

# Laboratório N° 7: Naive Bayes

---

Extração Automática de Informação  
2021/2022

Prof. Joaquim Filipe  
Eng. Filipe Mariano

## Objetivos

- Criar um classificador probabilístico de acordo com a teoria de decisão de Bayes.

## 1. Introdução

No laboratório anterior foi introduzida uma primeira abordagem a aplicar na classificação de documentos na aprendizagem supervisionada, através da identificação da similaridade entre documentos recorrendo à métrica da similaridade do cosseno. Neste laboratório irá ser introduzida uma nova forma de classificação de documentos através de métodos probabilísticos, assentes na teoria de decisão de Bayes.

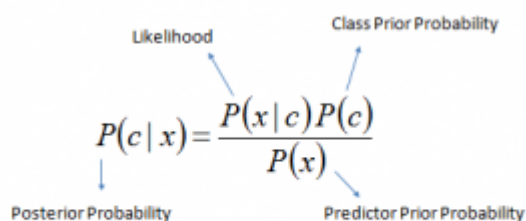
## 2. Teoria de Decisão de Bayes

A teoria de decisão de Bayes é uma abordagem estatística utilizada nos problemas de classificação em reconhecimento de padrões. É considerada uma das melhores abordagens quando a estrutura de probabilidade subjacente às categorias é conhecida perfeitamente. Apesar desse tipo de situação ocorrer com pouca frequência, permite determinar um classificador com o qual poderá ser comparado com outros.

Esta abordagem baseia-se na quantificação das compensações entre várias decisões de classificação usando a probabilidade e os custos que acompanham tais decisões. Parte do pressuposto de que o problema de decisão é colocado num modo probabilístico, e que todos os valores de probabilidade relevantes são conhecidos.

### 2.1. Naive Bayes para Texto

As características são representadas pelos termos presentes na *bag of words*, em que de acordo com a teoria de Bayes pode ser descrita pela seguinte fórmula:


$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Se considerarmos o  $x_1, x_2, \dots, x_n$  os termos da *bag of words* podemos observar que a fórmula é composta pelo produto da probabilidade do termo  $x_i$  na classe  $C$  com a probabilidade dessa classe.

A probabilidade da classe, descrita na fórmula por  $P(C)$  pode ser calculada a partir das observações realizadas, ou seja, através do conjunto de treino.

## 2.2. Exemplo Prático 1

Se considerarmos um conjunto de treino com 9 documentos em que 5 são para a classe **positive** e 4 para a classe **negative**, podemos concluir que a probabilidade a priori é  $P(\text{positive}) = 5/9$  e  $P(\text{negative}) = 4/9$ .

Agora que compreendemos como se calcula a probabilidade a priori da classe, iremos calcular a verosimilhança através da *bag of words*. Se imaginarmos a frase (para esta análise não irei retirar o **and** como stopword):

d1: Very good food and service.

Considerando os seguintes valores para o conjunto de treino:

Probability of  
Test Word  $j$

=

$$\frac{\text{counts of word "j" in class c}}{\text{counts of words in class c}}$$

		Count of Test Word "j" in Class c				
Class c		very	good	food	and	service
Positive Class		1	2	1	0	0
Negative Class		1	0	3	0	0

		Probability of Test Word "j" in Class c				
Class c		very	good	food	and	service
Positive Class		0.038461538	0.0769231	0.0384615	0	0
Negative Class		0.047619048	0	0.1428571	0	0

		product ( p of a test word "j" in class positive c ) = p (very) * p (good) * p (food) * p (and) * p (service)				
Class c		very	good	food	and	service
Positive Class		0.038461538	0.0769231	0.0384615	0	0
Negative Class		0.047619048	0	0.1428571	0	0

Deparamo-nos com uma situação que iremos multiplicar por 0, o que tornaria o resultado da nossa probabilidade também em 0. Desta forma, poderemos optar por uma abordagem mais simplista de se ignorar

essas palavras considerando apenas que estão contidas na *bag of words*. Caso contrário, terá de seguir uma abordagem denominada de **Laplace Correction** de modo a atribuir um valor muito baixo para essas palavras que ocorrem no documento que se pretende classificar, mas que não estavam contidas no conjunto de treino ( $|V|$  representa o número de palavras únicas existente em todo o conjunto de treino, independentemente da classe).

Probability of  
Unknown Test Word  $j$

=

$$\frac{0 + 1}{\text{counts of words in class } c + |V|}$$

No entanto, nesta perspetiva todos os termos do vocabulário deverão ser incrementados em **1** para ir ao encontro da fórmula das palavras desconhecidas.

Count of Test Word "j" in Class c + 1					
Class c	very	good	food	and	service
Positive Class	1 + 1	2 + 1	1 + 1	0 + 1	0 + 1
Negative Class	1 + 1	0 + 1	3 + 1	0 + 1	0 + 1

Count of Test Word "j" in Class c + 1					
Class c	very	good	food	and	service
Positive Class	2	3	2	1	1
Negative Class	2	1	4	1	1

Probability of Test Word "j" in Class c					
Class c	very	good	food	and	service
Positive Class	0.03389831	0.05084746	0.03389831	0.01694915	0.01694915
Negative Class	0.03703704	0.01851852	0.07407407	0.01851852	0.01851852

Agora já é possível fazer o produto de todos os termos porque já não existe valores a **0**. Na seguinte figura irá aparecer o resultado para cada classe da  $P(X | C)$ .

	product ( $p$ of a test word " $j$ " in class $c$ ) = $p$ (very) * $p$ (good) * $p$ (food) * $p$ (and) * $p$ (service)									
Positive Class	0.033898305	*	0.050847458	*	0.033898305	*	0.016949153	*	0.016949153	= 1.6785E-08
Negative Class	0.037037037	*	0.018518519	*	0.074074074	*	0.018518519	*	0.018518519	= 1.74229E-08

Multiplicando pela probabilidade a priori calculada anteriormente obtemos o resultado descrito no quadro seguinte:

$$p(i \text{ belonging to class } c) = \text{product}(p \text{ of a test word } "j" \text{ in class } c) * p \text{ of class } c$$

Positive Class	1.6785E-08	* 5 / 9	9.33E-09
Negative Class	1.74229E-08	* 4 / 9	7.74E-09

De acordo com os cálculos realizados, a  $P(\text{positive} \mid d1) > P(\text{negative} \mid d1)$ . O que se pode concluir que o **d1** que correspondia à frase **Very good food and service** é da classe **positive**.

### 3. Exercícios

1. Criar uma função que permita calcular a  $P(w)$ . Sendo que para o problema que temos vindo a trabalhar, apenas temos 2 classes: **happy** e **not happy**.
  - a. Esta função deverá receber como parâmetro de entrada qual a classe e filtrar a tabela **trainingset** por essa classe. Caso não seja passado nenhum parâmetro de entrada deverá retornar todos os resultados dessa tabela.
  - b. Para calcular a probabilidade a priori da classe  $P(w)$  terá de fazer a chamada da função anterior para a classe e sem parâmetro de entrada, dividindo o número de linhas da primeira chamada pelo número de linhas da segunda chamada.
  - c. Exportar esses valores no módulo **train** para serem utilizados no cálculo da  $P(W \mid X)$ , de acordo com a teoria de Bayes.
2. No módulo **classifier** criar uma função **classify** que irá realizar os cálculos necessário para identificar qual a classe do texto recebido como parâmetro de entrada. Pode recorrer à ocorrência do termo, ou substituir nas fórmulas o **count** por soma dos **tf-idf**. Ou seja, a soma dos **tf-idf** de todos os documentos do conjunto de treino para o termo **j** na classe **c** a dividir pela soma de todos os **tf-idf** da classe **c**.