Instituto Superior Técnico

# Aprendizagem Automática

## 4º Trabalho laboratorial
## Classificador de Bayes

*Autores :*
Daniel Rosa - 90041
Guilherme Viegas - 90090

Grupo 5

# 1 Bayes Classifiers

Os classificadores de Bayes e Naive Bayes utilizam probabilidades para decidir se determinado item pertence a determinada classe.

Primeiramente, o classificador de Bayes rege-se pela expressão situada na equação 1 (Lei de Bayes), onde X $\in [x_1, x_2, ..., x_n]$ é um dataset a estudar.

$$P(classe|X) = \frac{P(X|classe)P(classe)}{P(X)} \tag{1}$$

Seguidamente, após se calcular as probabilidades de o dataset pertencer a cada uma das classes, calcula-se o máximo das probabilidades e atribui-se o dataset a essa classe.

Agora para o Naive Bayes, faz-se a assumção que os eventos são independentes, isto é, a probabilidade de o item $x_i$ sabendo que é da classe z é independente da probabilidade de $x_j$ sabendo que é da classe z. Deste modo, simplifica-se o cálculo destas probabilidades, fazendo o que está expressado na equação 2.

$$P(classe|X) = P(classe)\prod_i \frac{P(x_i|classe)}{P(xi)} \tag{2}$$

Como o denominador se mantém constante para todas as classes, é possível desprezar-se, ficando

$$P(classe|X) \propto P(classe)\prod_i P(x_i|classe) \tag{3}$$

Posteriormente, como no classificador de Bayes, encontra-se qual a probabilidade maior para este dataset e atribui-se esse dataset como sendo dessa classe.

A maior diferença entre estes dois classificadores, é o facto de o Naive Bayes considerar que os eventos são independentes, desprezando toda e qualquer influência que os mesmos têm entre si. Com isto, apesar de não se verificar esta independência, o classificador de Naive Bayes é capaz de ser bastante preciso.

O Naive Bayes tem a vantagem de ter um menor custo computacional em comparação ao classificador de bayes, mas também de ser mais fácil de implementar e de necessitar de um training set com poucos dados para estimar o test set. No entanto, para dados que apareçam no conjunto de teste que não apareçam no conjunto de treino, este classificador atribui-os com probabilidade de 0, sendo necessário usar uma técnica de atenuação, como Laplace Smoothing.

# 2 A simple example

## 2.7

Observando ambos *accuracy scores* para os classificadores de Naive Bayes e Bayes, 94.67% e 96.67%, respetivamente, verifica-se que são muito próximos e, por isso, será preferível utilizar o Naive Bayes, pois é computacionalmente mais barato e a diminuição do *accuracy score* é pouco significativa.

# 3 Practical assignment

## 3.2 Testing

### 3.2.6

| Text | Real Language | Recognized Language | Score | Classification marg |
|---|---|---|---|---|
| Que fácil es comer peras | es | es | 0.670347 | 0.340693 |
| Que fácil é comer peras | pt | pt | 1 | 1 |
| Today is a great day for sightseeing | en | en | 1 | 1 |
| Je vais au cinéma demain soir | fr | fr | 1 | 1 |
| Ana es inteligente y simpática | es | es | 0.999971 | 0.999943 |
| Tu vais à escola | pt | fr | 0.79305 | 0.586101 |

### 3.2.7

Começando pela primeira frase, observa-se uma correta classificação, mas observa-se uma margem de erro relativamente pequena, o que nos indica que existe outra língua (pt) cuja respetiva frase apresenta trigramas idênticos à mesma na língua espanhola. As quatro frases seguintes encontram-se bem classificadas e com um *score* e uma *classification margin* bastante confortáveis.

Para a última frase verifica-se uma classificação incorrecta, isto pode dever-se a vários motivos, entre eles, um *dataset* de treino reduzido e principalmente a uma frase de teste muito pequena, ou seja composta por poucos trigramas. Para além disso a utilização de um algoritmo de Naive Bayes, que parte do pressuposto que a probabilidade da ocorrência de trigramas é independente entre si, leva a que, numa frase pequena, se façam assunções e predições erradas. Apesar das primeiras duas frases, em espanhol e português, serem parecidas, a frase em português tem características específicas desta língua, nomeadamente acentos no caractere 'e' que não se verificam em espanhol. Para a última frase, todos os trigramas constituintes existem em abundância em ambas as línguas, justificando também o erro na predição.

Uma forma de melhorar seria passar a usar palavras em vez de trigramas como datasets, pois são muito mais específicos de cada língua e aumentar o número de dados de treino.