Instituto Superior Técnico

Aprendizagem Automática

Laboratório 4 - Classificadores Naive Bayes

Autores:
Diogo Moura - nº 86976
Diogo Alves - nº 86980

Turno:
Terça 17h-18h30m

12 de Novembro de 2019

1)

1.1

Dada uma observação $x \in \mathbb{R}^d$ o classificador atribui uma classe $y \in \Omega$ a essa observação, $\Omega = \{w_0, ..., w_{k-1}\}$, isto é, uma de k classes.

$$X = (x_1, x_2, x_3, ..., x_n) \tag{1}$$

Para conseguir atinjir este objetivo, o classificador de Bayes separa o espaço dos vetores de entrada $x \in \mathbb{R}^d$ em k regiões dijuntas $\mathbb{R}_j$, $j \in \{0, ..., K-1\}$, cada uma delas associada a uma classe específica w_j, conhecidas como regiões de decisão:

$$R_j = \{x \in \mathbb{R}^d : f(x) = \omega_j\} \tag{2}$$

Conhecer as regiões de decisão é equivalente a conhecer o classificador $f(x)$.

É necessário construir uma função de custo para treinar o classificador.
Esta função tem diversos custos associados as pares $(y_a, \hat{y_b})$ e atribui uma penalização caso o classificador classifique mal uma determinada feature numa classe.
Numa função de perda binária, existe um custo de 1 associado ao erro e 0s na diagonal da matriz de custos (decisões sem erro).

Tabela 1: Exemplo de matriz de custos de uma função de perda binária

	$\hat{y_0}$	$\hat{y_1}$	$\hat{y_2}$
y_0	0	1	1
y_1	1	0	1
y_2	1	1	0

Numa função de perda geral, existem custos diferentes de 0 e diferentes de 1 associados aos tipos de erros e 0s na diagonal da matriz de custos (decisões sem erro).

Tabela 2: Exemplo de matriz de custos de uma função de perda geral

	$\hat{y_0}$	$\hat{y_1}$	$\hat{y_2}$
y_0	0	3	2
y_1	1	0	1
y_2	3	1	0

O classificador é a função que provoca a minimização do risco R, que equivale ao valor esperado da perda:

$$R = E\{L(y, \hat{y}(X))\} = \int \sum_{y \in \Omega} L(y, \hat{y}(X)) p(X, y) dX = \int [\sum_{y \in \Omega} L(y, \hat{y}(X)) P(y|X)] p(X) dX \tag{3}$$

$$f(X) = \underset{\omega_j \in \Omega}{argmin} [\sum_{y \in \Omega} L(y, \omega_j) P(y|X)] \tag{4}$$

Se a função de perda for binária:

$$f(x) = \underset{\omega_j \in \Omega}{argmin} [1 - P(\omega_j|X)] = \underset{\omega_j \in \Omega}{argmax} P(\omega_j|X) \tag{5}$$

O teorema de Bayes pode ser utilizado para calcular as probabilidades, que se pretendem maximizar $P(w_j|x)$, tendo-se nesse caso um classificador de Bayes:

$$P(\omega_j|X) = \frac{P(X|\omega_j) P(\omega_j)}{P(x)} \tag{6}$$

Na simplificação de Naive Bayes, assume-se que os acontecimentos são independentes, isto é que a probabilidade de um acontecimento não está dependente da a dos outros acontecimentos (observações), tendo-se nesse caso um classificador de Naive Bayes:

$$p(X|w_j) = p(x_1, ..., x_p|\omega_j) = \prod_{i=1}^{p} p(x_i|x_1, ..., x_i - 1, \omega_j) = \prod_{i=1}^{p} p(x_i|\omega_j) \tag{7}$$

Assim, temos que para o classificador de Naive Bayes. no caso de função de perda binária:

$$f(x) = \underset{\omega \in \Omega}{argmax} p(\omega|X) = \underset{\omega \in \Omega}{argmax} \prod_{i=1}^{p} \frac{p(x_i|\omega_k)p(\omega_k)}{p(x)} \tag{8}$$

Neste caso, sabendo p(x) e $p(\omega_k)$ conseguimos calcular as regiões de decisão do classificador.

Poderá ser necessário usar Suavização de LaPlace caso exista um tipo de feature que não tenha nenhuma ocorrência no conjunto de treino, o que originaria uma probabilidade de 0, usando uma aproximação de naive bayes. Neste caso, a Suavização de LaPlace acrescenta 1 à frequência de todas as observações do conjunto de treino.
Também poderá ser prático e até mesmo necessário usar o logaritmo para calcular estas probabilidades, dado que o produto das mesmas tende muito rapidamente para 0, dado que as probabilidades são menores que 1. Nesse caso:

$$f(x) = \underset{\omega_j \in \Omega}{argmax} \ log(p(\omega_j|x)) = \underset{\omega_j \in \Omega}{argmax} \sum_{i=1}^{p} log(\frac{p(x_i|\omega_j)p(\omega_j)}{p(x)}) \tag{9}$$

A probabilidade não normalizada é suficiente para obter o classificador, dado que dividi-la por uma probabilidade constante p(x) não tem influência no resultado:

$$p(x, w_j) = p(x|w_j)P(w_j) \tag{10}$$

Caso se use a probabilidade não normalizada:

$$f(x) = \underset{\omega \in \Omega}{argmax} \prod_{i=1}^{p} p(x_i|\omega_j)P(\omega_j) \tag{11}$$

ou

$$f(x) = \underset{\omega \in \Omega}{argmax} \sum_{i=1}^{p} log(p(x_i|\omega_j)p(w_j) \tag{12}$$

2)

2.6

Figura 1: Classificação dos conjuntos de treino e teste

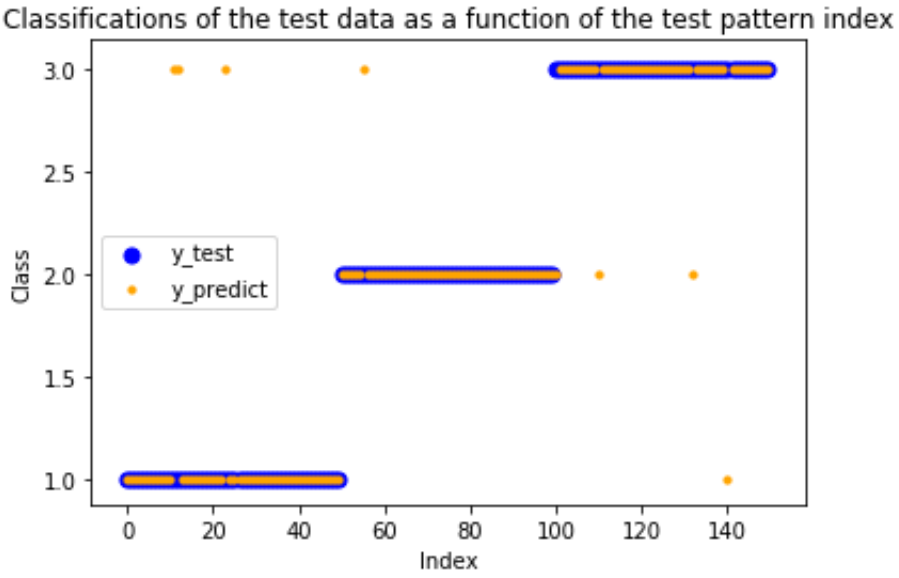

Figura 2: Classificação dos dados nas classes em função do índice do vetor

A percentagem de erros é bastante reduzida (5.33%). Apenas 8 pontos num total de 150 foram classificados incorretamente.
A maioria dos pontos que foram classificados incorretamente pertenciam à classe 1 e foram classificados na classe 3 ou pertenciam à classe 3 e foram classificados na classe 1. Este fenómeno acontece porque existe uma maior interseção entre estes dois conjuntos, como se pode ver na figura 1. Dado que os conjuntos não são separáveis por uma linha, este resultado é bastante aceitável.
Nem um classificador mais avançado que o de Naive Bayes, como uma Support Vector Machine, conseguiria classificar corretamente todos os dados.

3.2)

3.2.2.6

Tabela 3: Margem de Classificação

Text	Real Language	Recognized Language	Score	Classfication Margin
El cine esta abierto	es	es	0.999777	0.999555
Tu vais à escola hoje	pt	fr	0.793050	0.586101
Tu vais à escola hoje pois já estás melhor	pt	pt	0.999999	0.999999
English is easy to learn	en	en	0.999999	0.999999
Tu vas au cinéma demain matin	fr	fr	0.999999	0.999998
É fácil de entender	pt	es	0.548361	0.096729

3.2.2.7

Caso 1) "El cine está abierto":
Neste caso, como a palavra 'El' é um artigo definido da lingua espanhola, tem elevada frequência nos trigramas dessa mesma língua. No nosso conjunto de treino, tem uma frequência de 22187589, num total de 3483059442, o que equivale a uma frequência de 0.6%, que é bastante elevada, quando comparada com as outras línguas.
Outros trigramas também podem ter contribuído, como o "abi"ou "cin", mas por certo o que contribuiu mais significativamente foi o artigo definido "el". Por estas razões, o score e a margem de classificação são tão elevados, praticamente iguais a 1.

Caso 2) "Tu vais à escola hoje":
Neste caso, houve um erro de classificação.
Como "tu"é um pronome pessoal comum às línguas francesa e portuguesa e "vais"é um verbo também comum nestas línguas, para além da enorme ocorrência do trigrama "col"na lingua francesa, e somando o facto de que estas línguas de forma geral são bastante similares e têm bastantes fonemas e por conseguinte trigramas em

comum, a margem de classificação neste caso foi bastante reduzida (0.58) e o score também não muito elevado (0.79). Também o facto de a frase ser curta contribui para o erro do classificador, dado que em frases mais compridas, o efeito da similaridade entre as línguas se suaviza.

Caso 3) "Tu vais à escola hoje pois já estás melhor":
Neste caso, apesar de o início da frase ser igual ao caso anterior, a classificação já foi bem feita, com um score e margem de classificação de praticamente 1. Isto é devido ao facto de que a segunda parte da frase, que é diferente do caso anterior, já não tem tantas semelhanças com a língua francesa, sendo muito mais facilmente identificável como português.

Caso 4) "English is easy to learn":
Neste caso, a frase foi corretamente classificada, com uma score e margem de classificação de praticamente 1. Como a língua inglesa é uma língua bastante diferente das restantes línguas no conjunto de treino, por ser de origem anglo-saxónica, o que faz com que tenha menos trigramas em comum com as outras línguas, e por isso qualquer frase nesta língua será mais facilmente bem classificada.

Caso 5) "Tu vas au cinéma demain matin":
Neste caso como o fonema "in"aparece com bastante frequência nos trigramas da língua francesa, assim como "au", que é uma preposição, naturalmente que esta frase foi classificada corretamente como francesa com um score e uma margem praticamente iguais a 1.
De notar que o trigrama "in "aparece 4666003 vezes no conjunto de treino na língua francesa, ao passo que a que mais se aproxima é a espanhola se aproxima com 1549652 vezes, sendo mesmo assim esta frequência quase metade da francesa. O mesmo se verifica em relação à preposição "au".

Caso 6) "É fácil de entender":
Esta frase foi classificada incorretamente como espanhola, apesar de ser portuguesa.
Tal prende-se com o facto de ambas as línguas serem praticamente iguais e, aliás as línguas portuguesa e espanhola terem origem num mesmo idioma.
Por conseguinte têm bastantes trigramas em comum e não só esta frase, como todas as outras, deverão apresentar uma margem de classificação bastante reduzida. Apenas em textos mais compridos é que o classificador deverá obter margens de classificação e scores maiores.
Neste caso, a margem de classificação foi bastante reduzida, de 9%, devido à língua portuguesa também obter um score elevado.