Text classification and Naive Bayes

presentation by eshock

Introduction

Examples:
Sentiment Detection

Email Sorting

Vertical Search Engine

The Text Classification Problem

Document Space

Training Set

Classification function

Supervised Learning

Naive Bayes text classification

multinominal naive bayes

maximum a posteriori class

Estimation of P(t|c)

The Bernoulli model

The Bernoulli model estimates $P(t|c)$ as the fraction of documents of class c that contain term t, while the multinominal model estimates $P(t|c)$ as the fraction of tokens or fractions of positions in documents of class c that contain term t.

Properties of Naive Bayes

why is it NAIVE?
two naive assumptions:
conditional independence assumption

positional independence assumption

Though the probability estimates of NB are of low quality, its classification decisions are surprisingly good.

Feature Selection

Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification.

how to decide which terms to use?
compute the utility measure $A(t,c)$ and select k terms with the highest values of $A(t,c)$

how to compute A(t,c)

Mutual Information
MI measures how much information the presence / absence of a term contributes to making the correct classification decision on c.

X2 feature selection
applied to test the independence of two events

Frequency-based feature selection
select the terms that are most common in the class.
can be either defined as document frequency or as collection frequency.

Comparison of feature selection methods

Fig.13.8 page 275

Evaluation

effectiveness

precision, recall, F1, and accuracy