Instituto Superior Técnico

MEEC

Aprendizagem Automática

Lab 4

Bayes Classifier

Grupo 9

Manuel Diniz, 84125
Alexandre Rodrigues, 90002

Turno: 4ªf 11h00

Contents

Chapter 1

Classificador de Bayes

Chapter 2

Exemplo

Chapter 3

Reconhecimento de Linguagem

É agora aplicado um *naive Bayes classifier* a texto, de modo a se fazer o reconhecimento da linguagem em que este está escrito. O conjunto de treino é um conjunto de trigramas e o respetivo número de ocorrências nos textos originais.

3.1 Geração da matriz de treino

A partir dos dados, é gerada uma matriz de treino, com uma amostra (neste caso uma linguagem) em cada linha, e uma *feature* (neste caso, o número de ocorrências de um determinado trigrama) por coluna. Gera-se também um vetor com as *labels* corretas correspondentes às amostras.

3.2 Treino e teste

De seguida, o modelo é treinado e testado com os mesmos dados de treino, verificando que atribui a linguagem correta a cada conjunto de treino.

Testa-se agora o modelo em 6 frases fornecidas. Estas são processadas de modo a serem decompostas nos seus trigramas, e gera-se um vetor linha com o mesmo formato dos dados de treino, ou seja, com o número de ocorrências de cada trigrama em cada coluna (fazendo correspondência posicional entre os trigramas obtidos e os do conjunto de treno). Os resultados da previsão neste conjunto de dados estão na tabela

Texto	Linguagem real	Linguagem reconhecida	*Score*	Margem de classificação
Que fácil es comer peras.	es	es	0.6703	0.3407
Que fácil é comer peras.	pt	pt	0.9999	0.9999
Today is a great day for sightseeing.	en	en	1.0000	1.0000
Je vais au cinéma demain soir.	fr	fr	0.9999	0.9999
Ana es inteligente y simpática.	es	es	0.9999	0.9999
Tu vais à escola hoje	pt	fr	0.7930	0.5861

Todas as frases são identificadas corretamente com a exceção da última. A primeira é identificada com um *score* modesto, pois não existem traços fortes que a língua em questão seja espanhol e não português. A única diferença que a frase tem da seguinte, a palavra "es", forma os trigramas " es" e "es ", que também são bastante comuns na língua portuguesa ("estar" ou "antes", por exemplo).

O mesmo não é o caso na segunda frase, que tem um forte indicador da língua portuguesa, o trigrama " é ", que ocorre muito menos vezes no conjunto de treino espanhol. Deste modo, classifica a frase como português com uma confiança, ou *score* muito mais elevada.

Mais uma vez a terceira frase é classificada com elevada confiança, devido à presença de trigramas que são fortes indicadores da sua língua, como "ght" ou "day" neste caso.

A quarta frase também possui *score* elevado, com um dos trigramas chave sendo "oir", muito comum no francês.

A quinta possui um trigrama quase do exclusivo do espanhol, " y ", sendo que também é classificada com elevada segurança.

A sexta e última frase é identificada incorretamente como francês, se bem que com um *score* e margem de classificação reduzidos. Pode-se atribuir parcialmente à presença do trigrama " à ", que é muito comum no francês, para além de que quase todos os restantes trigramas são partilhados pelas duas linguagens de forma comum.