Naive Bayes Classifer（朴素贝叶斯）

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i \mid a_1, a_2, \ldots, a_n)$$

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \frac{P(a_1, a_2, \ldots, a_n \mid \omega_i) P(\omega_i)}{P(a_1, a_2, \ldots, a_n)}$$

理论上，应该按上面的公式进行实现，但一般联合概率是不可求的，除非样本数据量极大，属性较少。所以实际上，按下面的公式进行实现

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \underline{P(a_1, a_2, \ldots, a_n \mid \omega_i)} P(\omega_i)$$

Conditionally Independent

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i) \prod_j P(a_j \mid \omega_i)$$

将联合概率 → 多个边缘密度的乘积，该转换暗含了一个条件独立的假设

独立：（条件独立，指的是在G发生的时候A,B才是独立的）

$$P(A \cap B) = P(A)P(B|A) \quad + \quad P(B|A) = P(B)$$

$$\Downarrow$$

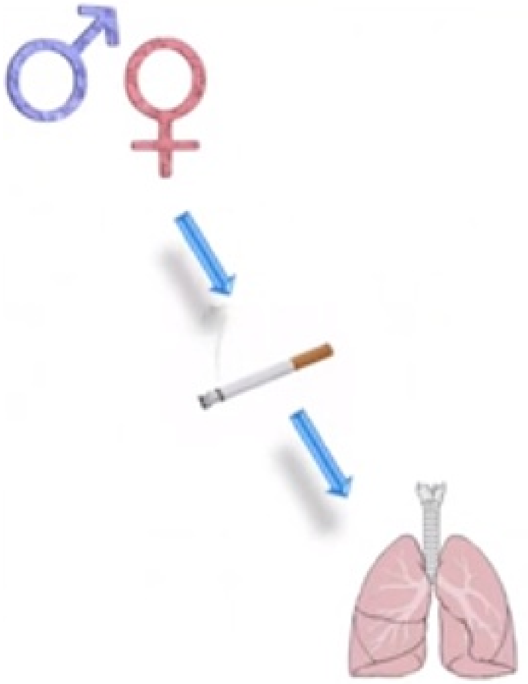
$$P(A \cap B) = P(A)P(B)$$

Conditionally Independent

$$P(A, B \mid G) = P(A \mid G) P(B \mid G) \Longleftrightarrow \underline{P(A \mid G, B)} = P(A \mid G)$$

$$P(A, B \mid G) = P(A, B, G) / P(G) = P(A \mid B, G) \times P(B, G) / P(G)$$
$$= \underline{P(A \mid B, G)} \times P(B \mid G)$$

举例：假如调查出的比率直观上表明男性的肺癌的概率是大于女性的，你会认为得肺癌和性别是不独立的。但是假设无论男女只要吸烟就会得肺癌，并不是说男性有更大的几率得肺癌，而是男性中是吸烟者的概率要大于女性，所以在你一直这个条件下，得肺癌和吸烟是相关的，而和性别是独立的。



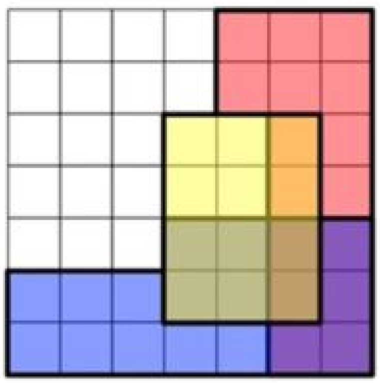
$$P(Cancer|Male) = 65/100{,}000$$
$$P(Cancer|Female) = 48/100{,}000$$

- Are the two events Male/Female and Cancer independent?
- Assume smoking is the sole contributing factor to cancer.

Conditionally Independent

$$P(Cancer|Male, Smoking) = P(Cancer|Smoking)$$

要注意问题本质上的因素



$P(R \cap B) = 6/49$

$P(R) = 16/49$

$P(B) = 18/49$

$P(R \cap B) \neq P(R)P(B)$

Not Independent

$P(R \cap B|Y) = 1/6$

$P(R|Y) = 1/3$

$P(B|Y) = 1/2$

$P(R \cap B|Y) = P(R|Y)P(B|Y)$

Conditionally Independent

- Two coins: fair vs. biased (two-headed)
- Select one coin at random and toss twice.
- A: First coin toss is head.
- B: Second coin toss is head.
- C: You selected the fair coin.

$$P(A) = P(B) = 0.5 \times 0.5 + 0.5 \times 1.0 = 0.75$$

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|\neg C)P(\neg C)} = \frac{0.5 \times 0.5}{0.5 \times 0.5 + 1 \times 0.5} = \frac{1}{3}$$

$$P(B|A) = \frac{1}{3} \times 0.5 + \frac{2}{3} \times 1.0 = \frac{5}{6} \neq P(B)$$ Not Independent

$$P(B|A, C) = P(B|C) = 0.5$$ Conditionally Independent

独立 不等于 不相关

$$\boxed{\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}}$$

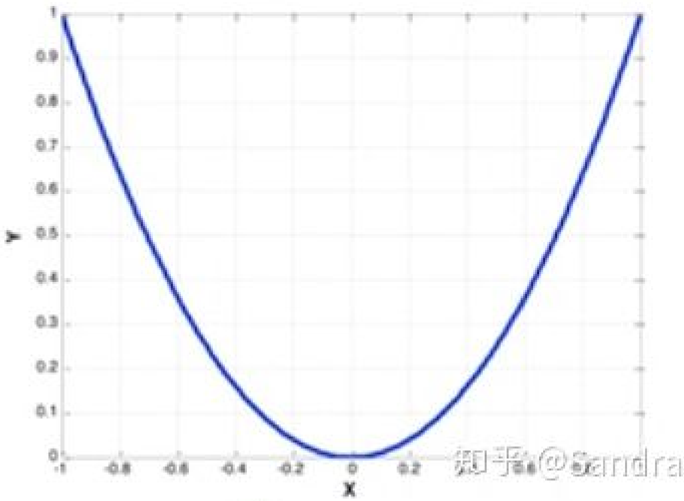
$X \in [-1,\ 1]$

$Y = X^2$

X	Y
1	1
0.5	0.25
0.2	0.04
0	0
-0.2	0.04
-0.5	0.25
-1	1

Cov (X,Y)=0 → X and Y are uncorrelated.

However, Y is completely determined by X.



事情是没有绝对性的，没见过的不代表不会出现或发生。为了避免概率连乘中出现0，是其他属性判断失效经常采用最下面的公式（拉普拉斯平缓）：

a_1	a_2	a_3	ω
	+		ω_1
			ω_2
	-		ω_1
	+		ω_1
			ω_2

$P(\omega_1) = 3/5; \qquad P(\omega_2) = 2/5$

$P(a_2 = '+' \mid \omega_1) = 2/3$

$P(a_2 = '-' \mid \omega_1) = 1/3$

Laplace Smoothing $$P(a_{jk} \mid \omega_i) = \frac{|a_j = a_{jk} \wedge \omega = \omega_i| + 1}{|\omega = \omega_i| + |a_j|}$$

若是连续性的属性呢？概率分布（高斯）函数-->计算概率

Given :

< Outlook = *sunny*, Temperature = *cool*, Humidity = *high*, Wind = *strong* >

Predict :

PlayTennis (*yes or no*)

Bayes Solution :

$P(PlayTennis = yes) = 9/14$

$P(PlayTennis = no) = 5/14$

$P(Wind = strong \mid PlayTennis = yes) = 3/9$

$P(Wind = strong \mid PlayTennis = no) = 3/5$

...

$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$

$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = 0.0206$

The conclusion is not to play tennis with probability : $\frac{0.0206}{0.0206 + 0.0053} = 0.795$

文章的筛选：（文章特点倒推分类）

a_1	a_2	a_3	a_4	...	a_n	ω
Long	long	ago	there	...	king	1
New	sanctions	will	be	...	Iran	0
Hidden	Markov	models	are	...	method	0
The	Federal	Court	today	...	investigate	0

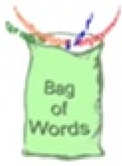
We need to estimate probabilities such as $P(a_2 = king \mid \omega = 1)$.

However, there are 2×n×|Vocabulary| terms in total. For n=100 and a vocabulary of 50,000 distinct words, it adds up to 10 million terms!

2：代表的是w1，w2；n：提取文章的前多少个单词（位置）；Vocabulary：常用英文单词数

文章于什么有关只根据关键字出现的次数有关，跟出现的位置无关（计算单词出现的次数）

- By only considering the probability of encountering a specific word instead of the specific word position, we can reduce the number of probabilities to be estimated.
- We only count the frequency of each word.
- Now, 2×50,000=100,000 terms need to be estimated.



$$P(V_K \mid \omega = \omega_i) = \frac{n_k + 1}{n + |Vocabulary|}$$

- n: the total number of word positions in all training samples whose target value is ω_i.
- n_k: the number of times word V_k is found among these n positions.

$$P(V_k \mid \omega = \omega_1) = \frac{n_k + 1}{n + |Vocabulary|}$$

--> 其中
$P(V_k|\omega=\omega_1)$
：我所感兴趣（或不感兴趣）文章中的特定的单词出现的概率； n：所有感兴趣的文章中单词的个数；
n_k
：某一个单词（1和Vocabulary是为了拉普拉斯平滑）。

应用：