Probabilistic Models and Machine Learning - Fall 2023
Homework 3

Due: Tuesday November 21st, 2023 – 11:59pm

The total page limit for the homework 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 course website. You should zip your writeup and code into one file and submit it on Gradescope. If you reference your code in your writeup, please also append your code as an appendix in the writeup pdf.

Problem

Apply an inference method discussed in class to a probabilistic model of a real-world data set. The data set should not be overly simplistic (e.g. it should contain multiple variables, and at least several hundred observations). Please do not repeat the (model, inference method) pair you used for the previous homework assignments. For instance if you implemented a Gibbs sampler to fit a Gaussian mixture model in Homework 2, do not implement a Gibbs sampler to fit a Gaussian mixture model for this assignment. A few possible suggestions are given below.

- Perform MAP estimation (or any inference method you prefer) for a matrix factorization model, for instance a Gausssian matrix factorization for the `MovieLens` data set or an ideal point model for the `Senate` data set.
- Perform posterior inference for a model related to your final project.

In your writeup, we would like to see:

- a brief description of your data set
- a clear statement of the probabilistic model and inference method you are using
- a plot showing the convergence of the inference algorithm
- an evaluation of the goodness of fit of the model (e.g. log posterior predictive)
- some explanation/jusification of your choice of hyperparameters/tuning parameters
- an interpretation of the latent variables (e.g. examining the topics in LDA).

You may also plot and discuss anything else that you think is interesting.

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.

MovieLens This data set contains movie ratings submitted by users on MovieLens, a movie recommendation website. The data set contains 100,000 ratings applied to 9,000 movies by 600 users. There is a README.txt file in `movielens.zip` that contains more information about the data set.