Probabilistic Models and Machine Learning Fall 2023: Homework 1

Due: Friday October 6th, 2023 – 11:59pm ET

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). Aim for clarity and conciseness (shorter explanations are better). Please use the LaTeX template on the website. You should zip your writeup and code into one file and submit it on Gradescope (you should be added already).

There are two problems.

Problem 1

Implement a stochastic MAP estimation for a regression model of your choice. You can implement a Bayesian linear regression model as we discussed in class, or a different regression model instead (e.g. a Bayesian logistic regression).

Apply your code to a real-world data set and discuss what you learned. You can plot and discuss whatever you like. Among the plots and the discussion, we would like to hear about:

- the influence of the prior on the model and how you selected the hyperparameters
- the predictive log-likelihood as a function of iteration
- an interpretation of the posterior coefficients
- the influence of batch size and learning rate on the optimization

You should provide plots and metrics that support your discussion.

You can find an example of dataset below. Feel free to use it or another dataset of your choice.

Getting through this exercise is important for having a good final project and, more generally, becoming fluent in the material. There are many "gotchas" in developing and deploying probabilistic models, which are only learned from experience. (For example, you may want to work in log space, only exponentiating when you need to.)

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.

Datasets for Problem 1

IMDB dataset The IMDB dataset contains 50k movie reviews and sentiment labels. The goal is to predict the binary sentiments (positive/negative) from the text of the reviews. Since the response is a binary variable, it may make sense to use a Bayesian logistic regression.

You'll probably want to encode each review using a bag-of-word representation, and then normalize the features using a tf-idf. Note that you may want to remove the most common words (stopwords) along with very rare words.

https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

Your 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:**

Problem 2

This problem is intended to help you brainstorm ideas for the project. Consider some data. If you have a data set in mind for your final project then we encourage you to use it for this exercise as well.

a) **Variables in the data.** What are the variables in the data and what are some of their relationships to each other? Do you expect some of the variables to be correlated? Do you expect others to be (conditionally) independent?

b) **Latent variables.** What are some latent variables you could introduce to capture the correlations between model variables? What are some latent variables that could summarize aspects of the data? What other latent variables could be hidden in the data?

c) **Research question.** Formulate several questions you might be able to answer with the data. Write down the three most interesting ones.