Math 189r
Homework 3
November 21, 2016

There are 4 problems in this set. You need to do 2 problems the first week and 2 the second week. Instead of a fifth or sixth problem, **you are encouraged to work on your final project**. Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though. When implementing algorithms you may not use any library (such as `sklearn`) that already implements the algorithms but you may use any other library for data cleaning and numeric purposes (`numpy` or `pandas`). Use common sense. Problems are in no specific order.

1 (Gaussian Mixture Model) Consider the generative process for a Gaussian Mixture Model:

(1) Draw $z_i \sim \text{Cat}(\boldsymbol{\pi})$ for $i = 1, 2, \dots, n$.

(2) Draw $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$ for $i = 1, 2, \dots, n$.

Note that z_i is unobserved but $\mathbf{x}_i$ is observed. Express this model as a directed graphical model, first 'unrolled' and then using Plate notation, before answering the following questions. Support all claims.

(a) Is $\boldsymbol{\pi}$ independent of $\boldsymbol{\mu}_{z_i}$ or $\boldsymbol{\Sigma}_{z_i}$ given your dataset $\mathcal{D} = \{\mathbf{x}_i\}$? Does the posterior distribution over $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ and $\boldsymbol{\pi}$ factorize? How does this change what inference procedure we need to use for this model?

(b) If z_i were observed, would this change? Would the posterior then factorize? *Hint:* what other model have we studied that corresponds to observing z_i?

(c) Find the maximum likelihood estimates for $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ *if* the latent variables z_i were observed.

2 (Linear Regression) Consider the Bayesian Linear Regression model with the following generative process:

(1) Draw $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_0)$

(2) Draw $\mathbf{y}_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ for $i = 1, 2, \dots, n$ where σ^2 is known.

Express this model as a directed graphical model using Plate notation. Is $\mathbf{y}_i$ independent of $\mathbf{w}$? Is $\mathbf{y}_i$ independent of $\mathbf{w}$ *given* $\mathcal{D} = \{\mathbf{x}_i\}$? Support these claims.

3 (Collaborative Filtering) Consider the 'ratings' matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ with the low rank approximation $\mathbf{R} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ live in $\mathbb{R}^{n \times k}$ where we have k latent factors. Define our optimization problem as

$$\text{minimize: } f(\mathbf{U}, \mathbf{V}) = \|\mathbf{R} - \mathbf{U}\mathbf{V}^\top\|_2^2 + \lambda\|\mathbf{U}\|_2^2 + \gamma\|\mathbf{V}\|_2^2$$

where $\|\cdot\|_2$ in this case is the Frobenius norm $\|\mathbf{R}\|_2^2 = \sum_{ij} \mathbf{R}_{ij}^2$. Derive the gradient of f with respect to $\mathbf{U}_i$ and $\mathbf{V}_j$. Derive a stochastic approximation to this gradient where you consider a single data point at a time.

4 (Non-Negative Matrix Factorization) Consider the dataset at `http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html`. Choosing an appropriate objective function and algorithm from Lee and Seung 2001[1] implement Non-Negative Matrix Factorization for topic modelling (choose an appropriate number of topics/latent features) and assert that the convergence properties proved in the paper hold. Display the 20 most relevant words for each of the topics you discover.

[1] `https://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf`