Natural Language Processing

Assignment- 9

TYPE OF QUESTION: MCQ

Number of questions: 10 **Total mark: 10 X 1 = 10**

Question 1. Which of the following is/are true about the LDA topic model?

1. Documents are a mixture of topics
2. Topics are a mixture of sentences
3. Using the probability distribution, topics generate the words.
4. LDA is a generative probabilistic model

Answer: 1,3, 4
Solution:
Option 2 is false as topics are not a mixture of sentences.

Question 2: In Topic modeling which hyperparameters tuning used for represents document-topic Density?

1. Dirichlet hyperparameter Beta
2. Dirichlet hyperparameter alpha
3. Number of Topics (K)
4. None of them

Answer: 2

Solution:
alpha is used to represent document-topic intensity

Question 3: Which of the following is/ are false about sLDA?

1. After training the LDA model, a supervised regression model is learned for mapping topic distributions to target class.
2. The target class is modeled as an observed random variable within the graphical model of LDA.
3. The target class is modeled as an unobserved random variable within the graphical model of LDA.
4. The target class is modeled as an unobserved fixed variable within the graphical model of LDA.

Answer: 1,3,4

Solution:
Refer lecture of week 9

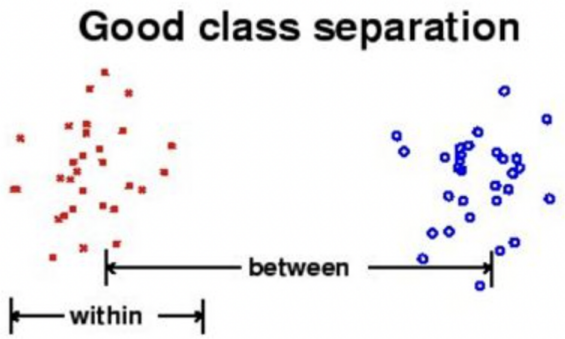
Question 4: Which of the following is/ are true?

1. LDA can not be applied in multiclass set up
2. PCA focuses on variance
3. Both PCA and LDA are unsupervised algorithm
4. The number of nonzero eigenvalues provide discriminatory directions in LDA

Answer: 2, 4
Solution:
1, 3 are false. LDA also can be extended to multiclass set up and it is a supervised algorithm

---

Question 5: Which of the following is true about LDA?



1. LDA aims to maximize the distance between class and minimize the within class distance
2. LDA aims to minimize both distance between class and distance within class
3. LDA aims to minimize the distance between class and maximize the distance within class
4. LDA aims to maximize both distance between class and distance within class

Answer: 1

Solution:

Option 1 is correct

For question 6 , 7 and 8 use the following information.
Suppose you are using Gibbs sampling to estimate the distributions, θ and β for topic models. The underlying corpus has 3 documents and 5 words, **{machine, learning, language, nature, vision}** and the number of topics is 2. At certain point, the structure of the documents looks like the following

Doc1: nature(1) language(1) vision(1) language(1) nature(1) nature(1) language(1) vision(1)
Doc2: nature(1) language(1) language(2) machine(2) vision(1) learning(2) language(1) nature(1)
Doc3: machine(2) language(2) learning(2) language(2) machine(2) machine(2) learning(2) language(2)

(number) –number inside the brackets denote the topic no. 1 and 2 denote whether the word is currently assigned to topics t1 and t2 respectively. η = 0.3 and α = 0.3

$$\beta_i^{(j)} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta} \qquad \theta_j^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

For question 6,7,8 calculate the value upto 4 decimal points and choose your answer

Question 6 : Using the above structure the estimated value of $\beta(2)_{\text{nature}}$ at this point is

1. 0.0240
2. 0.02459
3. 0.0260
4. 0.0234

Answer: 1

Solution:

	t1	t2
machine	0	4
nature	5	0
language	5	4
vision	3	0
learning	0	3

$\beta(2)_{nature} = (0+0.3)/(11+5*0.3) = 0.3/12.5 = 0.024$

Question 7 : Using the above structure the estimated value of θ_{t1}^{doc2}

1. 0.6562
2. 0.6162
3. 0.6385
4. 0.50000

Answer: 2

Solution:

	t1	t2
doc1	8	0
doc2	5	3
doc3	0	8

$\theta_{t1}^{doc2} = (5+0.3)/(8+2*0.3) = 5.3/ 8.6 = 0.6162$

Question 8 : Using the above structure the estimated value of θ_{t2}^{doc2}

1. 0.6562
2. 0.3975
3. 0.3837
4. 0.3707

Answer: 3

Solution:
Use the same formulae mentioned in Question 7 solution

Question 9 : Which of the following is/ are true ?

1. Dirichlet distribution is a family of exponential distribution
2. LDA is impacted by the order of documents
3. In LDA the number of latent clusters are identified automatically
4. All of the above are true

Answer: 1

Solution:
The order of documents does not matter in LDA, we need to identify the number of latent clusters in advance in the LDA topic model.

Question 10 :

In Gibbs sampling choose the correct option from below

1. It can not directly estimate the posterior distribution over z
2. It is a form of Markov chain Monte Carlo
3. Here sampling is done in parallel
4. Sampling is stopped before sampled values approximate the target distribution

Answer: 2

Solution:
In gibbs sampling, we do sequential sampling until the sampled values approximate the target distribution. This also can directly estimate the posterior distribution over z

---