Instituto Superior Técnico
Mestrado Integrado em Engenharia Electrotécnica e de Computadores

Aprendizagem Automática

2020/2021 - 1º Semestre

Laboratório 4

Grupo n.º 4
Simão Gonçalves - 90193
Miguel Amaral - 90150

Professora: Maria Margarida Campos da Silveira
Turno: Terça-feira, 11h00

1 Bayes Classifiers

1.1 Explicação dos classificadores de Bayes e naive Bayes

´

Um classificador f(x) divide o espaço de entrada de dimensão R^d em K regiões disjuntas R_j.

$$R_j = \left\{x \in R^d : f\left(x\right) = \omega_j\right\} \tag{1}$$

Onde $\omega_j, j \in \{1, \ldots, K-1\}$, é a classe associada à região R_j.

No caso de a função de perda associada à estimação da região à qual pertence determinada entrada x ser binária pode-se recorrer a um estimador do tipo

$$\hat{y} = \underset{\omega \in \Omega}{argmax} P(\omega|x) \tag{2}$$

Este estimador é um estimador de Bayes pois associa-se um input (x) à classe que tem uma maior probabilidade a posteriori, ou seja $P(\omega|x)$ é a distribuição das classes depois de se observar o vetor de features x. A probabilidade a posteriori para cada classe pode ser calculada através do teorema de Bayes.

$$P\left(\omega_i|x\right) = \frac{p\left(x|\omega_i\right) * P(\omega_i)}{p(x)} \tag{3}$$

Neste teorema $p\left(x|\omega_i\right)$ é a distribuição do vetor de features x associada à classe$\omega_i, P(\omega_i)$ a probabilidade a priori das classes e p(x) um termo de normalização que não influencia na decisão da classe à qual x pertence. Por vezes o vetor $x = [x_1, \ldots, \ x_p]^T$ contém muitas features, tornando-se difícil obter uma estimativa da distribuição $p\left(x|\omega_i\right)$. Para resolver situações deste tipo recorre-se a outro classificador, designado por Naive Bayes, neste pressupõe-se que as p features são todas condicionalmente independentes, podendo a distribuição $p\left(x|\omega_i\right)$ ser expressa da seguinte forma:

$$p\left(x|\omega_i\right) = \prod_{i=1}^{p} p\left(x_i|x_1, \ldots, \ x_{i-1}, \omega_k\right) = \prod_{i=1}^{p} p\left(x_i|\omega_k\right) \tag{4}$$

Assim torna-se apenas necessário calcular a distribuição condicional para cada feature individualmente. Recorrendo novamente ao teorema de Bayes calcula-se a distribuição a posteriori das classes.

2 A simple example

2.1 Resultados de ambos os classificadores

Começa-se por verificar a classificação dos dados em função dos resultados esperados, sempre que é observada uma esfera vermelha, esta indica um desvio em relação à reta verde que se encontra na mesma coordenada x:

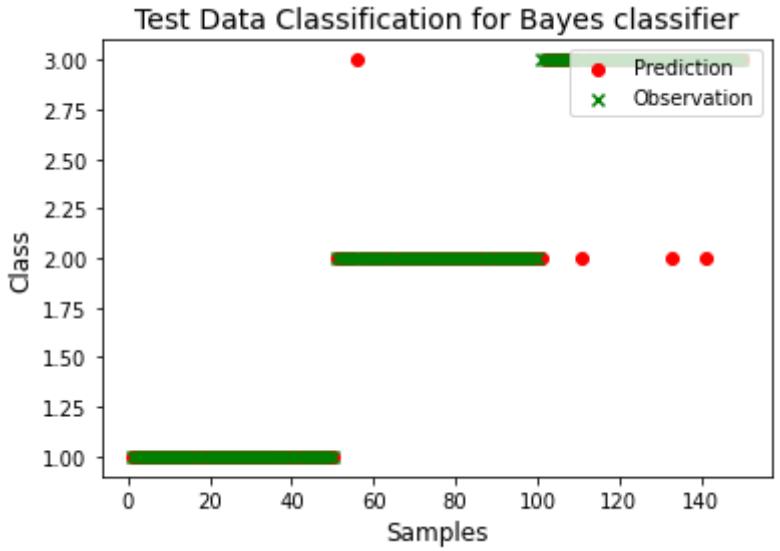

Figure 1: Classificação dos dados em função dos índices do vetor de informação, quando aplicado o *Bayes Classifier*

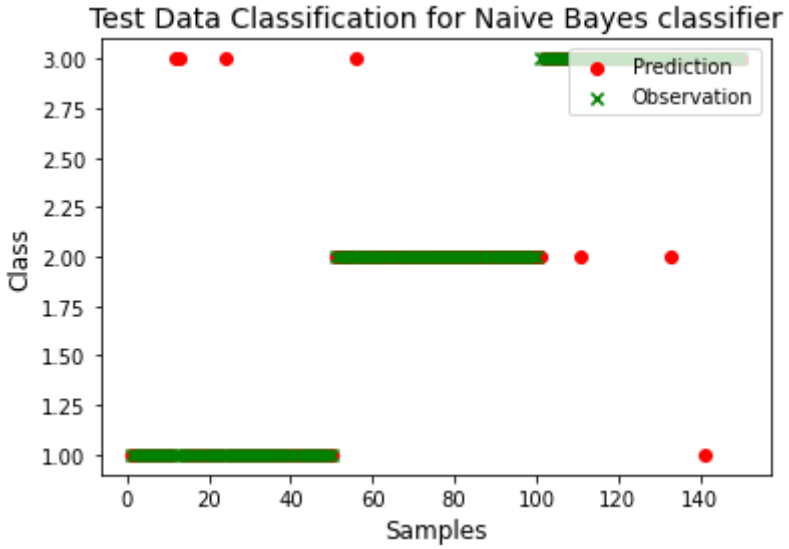

Figure 2: Classificação dos dados em função dos índices do vetor de informação, quando aplicado o *Naive Bayes Classifier*

É observada uma menor quantidade de erros no Bayes Classifier (3.33%), quando comparado ao Naive Bayes Classifier (5.33%).

Tal deve-se às features no Naive Bayes Classifier serem consideradas condicionalmente independentes (possuem covariância nula), contudo este não é o caso. Por esse motivo o Bayes Classifier possui uma maior imunidade ao erro tornando-o o método de classificação preferível para a informação tratada.

Para além disso a maior parte dos erros encontra-se presente entre as classes 2 e 3 onde existe uma maior sobreposição de dados, podendo gerar inconsistências nas classficações, como pode ser observado na figura seguinte:

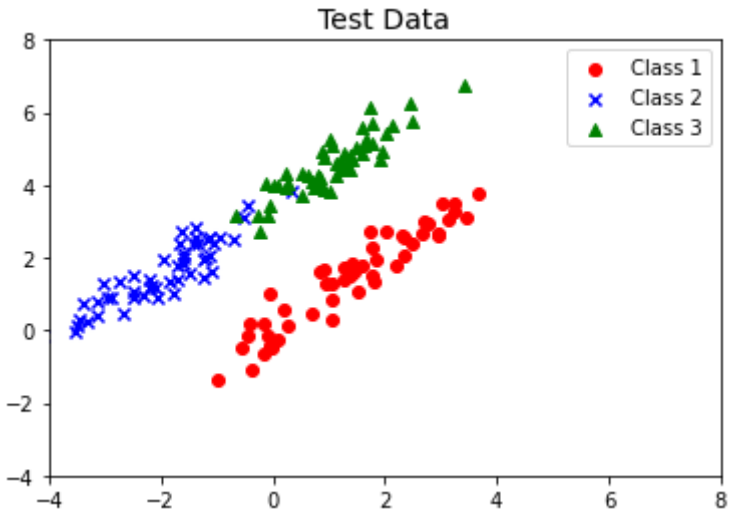

Figure 3: Data de teste

3 Pratical Assignment

3.1 Predictions, Score and Classification Margin

São apresentados os resultados obtidos através da apicação do classifcador de naive Bayes às 6 frases apresentadas.

Text	Real language	Recognized language	Score	Classification margin
Que fácil es comer peras.	es	es	0.67	0.34
Que fácil é comer peras.	pt	pt	0.99	0.99
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	0.99	0.99
Tu vais à escola hoje.	pt	fr	0.79	0.59

3.2 Análise de cada uma das frases

Frase 1) "Que fácil es comer peras"

Como "es" é um verbo pertencente à lingua espanhola a presença deste nos trigramas será elevada permitindo a correta identificação da frase. Contudo as restantes palavras da frase são comuns às línguas portuguesa e espanhola, o que causa o baixo score desta classificação.

Frase 2) "Que fácil é comer peras"

A frase apresentada é bastante semelhante à frase 1, contudo o verbo "é" é exclusivo à língua portuguesa e dessa forma garante uma probabilidade próxima de 100% da palavra pertencer a lingua portuguesa, aparecendo esta 1966239 vezes na nossa data.

Frase 3) "Today is a great day for sightseeing" Um dos trigramas mais presentes na língua inglesa é a palavra "for" e a terminação "ing" ou "ng" bastante comum na conjugação dos verbos, facilitando mais uma vez a identificação da linguagem. Mais uma vez o score é próximo de 100%

Frase 4) "Je vais au cinéma demain soir"

Nesta expressão encontram-se trigramas populares da língua francesa tais como "au" (uma proposição) e "in", também existe uma quantidade vasta de verbos que utilizam o trigrama "oir". Mais uma vez a certeza do classificador é quase total.

Frase 5) "Ana es inteligente y simpática."

Apesar de semelhante à frase 1, esta frase é classificada como pertencente à língua com uma certeza próxima de 100%, devido ao acréscimo de um monograma bastante popular na conjunção "y".

Frase 6) "Tu vais à escola hoje."

Esta frase foi a única classificada de forma incorreta, isto deve-se à quantidade de palavras partilhadas entre a língua francesa e portuguesa que se encontram presentes nesta frase. De forma a corrigir este erro, a frase pode ser restruturada para "Tu deslocaste-te à escola hoje" ou uma conjunção na primeira pessoal do singular tornando a identificação da frase mais evidente.