条件独立性假设

$$p(\mathbf{w}|ci) = p(w0,w1,\ldots,wN|ci) = p(w0|ci)p(w1|ci)\ldots p(wN|ci)$$

	w_1, w_2, w_3, w_4	...	w_m, w_{m+1}	...	w_n	c_i
d_1	[0, 0, 0, 0, 0, 1, 0, 0, 0, ...		1, 0, 0, 0, 0, 1, 0, 0, 0, 0]			0
d_2	[1, 0, 0, 0, 0, 1, 0, 0, 0, ...		0, 0, 1, 0, 1, 0, 0, 1, 0, 0]			1
d_3	[0, 0, 0, 0, 1, 0, 0, 1, 0, ...		0, 1, 0, 0, 0, 0, 1, 1, 0, 1]			0
d_4	[0, 1, 0, 0, 0, 0, 1, 0, 0, ...		0, 0, 0, 0, 0, 0, 0, 0, 1, 0]			1
d_5	[0, 1, 1, 1, 0, 0, 0, 0, 1, ...		0, 0, 0, 1, 0, 0, 0, 1, 0, 0]			0
	[0, 0, 0, 0, 0, 1, 0, 0, 0, ...		0, 0, 0, 0, 0, 0, 0, 0, 1, 0]			1

条件独立性假设

$$p(\mathbf{w}|ci) = p(w0,w1,...,wN|ci) = p(w0|ci)p(w1|ci)...p(wN|ci)$$

```
[0, 0, 0, 0, 0, 1,
[1, 0, 0, 0, 0, 1,
[0, 0, 0, 0, 1, 0,
[0, 1, 0, 0, 0, 0,
[0, 1, 1, 1, 0, 0,
[0, 0, 0, 0, 0, 1,
```

Naive-Bayes 中“naïve”的含义：🐸

根据统计学知识，我们知道如果每个特征需要 N 个样本，那么对于 10 个特征，我们就需要 N^{10} 个样本，对于包含 1000 个特征的词汇表就需要 N^{1000} 个样本。可以看到，所需要的样本数会随着特征数目真大而迅速增大。

但是，如果特征之间相互独立，那么样本数就可以从 N^{1000} 减少到 $1000 * N$。

所谓的独立指的就是统计意义上的独立，即一个特征或者单词出现的可能性与它和其他单词相邻没有关系。

👰个🍗：

在句子“it is a sunny day”中，day 出现在 sunny 后面和出现在 it 后面的概率相同。

当然啦，这个假设明显是不正确的。

但是，这个假设却就是我们所说的大名鼎鼎的 naive-bayes classification 里面的“naive”一词的含义。

条件独立性假设

$$p(\mathbf{w}|ci) = p(w0,w1,...,wN|ci) = p(w0|ci)p(w1|ci)...p(wN|ci)$$

```
[0, 0, 0, 0, 0, 1,
[1, 0, 0, 0, 0, 1,
[0, 0, 0, 0, 1, 0,
[0, 1, 0, 0, 0, 0,
[0, 1, 1, 1, 0, 0,
[0, 0, 0, 0, 0, 1,
```

Naive-Bayes 中“naïve”的含义（续）：🐸🐸

既然都说了 naïve 了，那就顺便再讲一下 naïve-bayes classifier 的另一个假设：

每个特征同等重要。

在很多任务中，其实我们不需要了解所有的特征，也许主要几十个特征我们就可以足以做出判断了。

🤦个🏉：

在垃圾邮件的分类任务中，也许一封邮件内有 1000 个单词（也就是说，有 1000 个特征），但是，我们通常只需要通过观察 20 个左右的特征就可以做出判断了。

条件独立性假设

$$p(\mathbf{w}|ci) = p(w0,w1,...,wN|ci) = p(w0|ci)p(w1|ci)...p(wN|ci)$$

```
[0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0]
[0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1]
[0, 1, 0, 0, 0, 0, 1, 0, 0, 0,                                                    0, 0, 0, 1, 0]
[0, 1, 1, 1, 0, 0, 0, 0, 1, 1,                                                    0, 0, 1, 0, 0]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0,                                                    0, 0, 0, 1, 0]
```

尽管，naïve bayes 存在上述的这些小瑕疵，但是作为一个基于概率论的数学模型，它的实际效果却是很好的。

条件独立性假设

$$p(\mathbf{w}|ci) = p(w0,w1,\ldots,wN|ci) = p(w0|ci)p(w1|ci)\ldots p(wN|ci)$$

```
[0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0] 0
[1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0] 1
[0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1] 0
[0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0] 1
[0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0] 0
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0] 1
```

$w_0|C_1$ num

$w_1|C_1$ num

$w_2|C_1$ num