



Part 8

Exercise 9.1, 9.2

- 9.1: In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one" search correspond to?
 - Slide 8 of relevance feedback
- 9.2: Give three reasons why relevance feedback has been little used in web search.

Exercise 9.3

- Under what conditions would the modified query q_m in Equation 9.3 be the same as the original query q_0 ?
- In all other cases, is q_m closer than q_0 to the centroid of the relevant documents?

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in D_{nr}} \vec{d}_j$$

Exercise 9.4

- Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs".
- The user examines two documents, d_1 and d_2 . She judges d_1 , with the content "*CDs cheap software cheap CDs*" relevant and d_2 with content "*cheap thrills DVDs*" nonrelevant.
- Assume that we are using direct term frequency (with no scaling and no document frequency).
- Using Rocchio relevance feedback as in Equation (9.3) what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.



Part 9

Exercise Multinomial

- Consider the following training set
 - D1: "American Boston American" -> Class=Y
 - D2: "American American Chicago" -> Class=Y
 - D3: "American Washington" -> Class=Y
 - D4: "Rome Italy American" -> Class=N
- Estimate $P(w|c)$ as described in the next slide ($\alpha = 1$)
- What is the class of the following test document:
 - D5: "American American American Rome Italy"

Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*
- Calculate required $P(c_j)$ and $P(x_k | c_j)$ terms
 - For each class c_j in C do
 - $docs_j \leftarrow$ subset of documents for which the target class is c_j

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- $Text_j \leftarrow$ single document containing all $docs_j$
 - for each word x_k in *Vocabulary*
 - $n_{jk} \leftarrow$ number of occurrences of x_k in $Text_j$
 - $n_j \leftarrow$ number of words in $Text_j$

$$P(x_k | c_j) \leftarrow \frac{n_{jk} + \alpha}{n_j + \alpha |Vocabulary|}$$

Assume $\alpha = 1$;
this is for
smoothing

Exercise Multivariate

- Consider the same training and test set of the previous example
- Predict the class of the test set using the Multivariate Bernoulli model
- What are the features X_i in this model? What are the values of these variables x_i ?
- Estimated $P(X_i=x_i | c_j)$ as follow – k is the number of possible values for X_i

$$\hat{P}(X_i = x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

- $N(X_i=x_i, C=c_j)$ is the number of documents that have value x_i for features X_i and are in class c_j .

Exercise 13.1, 13.2

- 13.1: Why is $|C||V| < |D|L_{\text{ave}}$ expected to hold for most text collections?
 - $|C|$ -#classes, $|V|$ -voc. size, $|D|$ -#documents – L_{ave} is avg. document length
- 13.2: Which of the documents in Table 13.5 have identical and different bag of words representations for (i) the Bernoulli model, and (ii) the multinomial model? If there are differences, describe them.

► **Table 13.5** A set of documents for which the NB independence assumptions are problematic.

- (1) He moved from London, Ontario, to London, England.
- (2) He moved from London, England, to London, Ontario.
- (3) He moved from England to London, Ontario.

Exercise 13.3

- Table 13.3 gives Bernoulli and multinomial estimates for the word the.
- Explain the difference:
 - $P(X = \text{the}|c) \approx 0.05$ Multinomial
 - $P(\text{'the' is present } |c) \approx 1.0$ Multivariate

► Table 13.3 Multinomial versus Bernoulli model.
multinomial model

	multinomial model	Bernoulli model
event model	generation of token	generation of document
random variable(s)	$X = t$ iff t occurs at given pos	$U_t = 1$ iff t occurs in doc
document representation	$d = \langle t_1, \dots, t_k, \dots, t_{n_d} \rangle, t_k \in V$	$d = \langle e_1, \dots, e_i, \dots, e_M \rangle,$ $e_i \in \{0, 1\}$
parameter estimation	$\hat{P}(X = t c)$	$\hat{P}(U_i = e c)$
decision rule: maximize	$\hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(X = t_k c)$	$\hat{P}(c) \prod_{t_i \in V} \hat{P}(U_i = e_i c)$
multiple occurrences	taken into account	ignored
length of docs	can handle longer docs	works best for short docs
# features	can handle more	works best with fewer
estimate for term the	$\hat{P}(X = \text{the} c) \approx 0.05$	$\hat{P}(U_{\text{the}} = 1 c) \approx 1.0$

Exercise

- Consider the example 13.3
- Compute the Chi-square statistic using equation (13.19) (use excel) and check that it is 284
- Compute E_{10} , E_{01} , E_{00} and check that you have the same results as in the example.

Example 13.3: Consider the class *poultry* and the term *export* in Reuters-RCV1. The counts of the number of documents with the four possible combinations of indicator values are as follows:

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27,652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774,106$

Exercise 13.6

- Assume a situation where every document in the test collection has been assigned exactly one class, and that a classifier also assigns exactly one class to each document.
- This setup is called one-of classification (Section 14.5, page 306).
- Show that in one-of classification the total number of false positive decisions (for all classes) equals the total number of false negative decisions (for all classes).