Indian Institute of Technology, Jodhpur

Dependable AI | Assignment 1

Topic : Bias - Detection, Mitigation and Evaluation

Total Marks : 180 + 20 Marks(Report) + 30 Marks(Viva)

Submission Policy and Requirements :

- Any kind of plagiarism is not accepted. We will strictly follow institute policies for plagiarism.
- Recommended programming languages: Python + Keras/TensorFlow/PyTorch.
- You may use any external libraries or GitHub codes. However, the evaluation will test your knowledge of the algorithm and the choice of hyperparameters. Do cite the libraries/codes.
- Submission should include: Working code for each of the parts separately and a report to show the analysis of results in each of the parts.

Assessment criterion:

The assessment will be done on the basis of the following components:

- Working codes
- Analysis and clarity of results (drawing comparisons across different parts) & clarity of the report
- Understanding the theoretical concepts and the choice of hyperparameters.

Guidelines for Submission:

- A single report(pdf) for all questions.
- Mention all the relevant results, comparisons as asked or wherever required for better understanding for the results.
- A single zip file containing the report, codes and readme if required

---

Q1. Analysis of Machine Learning model for bias. **[50 Marks]**

- Use the SVM model. (You can use sklearn library)
- Train the model on Bollywood celebrity dataset [Download Here]
- Choose any 10 classes from the dataset of your choice
- Report/ Show
    - Class-wise accuracy [10 Marks]
    - Overall accuracy [10 Marks]
    - Training loss vs Testing loss curve wrt epochs.[10 Marks]
    - Check if your model is biased or not by using at-least **2 metrics** ex. Confusion Matrix [15 Marks]
    - What type of bias you see(if any), explain. [5 Marks]

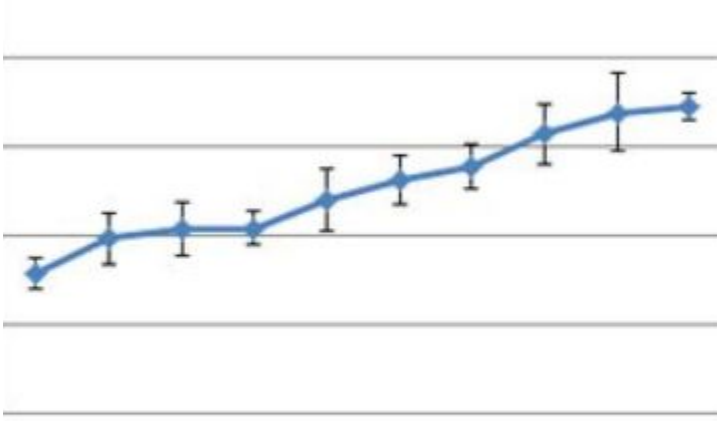
Q2. Evaluation Metrics for Bias Detection. **[50 Marks]**

Dataset : Fashion MNIST

Motive of this question is to use metrics to detect and evaluate the bias in the machine learning model.

- **AIM:** Perform a 2-class classification between **Pullover** and **Coat**
- **Training data split:** Take all the training samples corresponding to Pullover, and only 500 samples for Coat.
- **Testing data split:** Take all the testing samples corresponding to Pullover and Coat.

- **Cross-validation:** Repeat each of the experiments 2 more times by taking different 500 samples of Coat.
- **Algorithms:** Perform classification using:
    - A linear SVM. (You may use sklearn library)
    - A 5 layer neural network with architecture: [128 -- 128 -- 128 -- 64 -- 1] (These numbers denote the number of nodes in each layer)
- Perform a 2 class classification and report the performance as follows:
    - Report :
        - Testing performance (mean ± std) [5 Marks]
        - Comparison of Testing Accuracy and show Confusion matrix for the classification performed: SVM vs Neural Network [5 Marks]
    - ROC curve (just 1 ROC for each algorithm with error bars, as shown in below graph) and report EER (Equal Error Rate). [MUST be done from scratch] [15 Marks]



    - Draw the Precision-Recall curve. [MUST be done from scratch] [5 Marks]
    - Of the two curves stated above, which is more reliable for biased/imbalanced data? Why? Draw inferences from the observed results on why one approach performs worse/better than the other. [20 Marks]

Q3. Use the CIFAR10 dataset, take 5 classes and perform the classification. Print confusion matrix for 5 classes. [Download Dataset from here] **[20 Marks]**
Note: Code for Confusion matrix must be done from scratch, else no marks will be awarded.

Q4. **Dataset**: Labeled Faces in the Wild, [Link for dataset] **[50 Marks + 10 Marks Bonus]**
You need to perform binary classification for a person wearing 'sunglasses'/'Eyeglasses' in the dataset. There are number of attributes present for LFW dataset(please refer the link for details about the attributes labeling).[Link for attributes]

- Separate the dataset into 2 files. 65% and 35% of the total dataset.
- For 65% dataset, perform binary classification for the person wearing 'sunglasses' /'Eyeglasses' or not. [**Note**: In the dataset, eyeglasses and sunglasses are treated as different entities, you need to consider them as single entity]
- Use Dense Layer Model : [128 -- 128 -- 128 -- 64 -- 1]

- Keep the loss function as the “mean square error”.
- Do you think there is any kind of bias in the system? Evaluate the system using 3 different evaluation metrics to see if there is any bias or not. [10 Marks]
- Come up with a new evaluation metric to detect if there is a bias in the system.
  [Bonus 10 Marks]
- If you observed any bias in the system, mitigate the bias by: [30 Marks]
    - DATA method (Training using more data): You may use more data for training from that 35% data. Report the accuracy after mitigation and compare it with previous classification results.
    - ALGORITHMIC method: Alter loss function to incorporate more challenges. Use a multi-tasking approach to achieve your aim.Report the accuracy after mitigation and compare it with previous classification results.