Introduction to Information Retrieval
信息检索导论

Chunwei Yan

2013 年 1 月 15 日

1 文本分类问题

文本分类的目标是：利用某个分类函数 (classification function)，能够将文档映射到类别:

$$\Upsilon : X \to C \tag{1}$$

监督学习

2 Text classification and Naive Bayes

the probability of document d in class c is calculated as

$$P(c|d) \sim P(c)\Pi_{1\leq k\leq n_d}P(t_k|c) \tag{2}$$

- $P(c)$ is the prior probability of document occuring in class c.
- $P(t_k|c$ is the conditional probability of term k occuring in a document of class c.

the best class in NB is the most likely or MAP class C_{map}.

$$c_map = arg\max_{c\in C}\hat{P}(c|d) = arg\max_{c\in C}\hat{P}(c)\Pi_{1\leq k\leq n_d}\hat{P}(t_k|c) \tag{3}$$