1 Naive Bayes Classification

1.1 Definitions and Notations

Let $X = (x_1, x_2, \dots, x_n) \in \mathcal{C}^n$ be the vector of comments where $\mathcal{C}$ is a vector of words $(w_1, w_2, \dots, w_k)$ defining a comment and n is the length of X which is the number of comments in the data set. The position of the comments in X does not matter. We note the category vector as $Y = (y_1, y_2, \dots, y_6) \in \mathbb{M}_{n\times 6}([0,1])$ where each of them is predicted by the estimator vector $\hat{Y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_6) \in \mathbb{M}_{n\times 6}([0,1])$ associated to a bias vector $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_6) \in \mathbb{M}_{n\times 6}([0,1])$. The notation $Y \in \mathbb{M}_{n\times 6}([0,1])$ means that Y is a $n \times 6$ matrix where each value $Y_{i,j}$ is in $[0,1]$. Finally, let $C = (c_1, c_2, \dots, c_6)$ be the vector of class labels.

Let $f : \mathcal{C}^n \longrightarrow \mathbb{M}_{n\times 6}([0,1])$ be a model defined by $f(X) = \hat{Y}$ where $Y = \hat{Y} + \epsilon$ with $+$ the matrix addition operator. The goal is to find a model f such that the bias ϵ is minimized when f is applied on X.

1.2 Theorical Problem

Let $c \in C$ be a class label. In virtue of the Bayes theorem, we have

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

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

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

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

1. $\mathbb{P}(X) = 1$ because having a comment in X is always true.
2. Two comments $x_i, x_j \in X$ 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

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

which can be written equivalently using the definition of X as

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

Now, applying the second property gives

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

1.3 Bag of Words Model

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

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

To calculate that frequency, we have to calculate the ratio between the number of occurences of the word $w_{i,j}$ among the n comments, where the class label is c, and the total number of words in the n comments where the 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 occurences (z) a word $w_{i,j}$ is found among all comments classified as the class label c.

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

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

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

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

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

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 class label c, then $\widehat{P}(w_{i,j} \in X|c) = 0$. Having only one case like this without adding 1 causes

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