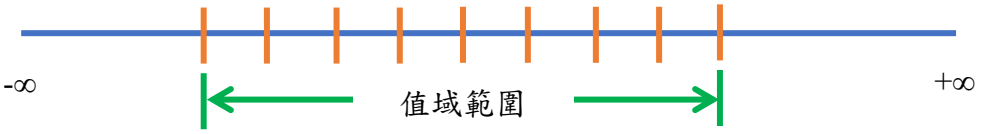
Artificial Intelligence Naïve Bayesian Project
Student ID: 0886006
Name: 張又允

(1) 資料前處理

- 處理文字屬性
  在訓練資料裡，有 Laterality 屬於文字屬性，其值有 L 與 R 兩種。在資料前處理過程中，先將 L 對應到數值 0、R 對應到數值 1，以利後續的公式計算。

- 連續屬性 (numerical features) 的前處理方式
  對於連續屬性，我們採取 discretization 的策略。
  細節如下：
  針對某個連續屬性，先對訓練資料的值域範圍，切割成等寬的 bins，然後再新增兩個在頭與尾的 bins 處理極端資料的狀況。示意圖如下:



- 離散屬性 (categorical features) 的前處理方式
  對於離散屬性，單純用整數的方式處理。另外，為了使接下來的計算的部分簡單統一化，對於離散屬性，我們可以想像成在資料的值域範圍上切割出等寬的 bins (bin width = 1)。每個 bin 所對應到的整數 x，取上下界[x - 0.5, x + 0.5)為 bin edges。同時，也在頭與尾的部分新增兩個 bins 處理極端資料的狀況。

(2) 計算公式

- Probability 計算
  在接下來所提到的 probability，計算方式皆使用下面所提到的方式:
  假設要計算的對象為 f 以及其特定值 c
  Step 1: 計算特定值 c 的出現次數 x
  Step 2: 計算全部的次數 y
  Step 3: 套用 Laplace correction 計算 probability

$$p(f = c) = \frac{x + 1}{y + n}$$

  其中 n 為 label 的總數。

- 計算 prior probability
  對於訓練資料裡給定的 target 值，分別計算 $p(\text{target} = 0)$ 與 $p(\text{target} = 1)$ 的 prior probability

- 計算 conditional probability $p(a \mid b)$
  對於訓練資料裡給定的 target 值，再針對每個屬性利用在資料前處理中所提及的方式計算$p(v \in \text{Bin}_x \mid \text{target} = i), i = 0, 1$

- 計算 posterior probability
  原公式：

$$p(\text{target} = i | f_1 = v_1, f_2 = v_2, \dots, f_n = v_n)$$
$$\propto p(f_1 = v_1 | \text{target} = i) \dots p(f_n = v_n | \text{target} = i) p(\text{target} = i)$$

  但實際上為了防止數值太小的問題，我們採取 log 的方式，將乘法轉化成加法，同時因為 log 的嚴格遞增性，對於我們最後比 posterior probability 的大小結果不會有影響。
  轉化公式:

$$\log[p(\text{target} = i | f_1 = v_1, f_2 = v_2, \dots, f_2 = v_2)]$$
$$\propto \log[p(f_1 = v_1 | \text{target} = i)] + \dots$$
$$+ \log[p(f_n = v_n | \text{target} = i)] + \log[p(\text{target} = i)]$$

- 預測結果
  針對每個 example $\mathbf{f} = (v_1, v_2, \dots, v_n)$，預測其 target = *t*

$$t = \underset{i}{\text{argmax}} \{\log[p(\text{target} = i | f_1 = v_1, f_2 = v_2, \dots, f_2 = v_2)]\}$$

(3) 驗證方法

- Cross-validation
  k-fold cross-validation
  Step 1: 先計算將訓練資料集分成 k 等分同大小的資料集所對應的資料及大小
  Step 2: Stratified subsets
  計算原訓練資料 target = 0 和 target = 1 的比例，並按此比例分配 examples 到每個 subset，確保每個 subset 的比例接近原本訓練資料集的比例。
  Step 3: 輪流挑選一個 subset 當成 validation set，其餘的 subsets 當成 training set，並計算相對應的準確率。
  Step 4: 平均每次得到的準確率，獲得 overall accuracy.

- 實測結果: 10-fold cross-validation

  Accuracy: 96.6667%

  Accuracy: 96.6667%

  Accuracy: 96.6667%

  Accuracy: 93.3333%

  Accuracy: 96.6667%

  Accuracy: 93.3333%

  Accuracy: 93.3333%

  Accuracy: 96.6667%

  Accuracy: 96.6667%

  Accuracy: 100%

  Overall accuracy: 96%