TECHNISCHE UNIVERSITÄT DRESDEN

Dresden Database Systems Group

Exercise 11

Information Retrieval

15..19. Text Classification

Warm up

Dresden Database Systems Group

Exercise 15..19.1

- Are the following statements true or false? Give reasons for your answer.

a) Classification is about assigning a document to one (or more) out of several predefined categories.

b) The accuracy of a classifier is always one minus its error rate.

c) When using supervised learning for classification, we first train a classifier on unlabeled documents.

d) Overfitting means that a classifier is too general.

e) With n-times k-fold cross-validation, we partition a set of n labeled documents into k (nearly) equally-sized subsets and for each subset S_i we train a classifier on the documents in S_i and evaluate its performance by applying it to the documents in S_i.

f) Given a set of labeled documents, the sequential covering algorithm determines a set of rules for rule-based text classification.

Warm up

Dresden Database Systems Group

Exercise 15..19.1

- Are the following statements true or false? Give reasons for your answer.

g) With probabilistic classification, a document d is assigned to category c_i if the probability $P(c_i \mid d)$ is the maximum of the probabilities $P(c_k \mid d)$ for all categories c_k.

h) Feature selection can reduce the training time of a classifier, but it cannot improve its quality.

i) The Rocchio approach for vector-based classification assumes that the document vectors in each category are close to each other, but distant from the document vectors in the other categories.

j) The kNN classifier assigns a doc. d to the k categories whose centroid vectors are closest to $\vec{d}$.

k) The idea of support vector machines (SVMs) is to separate the vector space using an optimal hyperplane and to assign a document d to one of two classes depending on whether $\vec{d}$ lies on the one or on the other side of the hyperplane.

ok

l) The SVM approach can only be applied to linearly separable datasets.

Multinomial Model vs. Bernoulli Model

Dresden Database Systems Group

Exercise 15..19.2

- Consider the following Bernoulli and multinomial estimates for the word “the”
  - multinomial model: $\hat{P}(t = \text{"the"} \mid c) \approx 0.05$
  - Bernoulli model: $\hat{P}(t = \text{"the"} \mid c) \approx 1.00$

- Explain the difference

? ?

Naive Bayes Classification

Dresden Database Systems Group

Exercise 15..19.3

- Based on the data in the table below,
  - estimate a **multinomial Naive Bayes classifier** and apply it to the test document
  - estimate a **Bernoulli Naive Bayes classifier** and apply it to the test document
- You don't need to estimate parameters that you don't need for classifying the test document

	docId	Tokens in document	In c = China?
Training set	1	Taipei Taiwan	Yes
	2	Macao Taiwan Shanghai	Yes
	3	Japan Sapporo	No
	4	Sapporo Osaka Taiwan	No
Test set	5	Taiwan Taiwan Sapporo	?

两套大公式，还得练

Feature Selection

Dresden Database Systems Group

Exercise 15..19.4

三种属性选择，目前没答案 NEXT

- Assume we have a collection of 100k documents
- Consider the following frequencies $N_{e_t e_c}$ for the class **`coffee`** for three terms

Term t	N_{00}	N_{10}	N_{01}	N_{11}	Frequency $N(t,c)$	Mutual Inf. $I(U_t;C_c)$	Chi-Square $\chi^2(t,c)$
`brazil`	98,012	102	1835	51		0,00155	818,9
`roasted`	99,824	143	23	10		0,00065	1964,3
`producer`	98,729	119	1118	34			

$N_{e_t e_c}$

$0 \rightarrow t \notin d$ $1 \rightarrow t \in d$ $0 \rightarrow d \notin c$ $1 \rightarrow d \in c$

a) Fill in the empty cells and select two of these three terms based on

frequency | mutual information | χ^2

b) What are the values of $I(U_t;C_c)$ and $\chi^2(t,c)$ if term and class are completely

dependent | independent