Assignment 3

Yang Can, HKUST

2021/10/27

Problem 1

Consider the problem motivating the James-Stein Estimator: $z_i|\mu_i \sim \mathcal{N}(\mu_i, 1)$, $i = 1, \ldots, N.$, how to obtain a good estimate of $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_N]^T$ from $\mathbf{z} = [z_1, \ldots, z_N]^T$?

(a) Assign prior distribution $g(\mu)$: $\mu_i \sim \mathcal{N}(0, \alpha^{-1}), i = 1, \ldots, N$. Now consider $\boldsymbol{\mu}$ as the latent variable and α^{-1} as the model parameter. Derive an EM algorithm for parameter estimation and then use the estimated parameter $\hat{\alpha}$ to obtain an estimate of $\boldsymbol{\mu}$.

(b) Implement the above EM algorithm and compare the estimate with the result obtained from the James-Stein estimator.

Problem 2

Let $\mathbf{y} = [y_1, \ldots, y_K]^T$ be a vector of random variables from the multinomial distribution

$$\text{Mult}(\mathbf{y}|\boldsymbol{\mu}, N) = \binom{N}{y_1\, y_2 \ldots y_K} \prod_{k=1}^{K} \mu_k^{y_k},$$

where$0 \leq \mu_k \leq 1$, $\sum_{k=1}^{K} \mu_k = 1$ and $\sum_{k=1}^{K} y_k = N$. Assume that the piror distribution of $\boldsymbol{\mu}$ is the Dirichlet distribution

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

where $\alpha_k > 0$.

(a) Suppose we have collected a data set $\mathcal{D} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, where

$$\mathbf{y}_i \sim \text{Mult}(\mathbf{y}_i|\boldsymbol{\mu}_i, N), \quad \boldsymbol{\mu}_i \sim \text{Dir}(\boldsymbol{\mu}_i|\boldsymbol{\alpha})$$

Derive an EM algorithm to estimate $\boldsymbol{\alpha} \in \mathbb{R}^K$, where $\boldsymbol{\mu}$ is viewed as the latent variable. Let $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \log p(\mathcal{D}|\boldsymbol{\alpha}) = \arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \log p(\mathbf{y}_i|\boldsymbol{\alpha})$ be the MLE, where

$$p(\mathbf{y}_i|\boldsymbol{\alpha}) = \int p(\mathbf{y}_i|\boldsymbol{\mu}_i) p(\boldsymbol{\mu}_i|\boldsymbol{\alpha}) d\boldsymbol{\mu}_i.$$

Can you guarantee that your EM algorithm converges to MLE $\hat{\boldsymbol{\alpha}}$ (or equivalently, is this a convex optimization problem)? Explain your reason.

(b) Consider the generative model is as follows:

$$\mathbf{y}_i \sim \text{Mult}(\mathbf{y}_i|\boldsymbol{\mu}_i, N), \quad \boldsymbol{\mu}_i \sim \text{Dir}(\boldsymbol{\mu}_i|\boldsymbol{\alpha}_i), \quad \boldsymbol{\alpha}_i = \exp(\boldsymbol{\beta}^T \mathbf{x}_i).$$

where the parameter $\boldsymbol{\alpha}_i$ in the Dirichlet prior is modulated by side information encoded in $\mathbf{x} \in \mathbb{R}^p$ and $\boldsymbol{\beta}$ is a $p \times K$ matrix. Please derive an EM algorithm to estimate $\boldsymbol{\beta}$.

(c) Suppose the data set $\mathcal{D} = \{\mathbf{y}_i, \mathbf{x}_i\}_{i=1,\ldots,n}$ is generated via the probabilistic model in (b), where $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,4}]^T$ is a vector of $K = 4$ random variables and $\sum_{k=1}^{K} y_{i,k} = 50$, $\mathbf{x}_i$ is a vector of $p = 5$ random variables. Apply your algorithm derived in (b) to the given data set $\mathcal{D}$ to estimate $\boldsymbol{\beta} \in \mathbb{R}^{5\times 4}$. The data set is given in *data.txt*.

Problem 3

Consider a three-variance-component model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e},$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of responses, $\mathbf{X}_1 \in \mathbb{R}^{n\times p_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n\times p_2}$ are the design matrices, $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ are two vectors of random effects and $\mathbf{e}$ is the vector of independent noise. We assume that

$$\boldsymbol{\beta}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I}_{p_1}), \boldsymbol{\beta}_2 \sim \mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I}_{p_2}), \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}_n).$$

Please derive a standard EM algorithm and a PXEM algorithm to estimate parameters $\boldsymbol{\theta} = \{\sigma_1^2, \sigma_2^2, \sigma_e^2\}$ based on the observed data set D, and compare their computational efficiency. You should see the monotonic increase of $\log p(D|\boldsymbol{\theta})$ and use the same criterion to check the convergence of your algorithms, i.e., the change of the marginal likelihood

$$\log p(D|\boldsymbol{\theta}_{new}) - \log p(D|\boldsymbol{\theta}_{old}) \leq 10^{-6}.$$

You need to apply your algorithm to *data_3vc.txt*, where $n = 300$, $p_1 = 500$, $p_2 = 500$. The first column correspond to $\mathbf{y}$ with column name 'y', the 2th–501th columns correspond to $\mathbf{X}_1$ with column names 'X1.1'–'X1.500', and the 502th–1001 columns correspond to $\mathbf{X}_2$ with column names 'X2.1'–'X2.500', respectively.

Problem 4

Besides the gradient boosting, Bagging (bootstrap aggregating of multiple trees) and Random Forest are two alternative approaches for ensembling learning. Let p be the number of variables for classification or regression problems. The only difference between Bagging and Random Forest is the *mtry* parameter, where $mtry = p$ in Bagging and $mtry = \sqrt{p}$ and $p/3$ for classification and regression in Random Forest, respectively. There is a claim that **the *mtry* parameter in Random Forest plays a role of inexplicit regularization**. Please provide your own view with some supporting evidence. You may use "Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success" https://arxiv.org/abs/1911.00190 as a reference. The answer to this question is quite open.

Requirement

- You need to submit a report, in which you should clearly describe your method and explain your idea. The code should also be included.
- You can use R, Python or Matlab for coding.
- Your report should be in the **pdf** or **html** format, which is automatically generated by either R markdown or Jupyter notebook.
- The report is due to Nov, 10, 11:59 pm, 2021 (HK time).