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

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

where $\Theta = \{\Theta_1, \ldots, \Theta_K\}$, $\Theta_k = (\alpha_k, q_{1k}, \ldots, q_{Mk})$, 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

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

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		

See Section 16.5 in The IR Book

Example taken from IR Book
Table 16.3

IR Book Table 16.3

Bernouilli Mixture Model after each iteration

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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
$r_{3,1}$		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
$q_{\text{africa},1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	0.200
$q_{\text{africa},2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	0.000
$q_{\text{brazil},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$q_{\text{brazil},2}$	0.000	0.167	0.195	0.213	0.214	0.196	0.167	0.167
$q_{\text{cocoa},1}$	0.000	0.400	0.432	0.465	0.474	0.508	0.600	0.600
$q_{\text{cocoa},2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	0.000
$q_{\text{sugar},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$q_{\text{sugar},2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	0.500
$q_{\text{sweet},1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	0.000
$q_{\text{sweet},2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	0.667

Class prior

Class Assignments r_{i1}

E-Step

M-Step

Word Class Conditionals $Q_{\text{word,class}}$

Sugar in class 2 only

IR Book Table 16.3

Bernouilli Mixture Model after each iteration

Class prior

Class Assignments r_{i1}

E-Step

M-Step

Word Class Conditionals $Q_{word,class}$

Sugar in class 2 only

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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
$r_{3,1}$		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						
$r_{6,1}$	1.00	1.00						
$r_{7,1}$	0.00	0.00						
$r_{8,1}$		0.00						
$r_{9,1}$		0.00						
$r_{10,1}$		0.50						
$r_{11,1}$		0.50						
$q_{\text{africa},1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	0.200
$q_{\text{africa},2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	0.000
$q_{\text{brazil},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$q_{\text{brazil},2}$	0.000	0.167	0.195	0.213	0.214	0.196	0.167	0.167
$q_{\text{cocoa},1}$	0.000	0.400	0.432	0.465	0.474	0.508	0.600	0.600
$q_{\text{cocoa},2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	0.000
$q_{\text{sugar},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$q_{\text{sugar},2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	0.500
$q_{\text{sweet},1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	0.000
$q_{\text{sweet},2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	0.667

(a)

docID	document text	docID	document text
1	hot chocolate cocoa beans	7	sweet sugar
2	cocoa ghana africa	8	sugar cane brazi
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		

Bernouilli Mixture Model after each iteration

IR Book Table 16.3

Class prior

Class Assignments r_{i1}

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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
$r_{3,1}$		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
$q_{\text{africa},1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	0.200
$q_{\text{africa},2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	0.000
$q_{\text{brazil},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Wor
Con
Q_w

Sugar in

TIM 251

► **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 $\epsilon = 0.0001$.

Bernouilli Mixture Model after each iteration

Class prior

Class Assignments r_{i1}

E-Step

M-Step

Word Class Conditionals $Q_{word,class}$

Sugar in class 2 only

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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
$r_{3,1}$		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
$r_{5,1}$								
$r_{6,1}$	1.00							
$r_{7,1}$	0.00							
$r_{8,1}$								
$r_{9,1}$								
$r_{10,1}$								
$r_{11,1}$		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	
$q_{brazil,1}$	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	
$q_{sugar,2}$	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	
$q_{sweet,2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

(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		

Bernouilli Mixture Model after each iteration

Class Priors

Class prior

Class Assignments r_{i1}

E-Step

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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.					
$r_{3,1}$		0.50	0.					
$r_{4,1}$		0.50	0.					
$r_{5,1}$		0.50	0.					
$r_{6,1}$	1.00	1.00	1.					
$r_{7,1}$	0.00	0.00	0.					
$r_{8,1}$		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
$q_{\text{africa},1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	
$q_{\text{africa},2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	
$q_{\text{brazil},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
$q_{\text{brazil},2}$	0.000	0.167	0.195	0.213	0.214	0.196	0.167	
$q_{\text{cocoa},1}$	0.000	0.400	0.432	0.465	0.474	0.508	0.600	
$q_{\text{cocoa},2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	
$q_{\text{sugar},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
$q_{\text{sugar},2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
$q_{\text{sweet},1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	
$q_{\text{sweet},2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

(a)

docID	document text
1	hot chocolate cocoa beans
2	cocoa ghana africa
3	beans harvest ghana
4	cocoa butter
5	butter truffles
6	sweet chocolate
7	sweet sugar
8	sugar cane brazil
9	sweet sugar beet
10	sweet cake icing
11	cake black forest

M-Step

Word Class Conditionals $Q_{\text{word,class}}$

Sugar in class 2 only

E-Step

M-Step

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
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	0.7					
$r_{3,1}$		0.50	0.					
$r_{4,1}$		0.50	0.					
$r_{5,1}$		0.50	0.					
$r_{6,1}$	1.00	1.00	1.					
$r_{7,1}$	0.00	0.00	0.					
$r_{10,1}$		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
$q_{\text{africa},1}$	0.000	0.100	0.134	0.158	0.158	0.169	0.200	
$q_{\text{africa},2}$	0.000	0.083	0.042	0.001	0.000	0.000	0.000	
$q_{\text{brazil},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
$q_{\text{brazil},2}$	0.000	0.167	0.195	0.213	0.214	0.196	0.167	
$q_{\text{cocoa},1}$	0.000	0.400	0.432	0.465	0.474	0.508	0.600	
$q_{\text{cocoa},2}$	0.000	0.167	0.090	0.014	0.001	0.000	0.000	
$q_{\text{sugar},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
$q_{\text{sugar},2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
$q_{\text{sweet},1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	
$q_{\text{sweet},2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

(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		

Soft cluster assigments for each document

Sugar in class 2 only

Bernouilli Mixture Model after each iteration

Class prior

Class Assignments r_{i1}

E-Step

Parameter	Iteration of clustering							
	0	1	2	3	4	5	15	25
α_1	0.50	0.45	0.53	0.57	0.58	0.54	0.45	
$r_{1,1}$		1.00	1.					
$r_{2,1}$		0.50	0.					
$r_{3,1}$		0.50	0.					
$r_{4,1}$		0.50	0.					
$r_{5,1}$		0.50	0.52	0.66	0.91	1.00	1.00	1.00
$r_{6,}$								
$r_{7,}$								
$r_{8,}$								
$r_{9,}$								
r_{10}								
r_{11}								
q_{af}								
q_{af}								
q_{br}								
q_{br}								
q_{cc}								
q_{cc}								
$q_{\text{sugar},1}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
$q_{\text{sugar},2}$	1.000	0.500	0.585	0.640	0.642	0.589	0.500	
$q_{\text{sweet},1}$	1.000	0.300	0.238	0.180	0.159	0.153	0.000	
$q_{\text{sweet},2}$	1.000	0.417	0.507	0.610	0.640	0.608	0.667	

nent 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		

Word Class conditionals Pr(Word=sugar|Class=2) = $q_{\text{sugar},2}$

Sugar in Doc7 and we only have one Doc in class 2

➔ **Iteration1:**

$q_{\text{sugar},2}$=1

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

➔ **?Smoothing**

➔ **E-Step is exactly the Bernoulli Naive Bayes**

M-Step

Word Class Conditionals $Q_{\text{word,class}}$

Sugar in class 2 only