Feedback — Week 2 Practice Quiz

Help Center

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Fri 19 Jun 2015 1:33 PM PDT**. You got a score of **12.00** out of **12.00**.

Question 1

You are given a unigram language model θ distributed over a vocabulary set V composed of **only** 4 words: "the", "machine", "learning", and "data". The distribution of θ is given in the table below:

w	$P(w\|\theta)$
machine	0.1
learning	0.2
data	0.3
the	0.4

$P(\text{“machine learning”}|\theta) =$

Your Answer		Score	Explanation
0.02	✔	1.00	
0.3			
0.004			
0.2			
Total		1.00 / 1.00	

Question Explanation

Since in a unigram language model words are assumed to be generated independently, we have
$P(\text{“machine learning”}|\theta) = P(\text{“machine”}|\theta)P(\text{“learning”}|\theta) = 0.1 * 0.2 = 0.02$

Question 2

Assume the same unigram language model as in Question 1. Then, $P(\text{“learning machine”}|\theta) =$

Your Answer		Score	Explanation
0.3			
0.02	✔	1.00	
0.2			
0.004			
Total		1.00 / 1.00	

Question Explanation

Due to the independence assumption, the word order does not matter when generating text based on a unigram language model. Thus, the answer is the same as that of Question 1.

Question 3

Assume the same unigram language model as in Question 1. Then, $P(\text{“learning machine learning”}|\theta) =$

Your Answer		Score	Explanation
0.02			
0.3			
0.004	✔	1.00	
0.2			
Total		1.00 / 1.00	

Question 4

Assume that words are being generated by a mixture of two unigram language models, θ_1 and θ_2, where $P(\theta_1) = 0.5$ and $P(\theta_2) = 0.5$. The distributions of the two models are given in the table below:

w	$P(w\|\theta_1)$	$P(w\|\theta_2)$
the	0.4	0.15
and	0.4	0.15
genes	0.05	0.3
biology	0.15	0.4

Then $P(\text{“biology”}) =$

Your Answer		Score	Explanation
0.275	✔	1.00	
0.55			
0.175			
0.15			
Total		1.00 / 1.00	

Question Explanation

$P(\text{“biology”}) = P(\text{“biology”}|\theta_1)P(\theta_1) + P(\text{“biology”}|\theta_2)P(\theta_2) = 0.15 * 0.5 + 0.4 * 0.5 = 0.275$

Question 5

Assuming the same given as in Question 4. Then $P(\text{“the biology”}) =$

Your Answer		Score	Explanation
1			
0.075625	✔	1.00	
0.275			
0.275625			
Total		1.00 / 1.00	

Question Explanation

$P(\text{“the biology”}) = P(\text{“the biology”}|\theta_1)P(\theta_1) + P(\text{“the biology”}|\theta_2)P(\theta_2) = P(\text{“the”}|\theta_1)P(\text{“biology”}|\theta_1)P(\theta_1) + P(\text{“the”}|\theta_2)P(\text{“biology”}|\theta_2$

Question 6

We want to run PLSA on a collection of N documents with a fixed number of topics k where the vocabulary size is M. What is the number of parameters that PLSA tries to estimate? Consider each $P(w|\theta_j)$ or $\pi_{d,j}$ as a separate parameter.

Your Answer		Score	Explanation
◉ $Mk + Nk$	✔	1.00	
○ Mk			
○ Nk			
○ MNk			
Total		1.00 / 1.00	

Question 7

Assume that the likelihood function of PLSA has multiple local maxima and one global maximum. There exists an initial set of parameters for which PLSA will converge to the global maximum of the likelihood function.

Your Answer		Score	Explanation
◉ True	✔	1.00	
○ False			
Total		1.00 / 1.00	

Question Explanation

If the initial parameters are “near” the optimal parameter values that achieve the global maximum of the likelihood function, then PLSA will converge towards the optimal parameters. In practice, determining initial parameter values that are near the global optimum is hard, thus we usually run PLSA several times using randomly sampled initial parameters and select those that attain the highest likelihood.

Question 8

In general, PLSA using the EM algorithm does not stop until it achieves the global maximum of the likelihood function.

Your Answer		Score	Explanation
◉ False	✔	1.00	
○ True			
Total		1.00 / 1.00	

Question Explanation

The EM algorithm is not guaranteed to achieve the global maximum. It will terminate when it reaches **any** maximum, whether local or global.

Question 9

Let $\theta_1, \ldots, \theta_k$ be the k unigram language models output by PLSA and V be the vocabulary set. Then, for any $i \in \{1, \ldots, k\}$, the following relation always holds : $\sum_{w \in V} P(w|\theta_i) = 1$.

Your Answer	Score	Explanation

False		
True ✔	1.00	
Total	1.00 / 1.00	

Question Explanation

This relation is valid since θ_i is a probability distribution.

Question 10

The EM algorithm **cannot** decrease the likelihood of the data.

Your Answer	Score	Explanation
False		
True ✔	1.00	
Total	1.00 / 1.00	

Question Explanation

The EM algorithm will keep on increasing the likelihood of the data until it converges.

Question 11

Suppose we have the following word counts for two documents d_1 and d_2.

Table 1: Counts for words in document set

Vocabulary Words	$c(w, d_1)$	$c(w, d_2)$	$P(w\|\theta_B)$
text	5	0	0.15
mining	4	0	0.05
the	4	4	0.50
fifa	0	4	0.10
football	0	3	0.20

We are interested in applying topic modeling to discover two topics, θ_0 and θ_1, in our corpus of two documents. Suppose that we run PLSA with the number of topics set to 2 (i.e. $k = 2$) while using an additional known (fixed) background word distribution θ_B as shown in Table 1. Using the EM algorithm, and after n iterations, the E-step gives the following estimates:

Table 2: Output of E-step after n iterations.

Documents	Words	$P(z_{w,d} = 0)$	$P(z_{w,d} = 1)$	$P(z_{w,d} = B)$
d_1	text	1.00	0.00	0.20
	mining	1.00	0.00	0.10
	the	0.60	0.40	0.90
d_2	the	0.40	0.60	0.90
	fifa	0.00	1.00	0.20
	football	0.00	1.00	0.20

Assume $\lambda_B = P(\theta_B) = 0.20$ and recall that $P(z_{w,d} = 0) + P(z_{w,d} = 1) = 1$ as discussed in the lectures. After completing the M-step, $P(\text{football}|\theta_1)$ =

Your Answer		Score	Explanation
0.2			
0.1			
0.4	✔	1.00	
0.3			
Total		1.00 / 1.00	

Question Explanation

We have that:
$P(\text{text}|\theta_1) \propto 5(1 - 0.20)(0.00) = 0.00$
$P(\text{mining}|\theta_1) \propto 4(1 - 0.10)(0.00) = 0.00$
$P(\text{the}|\theta_1) \propto 4(1 - 0.90)(0.60) + 4(1 - 0.90)(0.40) = 0.40$
$P(\text{fifa}|\theta_1) \propto 4(1 - 0.20)(1.00) = 3.20$
$P(\text{football}|\theta_1) \propto 3(1 - 0.20)(1.00) = 2.40$

Thus our normalizing term is $0.40 + 3.20 + 2.30 = 6.0$, then
$P(\text{football}|\theta_0) = 2.40/6.0 = 0.40$

Question 12

Assume the same given as in Question 11 and recall that $\pi_{d_2,0} + \pi_{d_2,1} = 1$. Then, $\pi_{d_2,1}$, rounded to the nearest hundredth, is equal to:

Your Answer		Score	Explanation
0.97	✔	1.00	
0.98			
0.99			
1			
Total		1.00 / 1.00	

Question Explanation

Computing the distribution for document 2 we have
$\pi_{d_2,1} \propto 3(1 - 0.20)(1.00) + 4(1 - 0.20)(1.00) + 4(1 - 0.90)(0.60) = 5.84$
$\pi_{d_2,0} \propto 3(1 - 0.20)(0.00) + 4(1 - 0.20)(0.00) + 4(1 - 0.90)(0.40) = 0.16$
Setting the normalization term as $5.84 + 0.16 = 6.00$, we can then compute the topic probabilities of document 2 as follows:
$\pi_{d_2,1} = 5.84/6.00 \approx 0.97$ $\pi_{d_2,0} = 0.16/6.00 \approx 0.03$