1 Naive Bayes Classification

Let's say that we work with a dataset of n observations (rows) and m output classes where we want to classify n texts.

1.1 Definitions and Notations

Let $X = (x_1, x_2, \dots, x_n) \in \mathcal{T}^n$ be the multiset of texts where $\mathcal{T}$ is a multiset of words $(w_1, w_2, \dots, w_k)$ defining a text. Note that the position of the texts in X does not matter.

We note

$$Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,m} & \\ y_{2,1} & y_{2,2} & \cdots & y_{2,m} & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,m} & \end{bmatrix} \tag{1}$$

the matrix of binary output values (explained variables) $y_{i,j} \in \{0, 1\}$ for $1 \leq i \leq n$ and $1 \leq j \leq m$.

Since the goal is to estimate Y because we are not supposed to know $y_{i,j}$, we note

$$\widehat{Y} = \begin{bmatrix} \widehat{y}_{1,1} & \widehat{y}_{1,2} & \cdots & \widehat{y}_{1,m} & \\ \widehat{y}_{2,1} & \widehat{y}_{2,2} & \cdots & \widehat{y}_{2,m} & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widehat{y}_{n,1} & \widehat{y}_{n,2} & \cdots & \widehat{y}_{n,m} & \end{bmatrix} \tag{2}$$

the estimator matrix of Y where $\widehat{y}_{i,j} \in [0, 1]$ because we want to give a probability.

Between the estimated and the true values, there is generally a bias that we note

$$\epsilon = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,m} & \\ \epsilon_{2,1} & \epsilon_{2,2} & \cdots & \epsilon_{2,m} & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \cdots & \epsilon_{n,m} & \end{bmatrix} \tag{3}$$

where $\epsilon_{i,j} \in [-1, 1]$ because the bias may be negative or positive. If $y_{i,j} = 1$ and the model estimated $\widehat{y}_{i,j} = 0.971$, then the bias is positive because $\epsilon_{i,j} = 1 - 0.97 = 0.03$. However, if $y_{i,j} = 0$ and $\widehat{y}_{i,j} = 0.12$, then the bias is negative because $\epsilon_{i,j} = 0 - 0.12 = -0.12$.

We deduce the vectored equation

$$Y = \widehat{Y} + \epsilon \tag{4}$$

where the operator $+$ is the element-wise matrix addition.

Let $f : \mathcal{T}^n \longrightarrow \mathbb{M}_{n \times m}([0, 1])$ be a model defined by $f(X) = \widehat{Y}$ where the notation $\mathbb{M}_{n \times m}([0, 1])$ means the set of matrix n by m for which each element is a real number in $[0, 1]$.

The goal is to find a model f such that the bias ϵ is minimized when f is applied on X. Obtaining $\epsilon = \mathbf{0}_{n\times m}$ means that the model f predict perfectly how the texts will be classified.

1.2 Theoretical Problem

Let $C = \{c_1, c_2, \dots, c_m\}$ be the set of output class labels and $c \in C$ be an output class label. In virtue of the Bayes theorem, we have

$$\mathbb{P}(c|X) = \frac{\mathbb{P}(X|c)\mathbb{P}(c)}{\mathbb{P}(X)}. \tag{5}$$

The goal of the Naive Bayes Classification is to find the output class label c that maximize the probability that a text $x_i \in X$ maps to the output class label c knowing X. In other terms, this means that

$$c_{max} = \arg\max_{c\in C} \mathbb{P}(c|X) = \arg\max_{c\in C} \frac{\mathbb{P}(X|c)\mathbb{P}(c)}{\mathbb{P}(X)}. \tag{6}$$

We extract 2 properties that will simplify the equation of c_{max}:

1. $\mathbb{P}(X) = 1$ because the probability of having a text in X is always 1.
2. Two texts $x_i, x_j \in X$ where $i \neq j$ are independent. This implies that $\mathbb{P}(x_i|c)$ is independent of $\mathbb{P}(x_j|c)$.

Applying the first property gives

$$c_{max} = \arg\max_{c\in C} \mathbb{P}(X|c)\mathbb{P}(c) \tag{7}$$

which can be written equivalently using the definition of X as

$$c_{max} = \arg\max_{c\in C} \mathbb{P}(x_1, x_2, \dots, x_n|c)\mathbb{P}(c). \tag{8}$$

Now, applying the second property in (8) gives

$$c_{max} = \arg\max_{c\in C} \mathbb{P}(c) \prod_{i=1}^{n} \mathbb{P}(x_i|c). \tag{9}$$

1.3 Bag of Words Model

Let $x_i = (w_{i,1}, w_{i,2}, \dots, w_{i_k}) \in \mathcal{T}$ be a text containing k words where a word $w_{i,j} \in W$ the set of words contained in X. Note that we assume that a text cannot be empty meaning that $\mathcal{T} \neq \emptyset$.

We want to use the maximum likelihood estimator $\widehat{P}$ defined as the frequency of a word $w_{i,j}$ among the n texts where $1 \leq j \leq k$ knowing the output class label $c \in C$. The estimator $\widehat{P}$ estimates the likelihood function $P(x_1, x_2, \dots, x_n; c) = \prod_{i=1}^{n} \mathbb{P}(x_i|c)$.

To calculate that frequency, we have to calculate the ratio between the number of occurrences of the word $w_{i,j}$ among the n texts, where the output class label is c, and the total number of words in the n texts where the output class label is c.

Let $f : W \times C \longrightarrow \mathbb{N}$ be a function defined as $f(w_{i,j}, c) = z$ that returns the number of occurrences (z) a word $w_{i,j}$ is found among all texts classified as the output class label c.

Therefore, the maximum likelihood estimator of $P(x_1, x_2, \ldots, x_n; c)$ is defined as

$$\widehat{P}(w_{i,j} \in X|c) = \frac{f(w_{i,j}, c) + 1}{\sum\limits_{w \in W} f(w, c) + 1}. \tag{10}$$

We also need the maximum likelihood estimator of $P(c) = \mathbb{P}(c)$ which is defined as the ratio between the number of texts classified as c and the number of texts n. Let $C_c = \{x_i \in X : x_i \mapsto c\}$ be the set of all texts x_i classified as c.

We note $|C_c|$ the cardinality of C_c. The estimator is defined as

$$\widehat{P}(c) = \frac{|C_c|}{n}. \tag{11}$$

The reason behind the Laplace smoothing, adding 1 to the numerator and denominator of $\widehat{P}(w_{i,j} \in X|c)$, is to handle the case when $f(w_{i,j}, c) = 0$. If a word $w_{i,j}$ is not found for a given output class label c, then $\widehat{P}(w_{i,j} \in X|c) = 0$. Having only one case like this without adding 1 causes

$$\widehat{P}(c) \prod_{i=1}^{n} \widehat{P}(w_{i,j} \in X|c) = 0. \tag{12}$$

1.4 Example