Naive Bayes Analysis（單純貝氏分析）

> Reference：
>
> 1. Naïve Bayes Analysis(單純貝氏) - Dr. Tun-Wen Pai
> 2. Naive Bayes, Clearly Explained!!! - StatQuest with Josh Starmer
> 3. Bayes' Theorem, Clearly Explained!!!! - StatQuest with Josh Starmer
> 4. Gaussian Naive Bayes, Clearly Explained!!! - StatQuest with Josh Starmer

貝氏定理

貝氏定理是關於隨機事件 A 和 B 的條件機率的一則定理，定義如下：

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, P(B) \neq 0$$

其中 $P(A)$ 是事件 A 的事前機率，$P(B)$ 是事件 B 的事前機率。

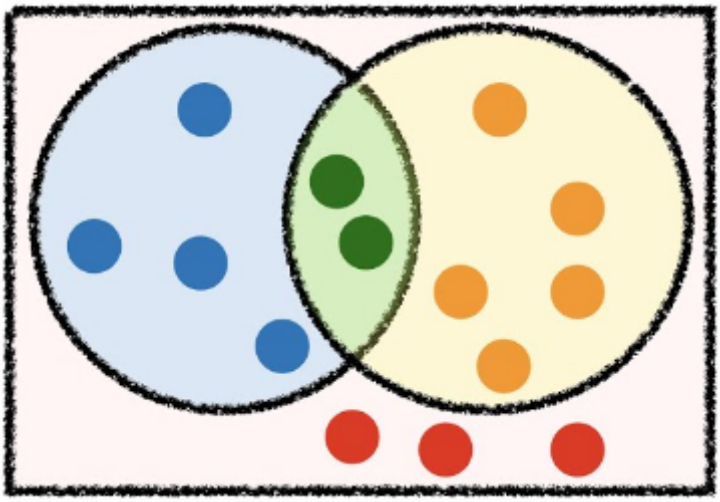
$P(B|A)$ 是事件 A 發生後事件 B 的條件機率（事後機率），同時，$P(B|A) = \frac{P(A \cap B)}{P(A)}$。

$P(A|B)$ 是事件 B 發生後事件 A 的條件機率（事後機率），同時，$P(A|B) = \frac{P(A \cap B)}{P(B)}$。

以一張文氏圖來說，我們可以令藍色區塊為 A，黃色區塊為 B，則若我們想要知道事件 B 發生後，發生 A 的機率

就相比我們以藍色區塊點的數量去除以綠色區塊點的數量 $P(G|B) = \frac{N(G)}{N(B)}$，就能得到先發生 B 後再發生 A 的機率。

> 文氏圖圖例，By StatQuest



多項單純貝氏分類

概述

當我們要對一筆資料 D 進行分類，我們可以根據資料的每個特徵 C_i，從訓練樣本中找出特徵 C_i 是事件 A 或事件 B 的條件機率 $P(C_i|A)$ 或 $P(C_i|B)$ 的分數，用分數來進行分類，定義如下。

$$P(D|E) = P(E) \times P(C_1|E) \times P(C_2|E) \times \ldots \times P(C_n|E)$$

其中 D 為一筆資料，E 為一個事件，$P(D|E)$ 為資料 D 歸類在事件 E 的分數，C_i 為某特徵。

分類會出現有零值與沒有零值的情況。

Example - 無零值

#	message	Happy
1	love happy joy joy love	Yes
2	happy love kick joy happy	Yes
3	love move joy good	Yes
4	love happy joy pain love	Yes
5	joy love pain kick pain	No
6	pain pain love kick	No
7	love pain joy love kick	?

我們想要找到第七筆應該被分類在哪一個類別（Happy = Yes/No），我們可以先算出事前機率：

$$P(yes) = \frac{4}{6} = \frac{2}{3}$$
$$P(no) = \frac{2}{6} = \frac{1}{3}$$

我們可以根據每個特徵來算出 D_7 特徵值的事後機率。

特徵值	Yes	No
good	1	0
happy	4	0
joy	5	1
kick	1	2
love	6	2
move	1	0
pain	1	4
總計	19	9

$$P(love|yes) = \frac{6}{19}, P(love|no) = \frac{2}{9}$$
$$P(pain|yes) = \frac{1}{19}, P(pain|no) = \frac{4}{9}$$
$$P(joy|yes) = \frac{5}{19}, P(joy|no) = \frac{1}{9}$$
$$P(kick|yes) = \frac{1}{19}, P(kick|no) = \frac{2}{9}$$

那我們就能夠根據定義來算出 D_7 特徵值的分數。

$$P(D_7|yes) = P(love|yes) \times P(pain|yes) \times P(joy|yes) \times P(love|yes) \times P(kick|yes) \approx 0.000004$$
$$P(D_7|no) = P(love|no) \times P(pain|no) \times P(joy|no) \times P(love|no) \times P(kick|no) \approx 0.00018$$

故 D_7 在 $P(D_7|no)$ 有較高的分數，故選 no。

Example - 有零值

#	message	Happy
1	love happy joy joy love	Yes
2	happy love kick joy happy	Yes
3	love move joy good	Yes
4	love happy joy pain love	Yes
5	joy love pain kick pain	No
6	pain pain love kick	No
7	love pain joy happy kick happy	?

我們想要找到第七筆應該被分類在哪一個類別（Happy = Yes/No），我們可以先算出事前機率：

$$P(yes) = \frac{4}{6} = \frac{2}{3}$$
$$P(no) = \frac{2}{6} = \frac{1}{3}$$

我們可以根據每個特徵來算出 D_7 特徵值的事後機率。

但此時會發現 happy 這個特徵值並沒有出現在 No 裡面，通常來說，我們會把每一個特徵值重複加上去，如下：

	修正前		修正後	
特徵值	Yes	No	Yes	No
good	1	0	2	0
happy	4	0	5	1
joy	5	1	6	2
kick	1	2	2	3
love	6	2	7	3
move	1	0	2	0
pain	1	4	2	5
總計	19	9	26	14

那我們就能夠根據定義來算出 D_7 特徵值的分數。

$$P(D_7|yes) = P(love|no) \times P(pain|yes) \times P(joy|yes) \times P(love|yes) \times P(kick|yes) \approx 9 \times 10^{-6}$$
$$P(D_7|no) = P(love|yes) \times P(pain|no) \times P(joy|no) \times P(love|no) \times P(kick|no) \approx 3 \times 10^{-6}$$

故 D_7 在 $P(D_7|yes)$ 有較高的分數，故選 yes。

常態單純貝氏分類

> 另一種單純貝氏分類，屬於補充資料，有時間補。