

Laboratório N° 8: Classificador de Bayes e Avaliação do Classificador

Extração Automática de Informação
2021/2022

Prof. Joaquim Filipe
Eng. Filipe Mariano

Objetivos

- Importância dos erros
- Decisão do classificador de bayes tendo em conta o risco de classificação mínimo
- Matriz de confusão
- *Precision*, *Recall* e *F1-score* do classificador

1. Introdução

No laboratório anterior foi introduzida a regra de decisão de bayes no qual o classificador de MAP se baseia. No entanto, o classificador MAP minimiza a probabilidade de erro de classificação sem atender à importância relativa dos erros. A solução passa por atribuir um custo a cada decisão e projetar o classificador para minimizar o custo médio (risco de classificação). Ao classificador que funciona de acordo com este critério, dá-se o nome de **Classificador de Bayes**.

Neste laboratório também se irá proceder ao cálculo de algumas métricas de modo a poder avaliar o classificador.

2. Tipos de Erro e Matriz de Confusão

A matriz de confusão pretende mostrar onde o sistema está a errar ao prever determinadas classes. É apresentada num formato de tabela com o mesmo número de linhas e colunas e que permite a visualização do desempenho de um classificador em aprendizagem supervisionada. Cada linha da matriz representa as instâncias de uma classe prevista, enquanto cada coluna representa as instâncias de uma classe real, ou vice-versa.

		Actual class		
		Cat	Dog	Rabbit
Predicted class	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

Figura 1: Matriz de confusão para 2 e 3 classes.

Tal como é possível observar pela figura anterior, por vezes a classe prevista pelo classificador é diferente da classe real, o que leva a um erro de classificação. É comum falar de dois tipos de erro na classificação:

- **Erro tipo I** (falso positivo)
Exemplo: Detetar um incêndio quando este não existe
- **Erro tipo II** (falso negativo)
Exemplo: Não detetar um incêndio quando na realidade este existe

		True condition	
		Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive	False positive, Type I error
	Predicted condition negative	False negative, Type II error	True negative

Figura 2: Identificação dos erros de tipo I e erros de tipo II.

3. Classificador de Bayes

O classificador MAP minimiza a probabilidade de erro de classificação sem atender à importância relativa dos erros. A solução passa por atribuir um custo a cada decisão e projetar o classificador para minimizar o custo médio (risco de classificação). Ao classificador que funciona de acordo com este critério, dá-se o nome de **Classificador de Bayes**.

A fórmula que minimiza o risco é:

$$\hat{\omega} = \omega_k : k = \arg \min_j c_j(x) = \arg \min_j \sum_i c_{ij} P(\omega_i | x)$$

Assim, no classificador de Bayes é tida em conta a matriz de confusão obtida de modo a minimizar o risco de classificação.

Observe o exemplo dado nas aulas teóricas para uma decisão binária (2 classes):

$c(\omega, \hat{\omega})$	$\hat{\omega}_1$	$\hat{\omega}_2$
ω_1	0.2	0.8
ω_2	0.9	0.1

$$c_1(x) = 0.2 P(\omega_1|x) + 0.9 P(\omega_2|x) =$$

$$= [0.2 P(x|\omega_1)P(\omega_1) + 0.9 P(x|\omega_2)P(\omega_2)] \frac{1}{p(x)}$$

$$c_2(x) = 0.8 P(\omega_1|x) + 0.1 P(\omega_2|x)$$

- A decisão consiste em determinar, dado um padrão x , qual dos custos de classificação é menor: $c_1(x)$ ou $c_2(x)$
- Dado que $c_j(x) = \sum_i c_{ij} P(\omega_i | x)$
- Aplicando as probabilidades a posteriori, a decisão será ω_1 se:
 $c_{11}p(x | \omega_1)P(\omega_1) + c_{21}p(x | \omega_2)P(\omega_2) < c_{12}p(x | \omega_1)P(\omega_1) + c_{22}p(x | \omega_2)P(\omega_2)$
e a decisão será ω_2 no caso contrário

Figura 3: Slides de EAI referentes à decisão do classificador de bayes.

4. Precisão (Precision), Cobertura (Recall) e F-measure (F1-score)

O *Precision* é uma métrica que indica o quão preciso o modelo está de acordo com os que foram previstos positivamente, quantos deles é que são de facto positivos. É uma boa medida para determinar quando os custos de Falso Positivo são altos. Por exemplo, na deteção de spam em emails um Falso Positivo significa que um email que não é spam foi identificado como spam e desta forma o utilizador pode perder emails importantes se a precisão não for alta.

O *Recall* calcula quantos dos Positivos Verdadeiros são capturados quando rotulados como Positivos Verdadeiros. Neste caso o *Recall* é uma métrica que terá grande importância quando há um alto custo associado aos Falsos Negativos. Por exemplo, se um doente for classificado como não doente, o custo

associado ao Falso Negativo será alto se tivermos em conta que podemos estar perante uma doença contagiosa.

O *F-measure* ou *F1-score* é utilizada quando se pretende encontrar um equilíbrio entre o *Precision* e o *Recall*. É uma métrica adequada para encontrar esse equilíbrio quando existe um grande número de Verdadeiros Negativos.

As fórmulas para cálculo destas métricas são:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figura 4: Métricas que permitem avaliar o classificador.

5. Exercícios

1. Criar um módulo de `stats` que permita exportar as seguintes funções:

- a. `confusionMatrix`: recebe um array de objetos com o documento, a classe prevista (utilizando um dos classificadores) e a classe real. Retorna uma matriz quadrada com os números referentes à previsão de cada classe e os que realmente eram verdadeiros.
- b. `precision`: a partir dos dados obtidos pela função `confusionMatrix` calcular a precisão do classificador.
- c. `recall`: a partir dos dados obtidos pela função `confusionMatrix` calcular a cobertura do classificador.
- d. `fMeasure`: a partir dos dados obtidos pela função `confusionMatrix`, `precision` e `recall` calcular a o *F1-score* do classificador.
- e. Considere a utilização destas funções através de dados provenientes da base de dados ou através de uma amostra que está em classificação num determinado momento.