Trabalho 2 de Inteligência Artificial - CLASSIFICADOR BAYESIANO

Miller Biazus

O trabalho consiste de um classificador de texto bayesiano para um conjunto de arquivos texto composto por 200 arquivos da classe Positivo e 200 arquivos da classe Negativo.

Os arquivos são separados em 10 grupos com 20 arquivos cada. Para teste, os grupos são selecionados um por um, enquanto os restantes compõem o grupo de treinamento. Cada grupo de teste, então, é colocado numa pasta 'positivoteste' ou 'negativoteste', dependendo de sua classe 'inicial'. Então, cada arquivo da pasta de teste gera dois resultados, que são as probabilidades de ser positivo e de ser negativo. A partir disto, cada grupo de teste recebe uma predição e são gerados números de acerto e erro.

Verdadeiro Positivo = pertence a classe positivo e predição retornou positivo
Falso Negativo = pertence a classe positivo, porém predição retornou negativo
Verdadeiro Negativo =pertence a classe negativo e predição retornou negativo
Falso Positivo = pertence a classe negativo, porém predição retornou positivo

*Legenda de cores das imagens

A matriz de confusão (para cada grupo de teste) é mostrada a seguir:

		PREDIÇÃO		
		CLASSE POSITIVO	**CLASSE NEGATIVO**	
	CLASSE POSITIVO			Desvio Padrão
	33-52	20	0	
	53-72	20	0	
	73-92	19	1	
	93-112	17	3	
	113-132	20	0	1.341641
	133-152	19	1	
	153-172	20	0	
	173-192	16	4	
	193-212	19	1	
	213-232	20	0	
REAIS	**CLASSE NEGATIVO**			Desvio Padrão
	33-52	15	5	
	53-72	11	9	
	73-92	10	10	
	93-112	16	4	
	113-132	15	5	3.039737
	133-152	13	7	
	153-172	17	3	
	173-192	18	2	
	193-212	13	7	
	213-232	8	12	

O desvio padrão é calculado a partir dos 10 valores retornados para cada classe.

De acordo com os resultados anteriores, pode-se calcular / mostrar alguns resultados relevantes, como matriz de confusão média e taxas de acertos:

		PREDIÇÃO	
		CLASSE POSITIVO	**CLASSE NEGATIVO**
REAIS	**CLASSE POSITIVO**	190	10
	CLASSE NEGATIVO	64	136

		PREDIÇÃO	
		CLASSE POSITIVO	**CLASSE NEGATIVO**
REAIS	**CLASSE POSITIVO**	95.00%	5.00%
	CLASSE NEGATIVO	32.00%	68.00%

ACURÁCIA	**81.50%**

Alguns outros cálculos são mostrados a seguir:

Accuracy = (TP + TN) / (P + N) = (190 + 136) / 400 = **0.815**
Precision = TP / (TP + FP) = 190 / 254 = **0.748**
Recall = TP / (TP + FN) = 190 / (190 + 10) = **0.95**
F-score = 2 * Precision * Recall / (Precision + Recall) = **0.837**

Sendo que:

- **Precisão** é a probabilidade de que um documento pego randomicamente seja relevante
- **Recall** é a probabilidade de que um documento relevante aleatório seja selecionado por uma pesquisa

A média harmônica entre esses 2 valores é chamada de **Medida F1** ou **Medida F**.

O coeficiente de correlação de Matthews calcula a qualidade de uma classificação entre duas classes.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

O cálculo:

0.95 x 0.68 - 0.32 x 0.05 / sqrt((0.95+0.32)(0.95+0.05)(0.68+0.32)(0.68+0.05))
= 0.646-0.016 / sqrt(1.27 x 1 x 1 x 0.73)
=0.63 / sqrt(0.9271)
= 0.63 / 0.962860322
= 0.6543