CS5154/6054 Quiz 9 Key, 9/20/2022

Exercise 18.4

Let

$$C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

be the term-document incidence matrix for a collection. Compute the co-occurrence matrix CC^T. What is the interpretation of the diagonal entries of CC^T when C is a term-document incidence matrix?

The rows of C are terms and columns are documents. CC^T = ((2 1 1) (1 1 0) (1 0 1)). CC^T_{ii} is the document frequency of term i, or the number of documents containing term i.

Exercise 18.6

Suppose that C is a binary term-document incidence matrix. What do the entries of C^TC represent?

C^TC is a document by document matrix. C^TC_{jj} is the number of terms document j contains (or the length of j). C^TC_{ij} is the number of term documents i and j share.

Exercise 18.7

Let

$$C = \begin{pmatrix} 0 & 2 & 1 \\ 0 & 3 & 0 \\ 2 & 1 & 0 \end{pmatrix}$$

be a term-document matrix whose entries are term frequencies; thus term 1 occurs 2 times in document 2 and once in document 3. Compute CC^T; observe that its entries are largest where two terms have their most frequent occurrences together in the same document.

CC^T = ((5 6 2) (6 9 3) (2 3 5)). The largest non-diagonal in CC^T is 6, from the first two terms with their most frequent occurrences together in document 2.

CS5154/6054 Quiz 10 Key, 9/22/2022

▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

(11.19)

	documents	relevant	nonrelevant	Total
Term present	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
Term absent	$x_t = 0$	$S - s$	$(N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	Total	S	$N - S$	N

Using this, $p_t = s/S$ and $u_t = (\mathrm{df}_t - s)/(N - S)$ and

$$c_t = K(N, \mathrm{df}_t, S, s) = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))} \quad (11.20)$$

To avoid the possibility of zeroes (such as if every or no relevant document has a particular term) it is fairly standard to add $\frac{1}{2}$ to each of the quantities in the center 4 terms of (11.19), and then to adjust the marginal counts (the totals) accordingly (so, the bottom right cell totals $N + 2$). Then we have:

$$\hat{c}_t = K(N, \mathrm{df}_t, S, s) = \log \frac{(s + \frac{1}{2})/(S - s + \frac{1}{2})}{(\mathrm{df}_t - s + \frac{1}{2})/(N - \mathrm{df}_t - S + s + \frac{1}{2})} \quad (11.21)$$

Adding $\frac{1}{2}$ in this way is a simple form of smoothing. For trials with cat-

Chinese	3	1
	0	0
Beijing	1	0
	2	1
Tokyo	0	1
	3	0

S = 3, N = 4

t	df_t	s	p_t	u_t	c_t (11.20)	c_t (11.21)
Chinese	4	3	3/3=1	1/1=1		
Beijing	1	1	1/3	0/1=0		
Shanghai	1	1	1/3	0/1=0		
Macao	1	1	1/3	0/1=0		
Tokyo	1	0	0/3=0	1/1=1		
Japan	1	0	0/3=0	1/1=1		
$\sum_t c_t$						

There are only three kinds of t: Chinese, {Beijing, Shanghai, Macao}, and {Tokyo, Japan} with c_t (11.21) log((3.5/0.5)/(1.5/0.5)) = log 3.5 – log 1.5 = log 7 – log 3, log((1.5/2.5)/(0.5/1.5)) = 2log 1.5 – log 0.5 – log 2.5 = 2 log 3 – log 5, and log((0.5/3.5)/(1.5/0.5)) = 2log 0.5 – log 3.5 – log 1.5 = - log 7 – log 3, respectively. Without an explicit query, $\sum_t c_t$ is over t = Chinese, Tokyo, and Japan (using the binary assumption) and the result is log 7 – log 3 + 2(- log 7 - log 3) = - log 7 – 3 log 3 = -5.24 for natural log.

CS5154/6054 Quiz 11 Key, 9/27/2022

Given the (pseudo) relevant set of documents, we no longer need the query to be in our equations and assumptions. The odds that a document represented by the vector x is relevant is (by Bayes rule)

$$O(R|x) = \frac{P(R=1|x)}{P(R=0|x)} = \frac{\frac{P(R=1)P(x|R=1)}{P(x)}}{\frac{P(R=0)P(x|R=0)}{P(x)}} = \frac{P(R=1)}{P(R=0}\frac{P(x|R=1)}{P(x|R=0)}$$

Under the Naïve Bayes assumption, we have (M is the vocabulary size and t is a term or a dimension of the vector x) The Independence assumption may also be acceptable.

$$\frac{P(x|R=1)}{P(x|R=0)} = \prod_{t=1}^{M} \frac{P(x_t|R=1)}{P(x_t|R=0)}$$

With this assumption, we have (in BIM, we assume that x is a boolean vector)

$$O(R|x) = O(R) \prod_{t:x_t=1} \frac{P(x_t=1|R=1)}{P(x_t=1|R=0)} \prod_{t:x_t=0} \frac{P(x_t=0|R=1)}{P(x_t=0|R=0)}$$

and using p_t and u_t, we have

$$O(R|x) = O(R) \prod_{t:x_t=1} \frac{p_t}{u_t} \prod_{t:x_t=0} \frac{1-p_t}{1-u_t}$$

Explain how this is turned into

$$O(R|x) = O(R) \prod_{t:x_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} \prod_{t=1}^{M} \frac{1-p_t}{1-u_t}$$

multiply and also divide $\prod_{t:x_t=1} \frac{1-p_t}{1-u_t}$

Since we are only interested in ranking, not really the odds, we ended up with

$$RSV_d = \log \prod_{t:x_t=1} \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=1} \log \frac{p_t(1-u_t)}{u_t(1-p_t)} = \sum_{t:x_t=1} c_t$$

This means, for the purpose of ranking test documents, with the document "Chinese Chinese Chinese Tokyo Japan" in the last quiz, we do not need to add the c_t's for terms Beijing, Shanghai, and Macao. Even for actually computing log odds for all test documents, we only need to add a term that can be precomputed and has nothing to do with the test documents. **Complete** that term with ??? replaced.

$$\log O(R) + \sum_{t=1}^{M} ???$$

Answer: $\log O(R) + \sum_{t=1}^{M} \log \frac{1-p_t}{1-u_t}$

CS5154/6054 Quiz 12 Key, 9/29/2022

NDCG *lative gain* (NDCG). NDCG is designed for situations of non-binary notions of relevance (cf. Section 8.5.1). Like precision at k, it is evaluated over some number k of top search results. For a set of queries Q, let $R(j,d)$ be the relevance score assessors gave to document d for query j. Then,

$$\mathrm{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)}-1}{\log_2(1+m)}, \tag{8.9}$$

where Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at k for query j is 1. For queries for which $k' < k$ documents are retrieved, the last summation is done up to k'.

m	1	2	3	4	5	6	7	8	9	10
$\log_2(1+m)$	1	1.58	2	2.32	2.58	2.81	3	3.17	3.32	3.46
$1/\log_2(1+m)$	1	0.63	0.5	0.43	0.39	0.36	0.33	0.32	0.3	0.29

Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1 R N R N N N N N R R

System 2 N R N N R R R N N N

Let us apply NDCG at k=10 for the binary relevance scores (R(j, d) is 1 if d at rank j is (R)elevant and 0 if it is (N)onrelevant) and fill out the table at the bottom.

First, write the cumulative gain (CG) at k=10 for the two top 10 results.

Then, what is DCG for the each of the two results, assuming |Q|= 1 and Z_{kj} = 1.

What is the ideal DCG (IDCG, when the result is RRRRNNNNN)? (Z_{kj} = 1/IDCG)

What are NDCG's for the two results? (Since we have the same information need, normalization is not needed for comparison of the two systems.)

What is the Reciprocal Rank (RR) score of each of the systems? (Since there is only one information need, these are also MRR's.)

	CG	DCG	IDCG	NDCG	MRR
System 1	4	1+0.5+0.3+0.29=2.09	1+0.63+0.5+0.43=2.56	0.82	1/1 = 1
System 2	4	0.63+0.39+0.36+0.33=1.71	2.56	0.67	1/2 =0.5

For DCG and IDCG, write down what you are adding up to show how you get the scores.