Natural Language Processing

Assignment- 9

TYPE OF QUESTION: MCQ

Number of questions: 8

[Question 7 and 8 carries 2 marks each] **Total mark: 6 X 1 + 2 X 2 = 10**

Question1. Which of the following is/are true?

1. Topic modelling discovers the hidden themes that pervade the collection
2. Topic modelling is a generative model
3. Dirichlet hyperparameter Beta used to represent document-topic Density?
4. None of the above

Answer: 1,2

Question2. Which of the following is/are true?

1. The Dirichlet is an exponential family distribution on the simplex positive and negative vectors sum to one
2. Correlated Topic Model (CTM) predicts better via correlated topics
3. LDA provides better fit than CTM
4. CTM draws topic distributions from a logistic normal

Answer: 2, 4

Solution: Refer Lecture 44

Question 3: You have a topic model with the parameters $\alpha = 0.89$ and $\beta = 0.04$. Now, if you want to have sparser distribution over words and denser distribution over topics, what should be the values for α and β?

1. Both α and β values should be decreased
2. Both α and β values should be increased
3. α should be decreased, but β should be increased

4. α should be increased, but β should be decreased

Answer: 4

Solution:
α : topic distribution
β : word distribution

Question4: Which of the following is/are false about LDA assumption?

1. LDA assumes that the order of documents matter
2. LDA is not appropriate for corpora that spans hundreds of years
3. LDA assumes that documents are a mixture of topics and topics are a mixture of words
4. LDA can decide on the number of topics by itself.

Answer: 1,4

Solution: Refer Lecture 44

Question 5: Which of the following is/are True about Relational Topic Model (RTM) ?

1. RTM formulation ensures that the same latent topic assignments used to generate the content of the documents
2. In RTM, link function models each per-pair binary variable as linear regression
3. In RTM, covariates are constructed by the Hadamard product
4. Link probability function is dependant on the topic assignments that generated their words

Answer: 1,3,4

Solution: Refer Lecture 45

Question 6:
Classically, topic models are introduced in the text analysis community for_____ topic discovery in a corpus of documents.

1. Unsupervised.

2. Supervised.
3. Semi-automated.
4. None of the above.

Answer - 1. Unsupervised

Question 7: Which of the following is/are False about Gibbs Sampling?

1. Gibbs sampling is a form of Markov chain Monte Carlo (MCMC)
2. Sampling is done sequentially and proceeds until the sampled values approximate the target distribution
3. It can not estimate the posterior distribution directly
4. Gibbs sampling falls under the category of variational methods

Answer: 3,4
Solution: Refer Gibbs Sampling slide

For question 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. $\eta = 0.3$ and $\alpha = 0.3$

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

Question 8 : 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
