UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MILLER BIAZUS

Classificador NB para a disciplina de Aprendizagem de Máquina

Trabalho Individual

Porto Alegre
2016

1 CLASSIFICADOR BAYESIANO

O trabalho apresenta um Classificador NB para a disciplina de Aprendizagem de Máquina. O Classificador deve classificar arquivos de texto nas classes *negativo* e *positivo*, de acordo com as palavras contidas em tais arquivos, independentemente da ordem. São recomendadas a utilização de rotinas LEARN_NAIVE_BAYES_TEXT e CLASSIFY_NAIVE_BAYES_TEXT, utilizando a notação logaritmica para o discriminante. O objetivo do trabalho é classificar um subconjunto de teste a partir de um conjunto de 24mil arquivos positivos e 24 mil arquivos negativos. Uma validação cruzada 10-fold também deve ser feita.

1.1 Algoritmo

Para a criação do classificador, foi utilizado o ambiente Linux, distribuição UBUNTU, e a linguagem de programação Python. Nenhuma biblioteca de inteligência artificial foi utilizada durante o trabalho.

Os algoritmos utilizados são os descritos na definição do exercício. O seguinte é feito: o programa divide o conjunto de arquivos-texto em 10 subconjuntos. Cada subconjunto é testado com o restante dos arquivos, ou seja, utiliza-se um subconjunto para teste e os outros 9 para treinamento, e assim sucessivamente, a fim de fazer a validação cruzada 10-fold.

Os arquivos-texto de treinamento, para cada validação cruzada, são lidos na linguagem Python, e cada palavra passa a fazer parte de um par key-value (palavra x número de ocorrências). Além disso outros parâmetros, como o número de palavras distintas, são utilizados a fim de realizar o cálculo de classificação.

É importante ressaltar que a determinação da classe é feita simplesmente a partir da maior probabilidade. Os resultados da classificação são mostrados na tabela 1.1. Os valores das probabilidades são convertidos, durante a execução, para a notação de logaritmos a fim de facilitar a representação dos mesmos. O programa de classificação segue anexado à mesma pasta deste relatório.

Tabela 1.1 – Resultados da classificação para cada subset de teste

		PREDIÇÃO		
		Positivo	Negativo	Arquivos de teste:
REAL	Positivo	1939	461	20 - 2419
	Negativo	471	1929	
REAL	**Positivo**	**1913**	**487**	**2420 - 4819**
	Negativo	**359**	**2041**	
REAL	Positivo	1800	600	4820 - 7219
	Negativo	316	2084	
REAL	**Positivo**	**1841**	**559**	**7220 - 9619**
	Negativo	**314**	**2086**	
REAL	Positivo	1772	628	9620 - 12019
	Negativo	315	2085	
REAL	**Positivo**	**1905**	**495**	**12020 - 14419**
	Negativo	**374**	**2026**	
REAL	Positivo	1872	528	14420 - 16819
	Negativo	338	2062	
REAL	**Positivo**	**1819**	**581**	**16820 - 19219**
	Negativo	**275**	**2125**	
REAL	Positivo	1766	624	19220 - 21619
	Negativo	334	2066	
REAL	**Positivo**	**1831**	**569**	**21620 - 24019**
	Negativo	**302**	**2098**	

1.2 Análise dos resultados

É possível obter, a partir da tabela 1.1, a matriz de confusão média para a classificação, a qual é apresentada a seguir.

		PREDIÇÃO	
		Positivo	Negativo
REAL	Positivo	1846	554
	Negativo	340	2060

Precisão: TP/(TP+FP) = 1846 /(1846+340)=1846 / 2186 = **0,8444647758462946**

Taxa de verdadeiros positivos (Recall): TP/(TP+FN) = 1846/2400 = **0,7691**

Taxa de falsos positivos: FP / (FP+TN) = 340/(340+2060)=**0,1417**

Medida-F: 2*Precisão*Recall / (Precisão+Recall)= 2*0,8445*0,7692 / (0,8445+0,7692) =1,2991788/1,6137 = **0,8050931399888455**

Desvio Padrão P = $(1939-1845,80)^2 + (1913-1845,80)^2 + \ldots + (1831-1845,80)^2/10 = 59.875$

Desvio Padrão N = 54.020