AI.FINTECH
AI 金融科技中心

Naive Bayes Classifier Introduction

National Kaohsiung University of Sciences and Technology
Department of Finance and Information, Professor
AI Fintech Center, Director
Lin, Ping-Chen

Outline

- 什麼是貝氏分類器(簡單or單純貝式分類器)
- 貝式分類器的假設
- 貝式分類器的公式
- 違約機率 - 案例
- 貝式分類器的類型
- 貝式分類器的優點
- 貝式分類器的缺點

Naive Bayes Classifier Introduction

什麼是貝氏分類器

- 貝式分類器是一種用於分類任務的監督式機器學習算法
- 它們使用機率原理來執行分類任務
- 是生成學習演算法系列的一部分
- 對給定類別或範疇的輸入分佈進行建模
- 與邏輯迴歸等判別分類器不同，不會學習哪些特徵在區分類別時最重要

貝式分類器的假設

- 貝式分類器在幾個關鍵假設下運作
- 假設在貝式模型中的預測變量是條件獨立的或與模型中的任何其他特徵無關
- 假設所有特徵對結果的貢獻是相等的
- 儘管這些假設在現實中經常被違反
- 透過使分類問題在計算上更容易處理來簡化分類問題
- 只需要為每個變量計算一個機率，使模型計算更容易。
- 儘管這種不現實的獨立性假設，分類算法在特別是小樣本情況下表現良好

貝式分類器的公式

- 類似於貝式定理，使用條件機率和先驗機率來計算後驗概率，公式如下:
- 事件A發生的概率在事件B已經發生的情況下:
- P(Al B) 是給定事件B發生後事件A發生的條件機率
- P(Bl A) 是另一個條件機率，即給定事件A發生後事件B發生的機率
- P(A) 跟P(B)是事件A和B發生的機率

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

違約機率 - 案例

- 假設所有預測變量是條件獨立的
- 特徵對結果的貢獻是相等的（無權重）

公式：$\mathbf{P}(\mathbf{C}\mid F_1, F_2, \ldots, F_n) = \dfrac{P(C)\mathbf{P}(F_1, F_2, \ldots, F_n \mid C)}{\mathbf{P}(F_1, F_2, \ldots, F_n)}$

(其中C表示類別變數；F_i為特徵變數)

Sex	Age	Occupation	Income	Failure/not
Male	21~30	Student	Low	No
Male	21~30	Worker	Low	No
Male	41~50	Worker	High	No
Male	31~40	Unemployed	Low	Yes
Male	31~40	Manager	High	No
Female	21~30	Student	Low	Yes
Female	21~30	Student	Low	No
Female	21~30	Unemployed	Low	No
Female	31~40	Worker	High	Yes
Female	41~50	Manager	High	No

Sex	Age	Occupation	Income	Predicted
Male	31~40	Student	Low	?

預測一個特定條件下借款人違約的概率?

違約機率 - 案例

如何計算具有以下條件的人違約的機率？

Sex	Age	Occupation	Income	Predicted
Male	31~40	Student	Low	?

- 違約概率：$\frac{3}{10}\times\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{1}{3}=0.0148$
  - P(違約) = $\frac{3}{10}$(10筆數據中有3筆)
  - P(收入較低| 違約) = $\frac{2}{3}$(3筆違約中2筆收入較低)
  - P(學生 | 違約) = $\frac{1}{3}$(3筆為學生資料中1筆違約)
  - P(31~40歲 | 違約) = $\frac{2}{3}$(3筆31~40歲資料中2筆違約)
  - P(男性| 違約) = $\frac{1}{3}$(3比男性資料中1筆違約)
- 不違約概率：$\frac{7}{10}\times\frac{4}{7}\times\frac{1}{7}\times\frac{2}{7}\times\frac{4}{7}=0.0093$

➔0.0148 > 0.0093，預測結果為違約

Sex	Age	Occupation	Income	Failure/not
Male	21~30	Student	Low	No
Male	21~30	Worker	Low	No
Male	41~50	Worker	High	No
Male	31~40	Unemployed	Low	Yes
Male	31~40	Manager	High	No
Female	21~30	Student	Low	Yes
Female	21~30	Student	Low	No
Female	21~30	Unemployed	Low	No
Female	31~40	Worker	High	Yes
Female	41~50	Manager	High	No

貝式分類器的類型

- **高斯貝式分類（GaussianNB）**
  - 適用於高斯分佈，即正態分佈和連續變量
  - 通過找到每個類別的平均值和標準差來擬合
- **多項式貝式分類（MultinomialNB）**
  - 假設特徵來自多項分佈
  - 在使用離散數據如頻率計數時很有用，通常應用於自然語言處理的案例如垃圾郵件分類
- **伯努利貝式分類（BernoulliNB）**
  - 另一種貝式分類器變體，適用於布林變數，即只有兩個值的變量，如真和假或1和0

貝式分類器的優點

- **不太複雜**
  - 與其他分類器相比，貝式分類器被認為是一種較簡單的分類器，因為參數更容易估計
- **良好的可擴展性**
  - 與邏輯迴歸相比，貝式分類器被認為是一種快速且高效的分類器，當條件獨立性假設成立時，它相當準確
  - 具有較低的存儲需求
- **能處理高維數據**
  - 文件分類等使用案例可能具有大量的維度，這對其他分類器來說可能很困難

貝式分類器的缺點

- **受到零機率影響**
- 發生在訓練集中不存在某個類別變量時
- **不現實的核心假設**
  - 儘管條件獨立性假設總體上表現良好，但該假設並不總是成立，導致分類錯誤

Reference

- https://towardsdatascience.com/naive-bayes-classifier-explained-54593abe6e18
- https://subashpalvel.medium.com/naive-bayes-classifier-an-introduction-37a221bef754
- https://www.ibm.com/topics/naive-bayes#Types+of+Na%C3%AFve+Bayes+classifiers

Thank you.