Minor Exam

ELL888

Indian Institute of Technology Delhi

- Date: 17/02/2022
- Submit by 23rd February 2022 midnight
- Submission on Gradescope
- Marks: 100

Plagiarism in any form is not acceptable, you will be straightforwardly awarded zero marks without explanation. You will receive full marks for the **right approach** and steps even if you fail to answer correctly.

For each of these problems. Chose a suitable dataset write your own code and demonstrate the effectiveness of your algorithm. Credits will be given on the choice of the dataset, efficiency of algorithm, and details of the discussion.

Graph Learning: Given $\mathbf{X} \in \mathbb{R}^{n\times d}$ whose rows reside on the vertices of an unknown graph.

$$X = \begin{bmatrix} - \mathbf{x}_1 \in \mathbb{R}^d - \\ - \mathbf{x}_2 \in \mathbb{R}^d - \\ \vdots \\ - \mathbf{x}_n \in \mathbb{R}^d - \end{bmatrix}$$

The graph learning from data simply means finding edge weight matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, where its element e.g., $w_{ij} = [A]_{ij}$ captures the relation between any two pair of vertices (i, j) or data points $(\mathbf{x}_i, \mathbf{x}_j)$. We learn the graph weights $\mathbf{w} \in \mathbb{R}^{n(n-1)/2}$ under certain assumptions, like smoothness, probabilistic distribution, nearest neighborhood, etc. The focus here is an undirected graph which implies the graph matrix is symmetric.
Now consider the following problems.

- **Q 1** Graph learning with missing data: **(25 Marks)**

$$X = \begin{bmatrix} \mathbf{x}_1 = [x_{11}, x_{12}, \bullet, x_{14}, \bullet, \bullet, x_{1d}] \\ \mathbf{x}_2 = [x_{21}, \bullet, x_{23}, \bullet, x_{25}, \bullet, \bullet, x_{2d}] \\ \vdots \\ \mathbf{x}_n = [\bullet, x_{n2}, \bullet, x_{n4}, \bullet, \bullet, x_{nd}] \end{bmatrix}$$

  where some parts of the data are missing at random and $\bullet$ indicates missing data. Explain in detail how can we learn graph matrix with missing data. Put a descriptive detail.

- **Q 2** Semi-Supervised Setting. Now consider a semi-supervised graph learning problem. Now with the data $\mathbf{X}$ sitting at n vertices we also have the label information for $k < n$ vertices. Simply, we have $\mathbf{X} \in \mathbb{R}^{n\times d}$ with label information for k vertices $\mathbf{y} \in \{0, 1\}^k$. Explain in detail how can we learn the graph matrix integrating the label information. The dataset available are $\left(\{\mathbf{x}_i, y_i\}_{i=1}^k, \{\mathbf{x}_j\}_{j=k+1}^n\right)$. **(25 Marks)**

- **Q 3** Consider a massive data scenario, where n and d both are very large and typically $n >> d$. Such that it is not possible to process the entire data every iteration. Explain how can we approach the graph learning problem under such a setting. (Hint: Stochastic gradient descent based approaches try to address such problems, where each iteration is performed considering only a small subset of the dataset namely minibatch.) (**25 Marks**)

- **Q 4** Consider a heterogeneous data scenario, where dataset $\mathbf{X}$ associated with vertices may belong to set of real $\mathbb{R}$, set of integers $\mathbb{I}$, and categorical $\mathbb{C} = \{c_1, c_2, \ldots, c_k\}$. More concretely, $\mathbf{X} \in \mathbb{R}^{n \times d_r} \times \mathbb{I}^{n \times d_i} \times \mathbb{C}^{n \times d_c}$ and $d = d_r + d_i + d_c$. For example $\mathbf{x}_1 = [x_{11}, x_{12}, \ldots, x_{1d_r} \in \mathbb{R}, x_{1d_r+1}, x_{1d_r+2}, \ldots, x_{1d_r+d_i} \in \mathbb{I}, x_{1d_r+d_i+1}, x_{1d_r+d_i+2}, \ldots, x_{1d_r+d_i+d_c} \in \mathbb{C}\,]$. Explain in detail how can we learn graph matrix with heteroegenous data seeting. Put a descriptive detail. (**25 Marks**)

$\star\star\star$

Good Luck!