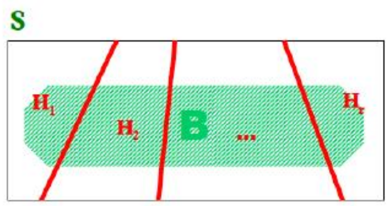
貝氏分類法 (Bayesian Classifier)

原理

由結果去追溯某個原因發生的機率，即由後天去推測先天。

設{H1, H2,…, Hr}為樣空間 S 中的分割(r>=2)，B 為 S 中的任意事件。直觀來說，H 集合為現有的特徵，B 為欲預測的值(label)。
若 P(B)>0，P(Hi)>0，i=1,2,…,r，j=1,2,…,r，則:

$$P(H_j \mid B) = \frac{P(H_j \cap B)}{P(B)} = \frac{P(H_j \cap B)}{\sum_{i=1}^{r} P(H_i \cap B)} = \frac{P(H_j)P(B \mid H_j)}{\sum_{i=1}^{r} P(H_i)P(B \mid H_i)}$$



$P(H_j)$：事前機率(先天機率)，依據現有資訊所求得的機率。

$P(H_j \mid B)$：事後機率，根據額外的資訊，經修正求得的機率。

由算式中可知，可以藉由事後機率 P(Hj|B)推算出 P(B|Hj)，即在現有特徵 H 下發生 B 的機率。

基本假設

1. 所有變數(特徵)對分類均是有用的
2. 變數(特徵)間相互獨立
3. 變數(特徵)間條件獨立

$$\mathbf{P(A \cap B \mid C) = P(A \mid C) \times P(B \mid C)}$$

證明

以 B = Yes 來說，其事後機率為：

$$P(B=Yes\mid H)=\frac{P(H\cap B=Yes)}{P(H)}=\frac{P(H\mid B=Yes)P(B=Yes)}{P(H)}$$

條件獨立

$$=\frac{\boxed{P(H_1\cap H_2\cap H_3\cap H_4\mid B=Yes)}P(B=Yes)}{P(H)}$$

$$=\frac{P(B=Yes)\boxed{P(H_1\mid B=Yes)P(H_2\mid B=Yes)P(H_3\mid B=Yes)P(H_4\mid B=Yes)}}{P(H)}$$

$$=\frac{P(B=Yes)\prod_{i=1}^{4}P(H_i\mid B=Yes)}{P(H)} \qquad \text{.........(1)}$$

以 B = No 來說，其事後機率為：

$$P(B=No\mid H)=\frac{P(H\cap B=No)}{P(H)}=\frac{P(H\mid B=No)P(B=No)}{P(H)}$$

條件獨立

$$=\frac{\boxed{P(H_1\cap H_2\cap H_3\cap H_4\mid B=No)}P(B=No)}{P(H)}$$

$$=\frac{P(B=No)\boxed{\prod_{i=1}^{4}P(H_i\mid B=No)}}{P(H)} \qquad \text{.........(2)}$$

國立聯合大學 資訊管理學系 |

不論是求解 B 為 Yes 或 No，分母 P(H)皆相同，皆可視為固定的常數，為計算的簡便性可以省略它，因此攻式(1)、(2)可修正為概似函數

$$P(B=Yes\mid H)=P(B=Yes)\prod_{i=1}^{4}P(H_i\mid B=Yes) \qquad \text{.........(3)}$$

$$P(B=No\mid H)=P(B=No)\prod_{i=1}^{4}P(H_i\mid B=No) \qquad \text{.........(4)}$$

比較(3)、(4)機率值大小，而 B 的結果取決於較大者

連續型態資料欄位

常態分配(高斯分配)最常被用來表示連續變數的類別條件機率

- **該分配有兩個參數：平均數 μ 與變異數 σ^2**
- **對於每個類別 y_j 而言，連續型資料欄位 X 的類別條件機率如下：**

$$P(X=x_i\mid Y=y_j)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- **參數 μ 可以經由連續型資料欄位 X 中、屬於類別 y_j 之樣本平均數 $\bar{X}_j$ 估計而來。而 σ^2 可以從樣本變異數 S_j^2 估計而來。**

當資料有遺漏時，對分類結果不會造成太大的影響

當訓練資料有資料遺漏時:

當進行頻率計數時，僅需對該遺漏值的計算，同時在機率計算中使用實際出現值之個數，而非訓練資料的總數。

當測試資料有資料遺漏時:

僅需簡單忽略這個屬性，機率會比先前計算值高很多，但不影響結果。

某個屬性的資料中不是每個類別都出現

當訓練資料過少且特徵個數過多時可能會產生此問題

使用Laplace Estimator評估條件機率的作法如下：

- **概念：對造成條件機率為0之屬性，將其所屬類別 b_j 之機率計算公式的分子、分母皆加上一數值，使該機率不為0**
  - **分子：加上1。**
  - **分母：加上q，其中 q 為該屬性內的不同資料個數。**

參考資料

1. Class Handout, Lee, Chia Jung professor, MDM64001, School of Big Data Management, Soochow University
2. 單純貝氏分類器 (https://zh.wikipedia.org/wiki/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%88%86%E7%B1%BB%E5%99%A8)
3. AI - Ch15 機器學習(3), 樸素貝葉斯分類器 Naive Bayes classifier (https://mropengate.blogspot.com/2015/06/ai-ch14-3-naive-bayes-classifier.html)