Probabilistic Models and Machine Learning - Fall 2023
Homework 2

Due: Friday November 3rd, 2023 – 11:59pm EST

The total page limit for Problem 1 is three pages, though you may use extra pages for figures, and your code can be any length. Please use the LaTeX template on the website. You should zip your writeup and code into one file and submit it on Gradescope.

Problem 1

Choose one:

- Implement Gibbs sampling for a mixture model. You can implement a Gaussian mixture model as we discussed in class, or a different mixture model instead. Apply your code to a real-world data set and discuss what you learned, and what challenges you encountered while building and fitting your model. You can plot and discuss whatever you like. Among the plots, we would like to see $\log p(x_{1:n}, z_{1:n}, \theta_{1:K}, \beta_{1:K})$ as a function of iteration. (This is one way to diagnose if the Gibbs sampler converges.)
- Implement variational inference for a mixed-membership model or for a mixture model. Apply your code to real-data and discuss what you learned, and what challenges you encountered while building and fitting your model. You can plot and discuss whatever you like. Among the plots, we would like to see a figure depicting the convergence of your inference procedure (for example, the ELBO as a function of iteration). We would also like to see a plot that checks the model on held-out data.

We encourage you to apply your code to datasets that aren't widely used in machine learning, and to share public datasets you learn about in the `datasets` Slack channel.

You are free to use any programming libraries of your choice. As a caveat though, relying on library code and not your own code might make it harder to interpret and discuss the results.

Problem 2

Write an aspirational abstract describing the final project you want to complete. In the abstract, you should clearly state your motivating problem, and how you will use probabilistic models and data to answer this problem. Note that you are not committed to deliver everything you mention on the abstract. Rather, preparing the abstract is a chance to think concretely and envision a successful final project. We encourage you to refer to computer science conferences (such as *Neural Information Processing Systems* and *International Conference of Machine Learning*) or journals (such as *The Annals of Applied Statistics*, *Journal of the American Statistical Association*, *Journal of Machine Learning Research*) to get a sense of how to write an abstract.

You can find examples of datasets for Problem 1 below. Feel free to other datasets of your choice.

Senate This dataset contains Senate voting data. After unzipping, you'll find two files in the folder senate: `votes.csv` and `senators.txt`.

Each of the n = 103 rows of `votes.csv` contains the voting record of a different member of the 113'th session of the United States senate. The columns correspond to d = 657 bills that were voted on: 1 indicates a 'yea' vote, 0 indicates a 'nay' vote, and -1 indicates that the particular member did not vote.

`senators.txt` contains the names of the 103 senators, in the same order they appear on `votes.csv`. Each senator's name contains his/her political party (D, R, or I) and state; for example, Schumer (D-NY) indicates that Schumer is a Democratic senator from New York.

AP This dataset contains the text of 2,246 articles from the Associated Press.

After unzipping, you'll find three files in the folder `ap`: `ap.dat`, `ap.txt`, and `vocab.txt`.

`ap.txt` is an XML file that contains the full text of every article.

You'll probably want to work with `ap.dat`, which contains the counts of each word for each article. Every line is a different article in a bag-of-words format. The first number in each line is the total number of words in that article. Following this number, the rest of the line contains word counts in the format `word_index:count`. For example, the line "`5 0:1 5:2 140:1 2031:1`" indicates a 5-word article that has 1 occurrence of word 0, 2 occurrences of word 5, 1 occurrences of word 140, and 1 occurrence of word 2031.

Finally, the file `vocab.txt` contains a list of each word, zero-indexed to match the indices in `ap.dat`.

Mixture models and mixed-membership models are appropriate for this dataset. Note that you may want to remove the most common words (stopwords) along with very rare words in order to ease computation and produce more interpretable mixtures.

NeurIPS Abstracts This dataset contains the abstracts of the 2,834 papers in the 2022 NeurIPS conference proceedings. After unzipping, you will find one text file per abstract, containing the plain text of that abstract.

If you use this dataset for Problem 1, it will be interesting to consider how the results of your analysis can help you identify conference papers that are relevant to your own final project.

Your own dataset We encourage you to use a dataset that you might want to use for your final project, or another dataset that you are curious about. Please make sure to quickly describe the data including:

- Description of the dataset:
- Description of the features (numbers, names if possible):
- Description of the response variable:
- Number of observations:
- (Optional) Link if the data is public: