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:

+∞

-∞

值域範圍

* + 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。

1. 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  
      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 與 的prior probability
   * 計算conditional probability *p*(*a* | *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
   * 計算posterior probability  
     原公式：  
     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:
   * 預測結果  
     針對每個example )，預測其target = *t*
2. 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%