Week 4 Quiz

TOTAL POINTS 10

1. What is NOT the motivation for text clustering? 1 point

- [ ] To link similar documents and remove duplicated documents
- [ ] To remove spam documents based on a small collection of human annotated spam documents
- [x] To create structure of text data
- [ ] To quickly get an idea about a large collection of documents

2. What is TRUE about the mixture model and topic modeling? 1 point

- [ ] Topic modeling can also be used for document clustering directly.
- [x] In topic modeling, the topic of each word is independently sampled, while in the mixture model, only one topic is drawn for each document.
- [ ] Only topic modeling can learn topics, while the mixture model does not yield such information after learning.

3. In the mixture model, if we want to encourage the formation of a large cluster: 1 point

- [ ] Use a smaller number of clusters for training
- [ ] Try different initialization
- [x] Add prior to $P(\theta)$ so that the distribution is skewed

4. In the EM algorithm, which step improves the model likelihood? 1 point

- [ ] E-step
- [x] M-step

5. True or false? In the EM algorithm, the model likelihood monotonically increases. 1 point

- [x] True
- [ ] False

6. What is the most difficult part of directly applying maximal likelihood to PLSA? 1 point

- [x] The objective function needs to sum over all topics for each word.
- [ ] The objective function needs to sum over all documents in the collection.
- [ ] The objective function needs to sum over all words for each document.

7. For the agglomerative clustering algorithm, which of the following is not TRUE? 1 point

- [x] The depth of the hierarchy is always $log_2(N)$ where N is the number of items.
- [ ] It's a bottom-up algorithm to form a hierarchy.
- [ ] The user needs to specify a similarity measurement.

8. Which evaluation method is best for clustering results of a large collection of documents? 1 point

- [x] Use the indirect evaluation method and test performance for an application with or without clustering.
- [ ] Use the direct evaluation method and create human annotations for each document in the collection.

9. Which of the following is NOT sensitive to outliers? 1 point

- [ ] Average-link
- [x] Single-link
- [ ] Complete-link

10. Which of the following is a generative classification algorithm? 1 point

- ○ K-NN
- ◉ Naive Bayes
- ○ SVM
- ○ Logistic Regression

☐ I, **BAL KRISHNA NYAUPANE**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code