

Test Collections

• Reuters-21578

- Reuters newswire articles classified by topic
 90 categories (multi-label)
- 90 categories (matt-label)
 9603 training documents / 3299 test documents (ModApte)
- ~27,000 features
- Ohsumed MeSH
 - Medical abstracts classified by subject heading
- 20 categories from "disease" subtree (multi-label)
- 10,000 training documents/ 10,000 test documents
 ~38,000 features
- WebKB Collection
 - WWW pages classified by function (e.g. personal HP, project HP)
 - 4 categories (multi-class)
 - 4183 training documents / 226 test documents
 - ~38,000 features

Example: Reuters Article (Multi-Label)

Categories: COFFEE, CRUDE

KENYAN ECONOMY FACES PROBLEMS, PRESIDENT SAYS

The Kenyan economy is heading for difficult times after a boom last year, and the country must tighten its belt to prevent the balance of payments swinging too far into deficit, President Daniel Arap Moi said. In a speech at the state opening of parliament, Moi said high coffee prices and cheap oil in 1986 led to economic growth of five pct, compared with 4.1 pct in 1985. The same factors produced a two billion shilling balance of payments surplus and inflation fell to 5.6 pct from 10.7 pct in 1985, he added.

"But both these factors are no longer in our favour ... As a result, we cannot expect an increase in foreign exchange reserves during the year," he said.

Example: Ohsumed Abstract

Categories:

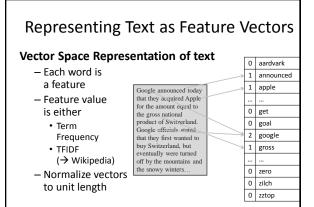
 Animal, Blood_Proteins/Metabolism, DNA/Drug_Effects, Mycotoxins/Toxicity, ...

How aspartame prevents the toxicity of ochratoxin A.

Creppy EE, Baudrimont I, Anne-Marie

Toxicology Department, University of Bordeaux, France

The ubiquitous mycotoxin ochratoxin A (OTA) is found as a frequent contaminant of a large variety of food and feed and beerage such as beer, coffee and win. It is produced as a secondary metabolize of models from Asperglikus and Penntillum genera. Coltaviscon A has been shown experimentally to inhibut protein synthesis by competition with pherglulanine its structural analogue and also to enhance caygen reactive radicals production. The combination of these basis mechanisms with the usureal long plasma had-lifs time (35 days in non-human primates and in human), the metabolisation of CTA into itil active derivatives and glutathone conjugate both potentially reactive with collular macromolecules including DNA could explan the nulliple tone effects, cytochrosity, genetosenticy, genetosenticy, materneticy and carcinogeneixty. A relation was first recognised between exposure to OTA in the Balkan geographical



Multi-Class via "One-against-rest"

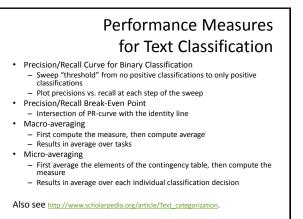
- Goal: $h: X \to Y$ with $Y = \{1, 2, 3, ..., k\}$
- Problem:
- Many classifiers can only learn binary classification rules
- Most common solution
- Learn one binary classifier for each class
- Put example into the class with the highest probability (or some approximation thereof)
- Example
 - Binarize: $(x, 2) \rightarrow (x, (-1, +1, -1, -1, -1, -1, -1, -1, -1))$
 - Training: one binary classifier h_i(x) for each class i
 - Prediction on new x: $h(x) = \operatorname{argmax}\{h_i(x)\}$
 - Assumes that classifier outputs a confidence score and not just +1 / -1

Multi-Label via "One-against-rest"

- Goal: $h: X \to Y$ with $Y = 2^{\{1,2,3,\dots,k\}}$
- Problem:
 - Many classifiers can only learn binary classification rules
- Most common solution
 - Learn one binary classifier for each label
 - Attach all labels, for which its binary classifier says positive
- Example
 - Binarize: (x, {2,7,9}) → (x, (-1,+1,-1,-1,-1,+1,-1,+1,-1))
 - Training: one binary classifier $h_i(x)$ for each class i
 - Prediction on new x: $h(x) = (h_1(x), h_2(x), ..., h_k(x))$

$\begin{array}{c|c} \hline & h(x) = +1 & h(x) = -1 \\ \hline & y = +1 & a & b \\ \hline & y = -1 & c & d \\ \hline \end{array}$ • Commonly used performance measures = ErrorRate = (b+c) / (a+b+c+d)

- Accuracy = (a+d) / (a+b+c+d)
- WeightedErrorRate = $(w_b^*b + w_c^*c) / (a+b+c+d)$
- Precision = a / (a+c)
- Recall = a / (a+b)
- F₁Score = 2 * (Precision * Recall) / (Precision + Recall)



	Experiment Results					
 90 categories 9603 training docs 3299 test docs ~27000 features 	WebKB Collection - 4 categories - 4183 training docs - 226 test docs - ~38000 features		Ohsumed MeSH – 20 categories – 10000 training docs – 10000 test docs – ~38000 features			
Microaveraged precision/recall breakeven point [0100]		Reuters	WebKB	Ohsumed		
Naïve Bayes (multinomial)		72.3	82.0	62.4		
Rocchio Algorithm (LDA)		79.9	74.1	61.5		
C4.5 Decision Tree		79.4	79.1	56.7		
k-Nearest Neighbors		82.6	80.5	63.4		
		87.5	90.3	71.6		

Comparison of Methods for Text Classification

	Naïve Bayes	Rocchio (LDA)	TDIDT C4.5	k-NN	svм
Simplicity (conceptual)	++	++	-	++	-
Efficiency at training	+	+		++	-
Efficiency at prediction	++	++	+		++
Handling many classes	+	+	-	++	-
Theoretical validity	-	-	-	0	+
Prediction accuracy	-	0	-	+	++
Stability and robustness	-	-	-	+	++