**Instituto Superior Técnico**

MEEC

Aprendizagem Automática

# Lab 4

## Bayes Classifier

**Grupo 9**

Manuel Diniz, 84125
Alexandre Rodrigues, 90002

**Turno:** 4ªf 11h00

# Contents

# Chapter 1

# Classificador de *Bayes*

O classificador de *Bayes* baseia-se num conjunto de classes que são utilizadas para prever valores de características de membros da própria classe. A ideia por detrás do classificador de *Bayes* é, se se souber previamente qual a classe, conseguir prever os valores de outras características da mesma classe. Se não se souber a classe, o terorema de Bayes (1.1) pode ser usado para a prever sabendo alguns dos valores das suas caracteristicas.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{1.1}$$

O classificador de *Bayes* é utilizado para criar um modelo probabilístico com base num conjunto de dados de treinos e, posteriormente, utilizado para prever a classificação de um novo conjunto teste.

O classificador de *Naive Bayes* é uma simplificação de classificador de Bayes, que assume que as variáveis sejam todas independentes umas das outras dada uma classe ($i$ classes).

$$P(Yi|X) = P(Yi)P(X|Yi) \tag{1.2}$$

Para o classificador de *Naive Bayes*, estima-se a densidade de probabilidade de um conjunto de dados. Para cada densidade faz-se uma distribuição normal (*Gaussiana*), com os valores para a média (*mean*) e desvio padrão ($\sigma$) calculados através dos mesmos (1.3)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x - mean)^2}{2\sigma^2}) \tag{1.3}$$

# Chapter 2

# Exemplo

Para demonstrar os classificadores de *Bayes* e *Naive Bayes*, recorreu-se a um exemplo simples com um conjunto de dados de treino com 3 classes e um outro conjunto de dados que posteriormente foi testado.

Os dados de treino e teste, já agrupados em 3 classes, estão apresentados na Figura 2.1 e 2.2.

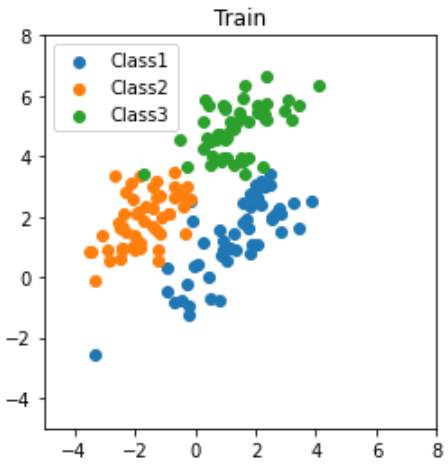


Figure 2.1: Scatter dos Dados de Treino

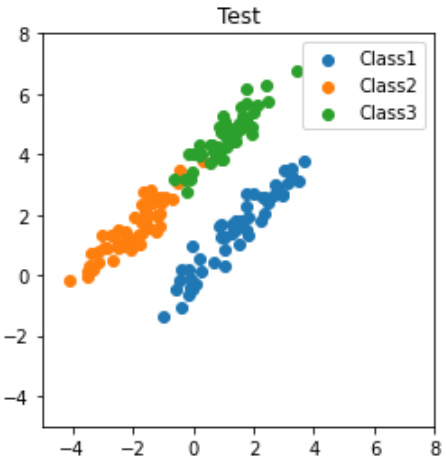


Figure 2.2: Scatter dos Dados de Teste

Desenvolveu-se um algoritmo que executa o método dos classificadores de *Naive Bayes* e obteve-se um novo gráfico dos dados de teste.

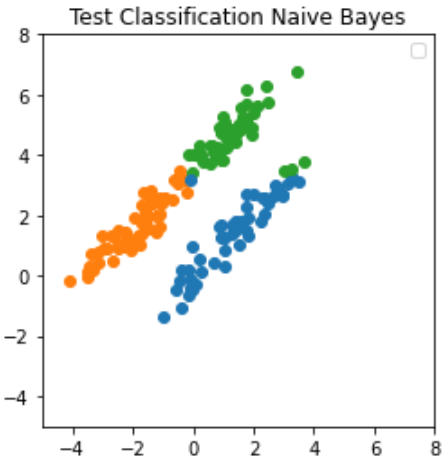


Figure 2.3: Scatter dos Dados de Teste (*Naive Bayes*)

Analisando e comparando com os dados de teste representado anteriormente, verifica-se que alguns dos pontos foram mal classificados. No entanto, este modelo apresenta uma percentagem de erro baixa, cerca de 5.3(3)%.

Seguidamente, criou-se um algoritmo para executar o método dos classificadores de *Bayes*.

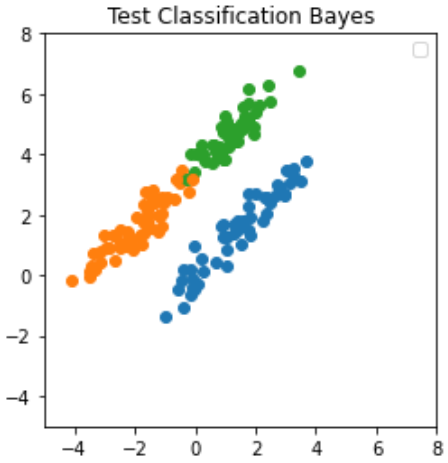


Figure 2.4: Scatter dos Dados de Teste (*Bayes*)

Comparando novamente com os dados de teste, o mesmo se verifica que no modelo anterior, alguns dos pontos foram mal classificados. A percentagem de erro deste modelo é de 3.3(3)%.

Verifica-se, por isso, que o método dos classificadores de *Bayes* tem uma melhor precisão.

# Chapter 3

# Reconhecimento de Linguagem

É agora aplicado um *naive Bayes classifier* a texto, de modo a se fazer o reconhecimento da linguagem em que este está escrito. O conjunto de treino é um conjunto de trigramas e o respetivo número de ocorrências nos textos originais.

## 3.1 Geração da matriz de treino

A partir dos dados, é gerada uma matriz de treino, com uma amostra (neste caso uma linguagem) em cada linha, e uma *feature* (neste caso, o número de ocorrências de um determinado trigrama) por coluna. Gera-se também um vetor com as *labels* corretas correspondentes às amostras.

## 3.2 Treino e teste

De seguida, o modelo é treinado e testado com os mesmos dados de treino, verificando que atribui a linguagem correta a cada conjunto de treino.

Testa-se agora o modelo em 6 frases fornecidas. Estas são processadas de modo a serem decompostas nos seus trigramas, e gera-se um vetor linha com o mesmo formato dos dados de treino, ou seja, com o número de ocorrências de cada trigrama em cada coluna (fazendo correspondência posicional entre os trigramas obtidos e os do conjunto de treno). Os resultados da previsão neste conjunto de dados estão na tabela

| Texto | Linguagem real | Linguagem reconhecida | *Score* | Margem de classificação |
|---|---|---|---|---|
| Que fácil es comer peras. | es | es | 0.6703 | 0.3407 |
| Que fácil é comer peras. | pt | pt | 0.9999 | 0.9999 |
| Today is a great day for sightseeing. | en | en | 1.0000 | 1.0000 |
| Je vais au cinéma demain soir. | fr | fr | 0.9999 | 0.9999 |
| Ana es inteligente y simpática. | es | es | 0.9999 | 0.9999 |
| Tu vais à escola hoje | pt | fr | 0.7930 | 0.5861 |

Todas as frases são identificadas corretamente com a exceção da última. A primeira é identificada com um *score* modesto, pois não existem traços fortes que a língua em questão seja espanhol e não português. A única diferença que a frase tem da seguinte é a palavra "es", forma os trigramas " es" e "es ", que também são bastante comuns na língua portuguesa ("estar" ou "antes", por exemplo).

O mesmo não é o caso na segunda frase, que tem um forte indicador da língua portuguesa, o trigrama " é ", que ocorre muito menos vezes no conjunto de treino espanhol. Deste modo, classifica a frase como português com uma confiança, ou *score*, muito mais elevada.

Mais uma vez a terceira frase é classificada com elevada confiança, devido à presença de trigramas que são fortes indicadores da sua língua, como "ght" ou "day" neste caso.

A quarta frase também possui *score* elevado, com um dos trigramas chave sendo "oir", muito comum no francês.

A quinta possui um trigrama quase do exclusivo do espanhol, " y ", sendo que também é classificada com elevada segurança.

A sexta e última frase é identificada incorretamente como francês, se bem que com um *score* e margem de classificação reduzidos. Pode-se atribuir parcialmente à presença do trigrama " à ", que é muito comum no francês, para além de que quase todos os restantes trigramas são partilhados pelas duas linguagens de forma comum.