2 \mid S)\, P(W_3 \mid S) \ldots P(W_n \mid S)$

$$P(W_i|S) = \frac{P(W_i, S)}{P(S)}$$

$$P(S) = \frac{\text{垃圾邮件的个数}}{\text{邮件总数}} \qquad P(W_i, S) = \frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{邮件总数}}$$

$$P(W_i|S) = \frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{垃圾邮件的个数}}$$

拉普拉斯平滑

- 给出一封邮件 D，D= { W_1, W_2,W_k} （k ≠ 1 , 2 , … , n）

$$P(W_k|S) = \frac{\text{包含 } W_k \text{ 的垃圾邮件个数}}{\text{垃圾邮件个数}} = 0 \qquad P(W_k|H) = \frac{\text{包含 } W_k \text{ 的正常邮件个数}}{\text{正常邮件个数}} = 0$$

- 所以相应的，$P(S \mid D) = 0$ & $P(H \mid D) = 0$

拉普拉斯平滑放大灯！

$$P(W_k|S) = \frac{\text{包含 } W_k \text{ 的垃圾邮件个数} + 1}{\text{垃圾邮件个数} + 1} \neq 0 \qquad P(W_k|H) = \frac{\text{包含 } W_k \text{ 的正常邮件个数} + 1}{\text{正常邮件个数} + 1} \neq 0$$

作业

- 在4000封邮件的训练集中训练朴素贝叶斯分类器，对1000封邮件的测试集进行分类，给出 分类 结果。

谢谢