Question 3, Assignment 2: CS 754, Spring 2024-25

Amitesh Shekhar
IIT Bombay
22b0014@iitb.ac.in

Anupam Rawat
IIT Bombay
22b3982@iitb.ac.in

Toshan Achintya Golla
IIT Bombay
22b2234@iitb.ac.in

February 19, 2025

Declaration: The work submitted is our own, and we have adhered to the principles of academic honesty while completing and submitting this work. We have not referred to any unauthorized sources, and we have not used generative AI tools for the work submitted here.

1. Perform a google search to find out a research paper that uses group testing in data science or machine learning. Explain (i) the specific ML/DS problem targeted in the paper, (ii) the pooling matrix, measurement vector and unknown signal vector in the context of the problem being solved in this paper, (iii) the algorithm used in the paper to solve this problem. You can also refer to references within chapter 1 of the book `https://arxiv.org/abs/1902.06002`. [16 points]

Soln:

Research Paper:- **Neural Group Testing to Accelerate Deep Learning**, *Weixin Liang and James Zou, 2021*

(a) **The specific ML/DS problem targeted in the paper**:- The paper addresses the computational inefficiency in deep learning, particularly in large-scale neural networks that require significant resources. We're interested in the efficient data processing and reducing the computational load during the inference phase without sacrificing performance. To achieve this, we use Group testing to accelerate the detection of relevant patterns.

(b) **Pooling Matrix, Measurement Vector, and Unknown Signal Vector**:-
1. Pooling Matrix P: The pooling matrix is used to map the groupings of samples (or features) for group testing. It is a binary matrix where rows represent pooled groups and columns represent the individual samples. If sample x_j is part of group i, then $P_{ij} = 1$, otherwise $P_{ij} = 0$.

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 1 & \dots \\ 1 & 1 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

2. Measurement Vector y:
The measurement vector contains the outputs (responses) corresponding to each pooled group. If the group is defective, the measurement will be 1, otherwise 0.

$$\mathbf{y} = \mathbf{P}\mathbf{x} + \eta$$

Where:

- $\mathbf{x}$ is the unknown signal vector (containing the true feature values of the individual samples). $x \in \mathbb{R}^{n\times 1}$
- $\mathbf{n}$ is noise (representing errors or uncertainties in measurement). $\eta \in \mathbb{R}^{m\times 1}$
- $\mathbf{Px}$ gives the measurement results for each pooled group, where $P \in \mathbb{R}^{m\times n}$ &
- $y \in \mathbb{R}^{m\times 1}$

3. Unknown Signal Vector x:
The unknown signal vector $\mathbf{x}$ contains the individual features or signals for each sample. We aim to reconstruct this vector from the pooled group tests.

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

(c) **Algorithm Used**:-
The paper introduces a neural network and group testing based algorithm. The algorithm is described below:-

1. Group Testing Neural Network (GTNN):

A neural network architecture is trained to predict the unknown signal vector $\mathbf{x}$ given the measurements $\mathbf{y}$. Computational costs are reduced by incorporating pooling strategies, where input data is processed in groups rather than individually. The neural network solves the optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{P}\mathbf{x} - \mathbf{y}\|_2^2$$

2. Regularization and Loss Function:

A regularized loss function is often used to avoid overfitting and to promote sparsity in the solution (since group testing generally assumes most of the signal vector components are zero):

$$\mathcal{L}(\hat{\mathbf{x}}) = \|\mathbf{P}\hat{\mathbf{x}} - \mathbf{y}\|_2^2 + \lambda\|\hat{\mathbf{x}}\|_1$$

Where:

- λ is a regularization parameter controlling the sparsity of the estimated signal vector $\hat{\mathbf{x}}$.
- $\|\hat{\mathbf{x}}\|_1$ is the l_1-norm, promoting sparsity in the vector, which is typical for group testing problems.

3. Reconstruction of the Signal:

Once the neural network is trained, the estimated signal $\hat{\mathbf{x}}$ is used to reconstruct the individual features of the samples. The network learns to map the pooled measurements back to the original feature space, allowing for accurate reconstruction of $\mathbf{x}$.

In some cases, iterative methods like **belief propagation** or **compressed sensing algorithms** are employed to further refine the estimates of $\hat{\mathbf{x}}$.

(d) **Conclusion**:- This method enables deep learning models to process large datasets efficiently by reducing the number of measurements required, making it highly applicable in domains where collecting large amounts of labeled data is costly or time-consuming.