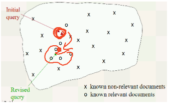
Relevance feedback + Text classification (Chapter 9+13)

$\alpha = 1$ $+ 0{,}7 \cdot \frac{1}{\ldots}\sum \ldots$ $- 0{,}15 \frac{1}{\ldots}\sum \ldots$ γ

Definition 1 (Rocchio relevance feedback)
Rocchio relevance feedback has the form



$$q_m = \alpha q_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_r \in D_r} \vec{d}_r - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_{nr} \in D_{nr}} \vec{d}_{nr}$$

where q_0 is the original query vector, D_r is the set of relevant documents, D_{nr} is the set of non-relevant documents and the values α, β, γ depend on the system setting.

PSEUDO-RELEVANCE FEEDBACK

Exercise 9/1

What is the main purpose of Rocchio relevance feedback?

$q_0 \rightarrow d_1, d_2, \ldots$ D_r $\rightarrow q_1 = \alpha q_0 + \beta \cdots + \gamma \cdots$

d_{n-1}, d_n D_{nr}

d'_1
d'_2

Exercise 9/2

A user's primary query is *cheap CDs cheap DVDs extremely cheap CDs*. The user has a look on two documents: doc1 a doc2, marking doc1 *CDs cheap software cheap CDs* as relevant and doc2 *cheap thrills DVDs* as non-relevant. Assume that we use a simple *tf* scheme without vector length normalization. What would be the restructured query vector after considering the Rocchio relevance feedback with values $\alpha = 1$, $\beta = 0.75$, and $\gamma = 0.25$?

We rewrite the exercise to the table for an easier processing.

	relevant	non-relevant	
terms	doc1	doc2	query
CDs	2	0	2
cheap	2	1	3
software	1	0	0
thrills	0	1	0
DVDs	0	1	1
extremely	0	0	1

Table 1:

$q_0 = [2\ 3\ 0\ 0\ 1\ 1]$

$doc_1 = [2\ 2\ 1\ 0\ 0\ 0]$

$doc_2 = [0\ 1\ 0\ 1\ 1\ 0]$

$q_m = \alpha q_0 + \beta\, doc_1 - \gamma\, doc_2$

$= 1 \cdot q_0 + 0{,}75\, doc_1 - 0{,}25\, doc_2$

$= [3{,}5\ \ 4{,}25\ \ 0{,}75\ \ -0{,}25\ \ 0{,}75\ \ 1]$

$\mathrm{cossim}(q_m, \ldots) \in [-1; 1]$

Ack: 11

Text classification and Naive Bayes (Chapter 13)

Definition 2 (Naive Bayes Classifier)
Naive Bayes (NB) Classifier assumes that the effect of the value of a predictor x on a given class c is class conditional independent. Bayes theorem provides a way of calculating the posterior probability $P(c|x)$ from class prior probability $P(c)$, predictor prior probability $P(x)$ and probability of the predictor given the class $P(x|c)$

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

and for a vector of predictors $X = (x_1, \ldots, x_n)$

$$P(c|X) = \frac{P(x_1|c) \ldots P(x_n|c)P(c)}{P(x_1) \ldots P(x_n)}.$$

The class with the highest posterior probability is the outcome of prediction.

$\vec{x} = [x_1\ x_2\ x_3\ x_4\ \cdots\ x_n]$

$P(x_1, x_2) = P(x_1) \cdot P(x_2)$

$P(x_1, x_2, x_3) = P(x_1, x_2) \cdot P(x_3)$

$$p(x_i \mid \cancel{x_{i+1}, \ldots, x_n}, C_k) = p(x_i \mid C_k)$$

$P(x_i \mid x_{i+1}) = \frac{P(x_i, x_{i+1})}{P(x_{i+1})} = \frac{P(x_i) \cdot \cancel{P(x_{i+1})}}{\cancel{P(x_{i+1})}} = P(x_i)$

Exercise 13/2

Considering the table of observations, use the Naive Bayes classifier to recommend whether to *Play Golf* given a day with *Outlook = Rainy*, *Temperature = Mild*, *Humidity = Normal* and *Windy = True*. Do not deal with the zero-frequency problem.

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No
x_1	x_2	x_3	x_4	c

Table 1: Exercise.

$P(\text{Yes} \mid \text{Rainy}, \text{Mild}, \text{Normal}, \text{True})$

$\propto P(\text{Yes}) \cdot P(\text{Rainy} \mid \text{Yes}) \cdot P(\text{Mild} \mid \text{Yes}) \cdot P(\text{Normal} \mid \text{Yes}) \cdot P(\text{True} / \text{Yes})$

$\propto 9/14 \cdot 2/9 \cdot 4/9 \cdot 6/9 \cdot 3/9 = 0{,}014$

$P(\text{No} \mid \ldots) \propto 5/14 \cdot 3/5 \cdot 2/5 \cdot 1/5 \cdot 3/5 = 0{,}010$

$P(\text{yes} \mid \ldots) = \frac{0{,}014}{0{,}014 + 0{,}010} = 57{,}84\,\%$

$P(\text{No} \mid \ldots) = 42{,}16\,\%$