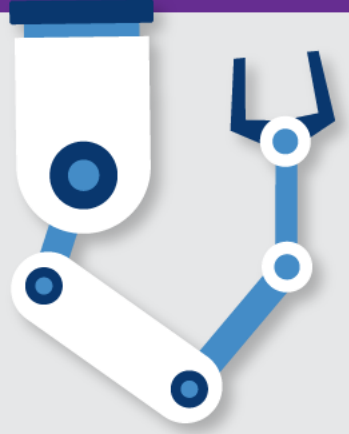
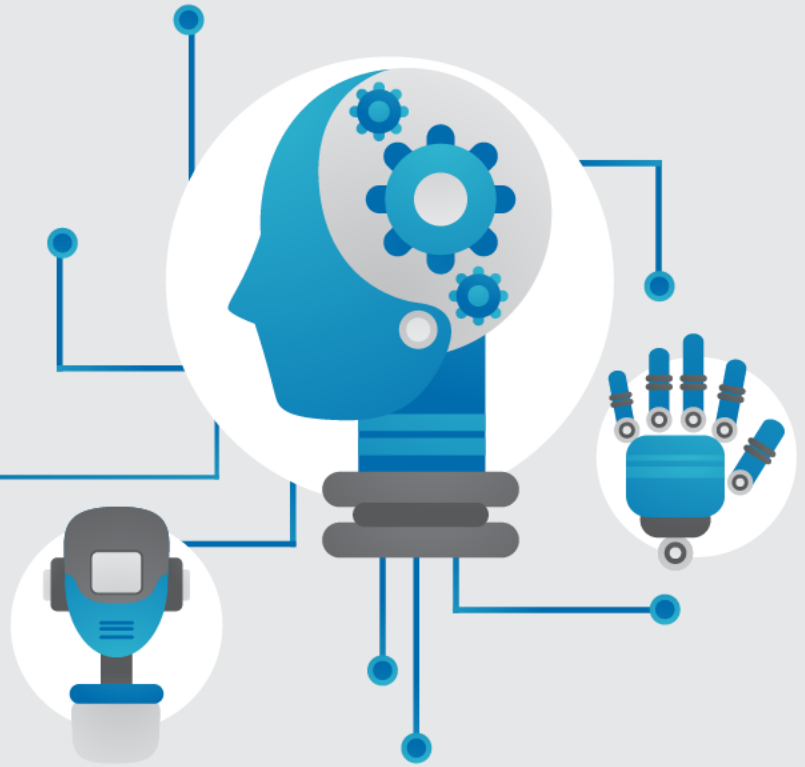


貝氏分類器



貝氏分類器

- **單純貝氏分類器（Naïve Bayes' Classifier）**是假設特徵之間互相獨立，運用貝葉斯定理（Bayes' Theorem）為基礎的簡單機率分類器。

- 貝氏分類器在20世紀60年代初引入到資料檢索領域，目前仍然是文件分類的一種熱門方法。

- 文件分類是以詞頻為特徵判斷文件所屬類別或其他（如垃圾郵件、合法性、體育或政治等）的問題。

貝氏定理

- **貝氏定理（Bayes' theorem）**是機率論中的一個定理，描述在已知一些條件下，某事件的發生機率。

$$P(A|B) = P(B|A)\,P(A) / P(B)$$

P(A)、P(B)：是A、B的**先驗機率**
P(A|B)：已知B發生後，A的條件機率。稱作 A的**後驗機率**。
P(B|A)：已知A發生後，B的條件機率。稱作 B的**後驗機率**。

貝氏定理

› 根據條件概率的定義

事件B發生的條件下事件A發生的概率是 $P(A|B) = P(A \cap B) / P(B)$

事件A發生的條件下事件B發生的概率是 $P(B|A) = P(A \cap B) / P(A)$

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

所以得到貝氏定理 $P(A|B) = P(B|A)\ P(A) / P(B)$

貝氏分類器概念

- 假設我們根據一些**屬性**（features）將一些物件進行**分類**（classes）

- 給定一組屬性F，可計算具有屬性F的物件O屬於類別C的機率，我們表示為$P(C|F)$

- 根據所有類別C_i的$P(C_i|F)$，物件O就分類屬於那個$P(C_i|F)$值最高的類別。

貝氏分類器概念

- 例如紅酒可以分類**餐酒**（Table wine）、**優質酒**（ Premium）和**暢飲酒**（Swill）。酒的屬性包含**酸度**（acidity），**酒體**（body），**色澤**（color）和**價格**（price）

- 假設有一種酒W具有屬性F，而且我們知道 $P(T|F) = 0.5$，$P(P|F) = 0.2$ and $P(S|F) = 0.3$

➡ 於是我們可以將W分類為**餐酒**，因為$P(T|F)$最高

多分類貝氏定理

- 假設給定n個兩兩不相交的類別$C_1, C_2, \ldots, C_n$和一組屬性F
- 一個物件具備屬性F，分類為C_j的機率

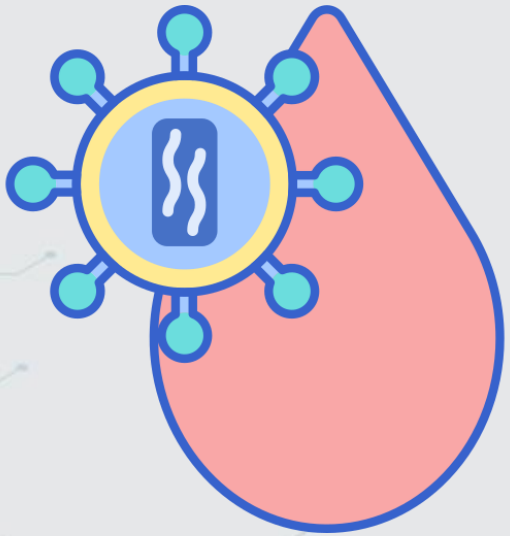
$P(C_j \mid F) = X / Y$，其中

$X = P(F \mid C_j)\, P(C_j)$，$Y = \sum_{i=1}^{n} P(F \mid C_i)\, P(C_i)$

快篩檢驗

- 有一種愛滋病（HIV）的快篩檢驗方法稱為**酵素免疫分析法**（ELISA）

- 兩種類別
  - H：有HIV病毒
  - H^c ：沒有HIV病毒



- 屬性

  Pos：ELISA檢驗為陽性

快篩檢驗

- 實驗資訊

P(H)=0.15, P(H^c)=0.85, P(Pos|H)=0.95, P(Pos|H^c)=0.02

$$P(H \mid Pos) = \frac{P(Pos \mid H)P(H)}{P(Pos \mid H)P(H) + P(Pos \mid H^c)P(H^c)}$$

$$= \frac{(0.95)(0.15)}{(0.95)(0.15) + (0.02)(0.85)} = 0.893.$$

貝氏分類器種類

- 當單純貝氏分類器所給定的特徵是連續的數值，我們使用高斯貝氏分類器（GaussianNB）。
- 假設變數為**常態分配**的情況下，以樣本的資料的**標準差**及**變異數**來計算機率，算式如下

$$P(x \mid y=c, D) = \prod_{j=1}^{n} N(x_j \mid \mu_{jc}, \delta_{jc}^2)$$

- 當資料為二元值時，可採用**伯努力（BernouliNB）分類器**
- 在離散資料方面，也可採用**多項式（MultinomialNB）分類器**

貝氏分類器優缺點

優

- 樸素貝氏模型發源於古典數學理論，有堅實的數學基礎，及穩定的分類效率。
- 對小規模的數據表現很好，能夠處理多分類任務，適合增量式訓練。
- 對缺失數據不太敏感，算法也較簡單。

缺

- 無法學習特徵間的相互作用
- 需要計算先驗機率