EM Algorithm for a mixture of Bernoulli distributions

EXPECTATION-MAXIMIZATION ALGORITHM A commonly used algorithm for model-based clustering is the *Expectation-Maximization algorithm* or *EM algorithm*. EM clustering is an iterative algorithm that maximizes $L(D|\Theta)$. EM can be applied to many different types of probabilistic modeling. We will work with a mixture of multivariate Bernoulli distributions here, the distribution we know from Section 11.3 (page 222) and Section 13.3 (page 263):

W_k is the cluster

(16.14)
$$P(d|\omega_k;\Theta) = \left(\prod_{t_m \in d} q_{mk}\right) \left(\prod_{t_m \notin d} (1 - q_{mk})\right)$$

where $\Theta = \{\Theta_1, \dots, \Theta_K\}$, $\Theta_k = (\alpha_k, q_{1k}, \dots, q_{Mk})$, and $q_{mk} = P(U_m = 1 | \omega_k)$ are the parameters of the model.³ $P(U_m = 1 | \omega_k)$ is the probability that a document from cluster ω_k contains term t_m . The probability α_k is the prior of cluster ω_k : the probability that a document d is in ω_k if we have no information about d.

The mixture model then is:

See Section 16.5 in The IR Book

(16.15)
$$P(d|\Theta) = \sum_{k=1}^{K} \alpha_k \left(\prod_{t_m \in d} q_{mk} \right) \left(\prod_{t_m \notin d} (1 - q_{mk}) \right)$$

Example: The EM clustering algorithm

- **Bernoulli Mixture Model Example**
- **Example taken from IR Book: Table 16.3**
- http://nlp.stanford.edu/IR-book/pdf/16flat.pdf

(a)	docID	document text	docID	document text
	1	hot chocolate cocoa beans	7	sweet sugar
	2	cocoa ghana africa	8	sugar cane brazil
	3	beans harvest ghana	9	sweet sugar beet
	4	cocoa butter	10	sweet cake icing
	5	butter truffles	11	cake black forest
	6	sweet chocolate		
		l	See Sec	tion 16.5 in The IR Book

Example taken from IR Book Table 16.3

IR	Bo	ol	<
Ta	ble	1	6.3

Class prior

Class Assignments

r_{i1}

E-Step

M-Step

Word Class Conditionals

 $\mathbf{Q}_{\text{word,class}}$

Sugar in class 2 only

TIM 251: Large-Scale

Bernouilli Mixture Model after each iteration

ı	D (DCII	<u>IOUIIII</u>	<u> </u>	<u>ç ivlout</u>	<u>i aitei</u>	Cacili	teration
	Parameter			Ite	ration o	f cluster	ing		teratio
		0	1	2	3	4	5	15	25
5r	α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
	$r_{1,1}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$r_{2,1}$		0.50	0.79	0.99	1.00	1.00	1.00	1.00
ts	•		0.50	0.84	1.00	1.00	1.00	1.00	1.00
	$r_{4,1}$		0.50	0.75	0.94	1.00	1.00	1.00	1.00
	$r_{5,1}$		0.50	0.52	0.66	0.91	1.00	1.00	1.00
	$r_{6,1}$	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.00
	$r_{7,1}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{8,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{9,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{10,1}$		0.50	0.40	0.14	0.01	0.00	0.00	0.00
	$r_{11,1}$		0.50	0.57	0.58	0.41	0.07	0.00	0.00
	q _{africa,1}	0.000	0.100	0.134	0.158	0.158	0.169	0.200	0.200
	$q_{ m africa,2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	0.000
	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$q_{ m brazil,2}$	0.000	0.167	0.195	0.213	0.214	0.196	0.167	0.167
	q _{cocoa,1}	0.000	0.400	0.432	0.465	0.474	0.508	0.600	0.600
•	$q_{\rm cocoa,2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	0.000
	g _{sugar,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ly	$q_{ m sugar,2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	0.500
	$q_{ m sweet,1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	0.000
ıle	$q_{ m sweet,2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	0.667

ID Dook			Berr	nouilli	<u>Mixtur</u>	e Mode	el aft	er eac	<u>h iteratio</u> ı
IR Book	Parameter			Ite	ration o	f cluster	ing		
Table 16.3		0	1	2	3	4	5	15	25
Class prior	α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.4	5
F	$r_{1,1}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00
Class	$r_{2,1}$		0.50	0.79	0.99	1.00	1.00	1.00	1.00
Assignments	$r_{3,1}$		0.50	0.84	1.00	1.00	1.00	1.00	1.00
r _{i1}	$r_{4,1}$		0.50	0.75	0.94	1.00	1.00	1.00	1.00
E-Step	$r_{5,1}$		0.50	0.50	0//	0.01	1.00	1.00	1.00
r-oreh	$r_{6,1}$	1.00	1.00 (a)	docID 1	documen	it text olate cocoa	hoone		document text sweet sugar
	$r_{7,1}$	0.00	0.00	2		ana africa	Dearis		sweet sugat sugar cane braz
	$r_{8,1}$		0.00	3	beans ha	rvest ghana	a		sweet sugar be
	$r_{9,1}$		0.00	4 5	cocoa bu				sweet cake icin cake black fore
	$r_{10,1}$		0.50	6	sweet cho			11	cake black fore
	r _{11,1}		0.50						
	q _{africa,1}	0.000	0.100	0.134	0.158	0.158	0.16	9 0.20	0.200
	q _{africa,2}	0.000	0.083	0.042	0.001	0.000	0.00	0.00	0.000
M-Step	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.00	0.00	0.000
Word Class	q _{brazil,2}	0.000	0.167	0.195	0.213	0.214	0.19	6 0.16	0.167
	q _{cocoa,1}	0.000	0.400	0.432	0.465	0.474	0.50	8 0.60	0.600
Conditionals	q _{cocoa,2}	0.000	0.167	0.090	0.014	0.001	0.00	0.00	0.000
$Q_{word,class}$	g _{sugar,1}	0.000	0.000	0.000	0.000	0.000	0.00	0.00	000.000
Sugar in class 2 only	q _{sugar,2}	1.000	0.500	0.585	0.640	0.642	0.58	9 0.50	0.500

1.000

1.000

 $q_{\text{sweet,1}}$

 $q_{\mathrm{sweet,2}}$

0.300

0.417

0.238

0.507

0.180

0.610

0.159

0.640

0.153

0.608

0.000

0.667

0.000

0.667

TIM 251: Large-Scale

Bernouilli Mixture Model after each iteration

	Daganakag		<u> </u>					<u> cacii i</u>	teration
IR Book	Parameter			ite	ration o	r ciuster	ing		
Table 16.3		0	1	2	3	4	5	15	25
Class prior	α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
0.000 p	$r_{1,1}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00
Class	$r_{2,1}$		0.50	0.79	0.99	1.00	1.00	1.00	1.00
Assignments	$r_{3,1}$		0.50	0.84	1.00	1.00	1.00	1.00	1.00
r _{i1}	$r_{4,1}$		0.50	0.75	0.94	1.00	1.00	1.00	1.00
-11	$r_{5,1}$		0.50	0.52	0.66	0.91	1.00	1.00	1.00
	$r_{6,1}$	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.00
	$r_{7,1}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{8,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{9,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{10,1}$		0.50	0.40	0.14	0.01	0.00	0.00	0.00
	$r_{11,1}$		0.50	0.57	0.58	0.41	0.07	0.00	0.00
	q _{africa,1}	0.000	0.100	0.134	0.158	0.158	0.169	0.200	0.200
	q _{africa,2}	0.000	0.083	0.042	0.001	0.000	0.000	0.000	0.000
	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Sugar in

▶ Table 16.3 The EM clustering algorithm. The table shows a set of documents (a) and parameter values for selected iterations during EM clustering (b). Parameters shown are prior α_1 , soft assignment scores $r_{n,1}$ (both omitted for cluster 2), and lexical parameters $q_{m,k}$ for a few terms. The authors initially assigned document 6 to cluster 1 and document 7 to cluster 2 (iteration 0). EM converges after 25 iterations. For smoothing, the r_{nk} in Equation (16.16) were replaced with $r_{nk} + \epsilon$ where $\epsilon = 0.0001$.

TIM 251

Bernouilli	Mixture	Model	after	each	itera	tion
DCITIOUIII	MINTAL	MOGC	aitti	Cacii	ILC I G	

-			<u> Deri</u>	<u> 10uiiii </u>	<u> IVIIXTUR</u>	<u>e woae</u>	<u>er arter</u>	<u>each</u>	<u>iteratio</u> r
	Parameter			Ite:	ration o	f cluster	ing		
		0	1	2	3	4	5	15	25
Class prior	α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
Glado prior	r _{1,1}		1.00	1.00	1.00	1.00	1.00	1.00	1.00
Class	$r_{2,1}$		0.50	0.79	0.99	1.00	1.00	1.00	1.00
Assignments			0.50	0.84	1.00	1.00	1.00	1.00	1.00
	$r_{4,1}$		0.50	0.75	0.04	1.00	1.00	1.00	1.00
r _{i1}	$r_{5,1}$		` '		ıment text		docID	documen	
E-Step	$r_{6,1}$	1.00	1			ocoa beans	7	sweet sug	
-	•	0.00	2		a ghana afi		8	sugar can	
	$r_{7,1}$	0.00	3		s harvest g	hana	9	sweet sug	
	$r_{8,1}$		4 5		a butter er truffles		10 11	sweet cak cake black	
	$r_{9,1}$		6		er trumes et chocolate	.	11	cake black	KTOTEST
	$r_{10,1}$				CHOCOLAN	•	I		
	$r_{11,1}$		0.50	0.57	0.58	0.41	0.07	0.00	0.00
	q _{africa,1}	0.000	0.100	0.134	0.158	0.158	0.169	0.200	
	9 _{africa,2}	0.000	0.083	0.042	0.001	0.000	0.000	0.000	
M-Step	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	·
<u>-</u>	q _{brazil,2}	0.000	0.167	0.195	0.213	0.214	0.196	0.167	
Word Class	q _{cocoa,1}	0.000	0.400	0.432	0.465	0.474	0.508	0.600	·
Conditionals	$q_{\rm cocoa,2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	
$Q_{word,class}$	9sugar,1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Sugar in class 2 only	q _{sugar,2}	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
	q _{sweet,1}	1.000	0.300	0.238	0.180	0.159	0.153	0.000	Ţ
TIM 251: Large-Scale	q _{sweet,2}	1.000	0.417	0.507	0.610	0.640	0.608	0.667	.

Bernouilli	Mixturo	Madal	oftor	aaah	itara	Han
Demoulli	MIXLUIE	Model	aitei	each	itera	LIOH

ı	ъ .	<u> </u>	Berr					eacn	<u>iteratio</u> i
	Parameter			Ite	ration o	t cluster	ing		
		0	1	2	3	4	5	15	25
Clase assi prisr	α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
0.0.00 p.1.01	$r_{1,1}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00
Class	$r_{2,1}$		0.50	0.70	0.00	1.00	1.00	1.00	1 00
Assignments			0.50	0. (a) $\frac{6}{1}$		ment text nocolate coco	doo a beans 7	cID docum sweet s	ent text augar
	$r_{4,1}$		0.50	0.	2 cocoa	ghana africa	8	sugar c	ane brazil
r _{i1}	$r_{5,1}$		0.50	0.		harvest ghar butter	na 9 10	sweet o	ugar beet ake icing
E-Step	$r_{6,1}$	1.00	1.00	1. 5		r truffles chocolate	11	cake bl	ack forest
	$r_{7,1}$	0.00	0.00	0.	0.00	0.00	0.00	0.00	0.00
	$r_{8,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{9,1}$		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$r_{10,1}$		0.50	0.40	0.14	0.01	0.00	0.00	0.00
	r _{11,1}		0.50	0.57	0.58	0.41	0.07	0.00	0.00
	q _{africa,1}	0.000	0.100	0.134	0.158	0.158	0.169	0.200	
	q _{africa,2}	0.000	0.083	0.042	0.001	0.000	0.000	0.000	
M-Step	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	·
-	q _{brazil,2}	0.000	0.167	0.195	0.213	0.214	0.196	0.167	
Word Class	q _{cocoa,1}	0.000	0.400	0.432	0.465	0.474	0.508	0.600	·
Conditionals	$q_{\rm cocoa,2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	
$Q_{word,class}$	g _{sugar,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
Sugar in class 2 only	$q_{ m sugar,2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
	q _{sweet,1}	1.000	0.300	0.238	0.180	0.159	0.153	0.000	<u> </u>
TIM 251: Large-Scale	q _{sweet,2}	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

Г	Parameter			Ite	ration o	f cluster	ing		
		0	1	2	3	4	5	15	25
- [α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
Г	$r_{1,1}$		1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$r_{2,1}$		0.50	0.70	0.00	1.00	1 00	1.00	1.00
	$r_{3,1}$		0.50			ment text hocolate coco	doo a beans 7	cID docum sweet s	ent text Sugar
	$r_{4,1}$		0.50			ghana africa		sugar o	cane brazil sugar beet
	$r_{5,1}$		0.50	0.	4 cocoa	s harvest ghai i butter	10	sweet	cake icing
	$r_{6,1}$	1.00	1.00			r truffles t chocolate	11	cake bl	ack forest
	$r_{7,1}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		4		!	-4- 4				
	Soft clu	ıster	ass	igme	ents t	or ea	ich a	locui	ment
	$r_{10,1}$		0.50	0.40	0.14	0.01	0.00	0.00	0.00
	$r_{11,1}$		0.50	0.57	0.58	0.41	0.07	0.00	0.00
	$q_{ m africa,1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	
	q _{africa,2}	0.000	0.083	0.042	0.001	0.000	0.000	0.000	
	q _{brazil,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
	q _{brazil,2}	0.000	0.167	0.195	0.213	0.214	0.196	0.167	
	q _{cocoa,1}	0.000	0.400	0.432	0.465	0.474	0.508	0.600	·
	q _{cocoa,2}	0.000	0.167	0.090	0.014	0.001	0.000	0.000	
	q _{sugar,1}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
y	$q_{ m sugar,2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
	q _{sweet,1}	1.000	0.300	0.238	0.180	0.159	0.153	0.000	·
le	q _{sweet,2}	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

E-Step

M-Step

Sugar in class 2 only

TIM 251: Large-Scale

Bernouilli Mixture Model after each iteration

						1177661	<u> </u>	<u> </u>		- Ca-oii	<u>itoratio</u>
	Parameter			I	ter	ation o	of cluster	ring			
		0	1	2		3	4	5		15	25
Class prior	α_1	0.50	0.45	0.53		0.57	0.58	0.54	4	0.45	nent text
Glade prior	r _{1,1}		1.00	1.	1 2		chocolate coco a ghana africa		7 8	sweet	sugar cane brazil
Class	$r_{2,1}$		0.50	0.	3	bean	s harvest gha		9	sweet	sugar beet
Assignments			0.50	0.	4 5		a butter er truffles		10 11		cake icing lack forest
r	$r_{4,1}$		0.50	0.	6	swee	et chocolate				
F Ston	$r_{5,1}$		0.50	0.52		0.66	0.91	1.00)	1.00	1.00
L-216 0	*										

Word Class conditionals Pr(Word=sugar|Class=2) = q_{sugar,2} Sugar in Doc7 and we only have one Doc in class 2

Iteration1: r9.

q_{sugar,2}=1 r_{10}

 r_{11}

 $q_{\rm af}$

 $q_{\rm bi}$

 $q_{\rm bi}$

 q_{cc}

1/1 #of docs in class 2 with Sweet / # of docs in class 2

9af → ?Smoothing

M-Step

Word Class Conditionals

Q_{word,class}

TIM 251: Large-Scale

Sugar in class 2 only

g_{cc} → E-Step is exactly the Bernoulli Naive Bayes

a _{sugar} 1	10.000	0.000	0.000	0.000	0.000	0.000	0.000
$q_{ m sugar,2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500
q _{sweet,1}	1.000	0.300	0.238	0.180	0.159	0.153	0.000
q _{sweet,2}	1.000	0.417	0.507	0.610	0.640	0.608	0.667