cc

Machine Learning
CS456- 2023

COMPUTER VISION: MNIST CLASSIFICATION

The MNIST dataset consists of a large collection of handwritten digits, specifically digits from 0 to 9. It is commonly used for tasks like image classification and digit recognition. The dataset is divided into two main parts: a training set and a test set.
The dataset in this link provides the MNIST data in CSV format. Each row in the CSV file represents an image, and the columns contain the pixel values for each pixel in the image.
The CSV files in this dataset include:

1. mnist_train.csv: This file contains the training set, consisting of a large number of labelled images. Each row represents an image, with the first column representing the label (the digit from 0 to 9) and the remaining columns representing the pixel values.
2. mnist_test.csv: This file contains the test set, which is used for evaluating the performance of machine learning models. It follows the same format as the training set, with the first column representing the label and the remaining columns representing the pixel values.

The pixel values in the CSV files are represented as grayscale intensity values ranging from 0 to 255, where 0 represents black and 255 represents white. Each image in the dataset has a fixed size of 28x28 pixels.

(1) Objectives:

- Acquire proficiency in using the Tensorflow framework and the Scikit-Learn library.
- Familiarize yourself with Artificial Neural Networks (ANNs) and K-Nearest Neighbors (K-NN).
- Attain a comprehension of the concept of hyperparameter tuning.

(2) Approaching the project:

(a) Data Exploration and preparation:

- Load the dataset and perform initial data exploration.
- Begin by familiarizing yourself with the dataset.
- Identify the number of unique classes.
- Identify the number of features.
- Check for missing values.
- Normalize each image by dividing each pixel by 255.
- Resize images to dimensions of 28 by 28. After resizing, visualize some images to verify the correctness of the reshaping process.
- Split the training data (mnist_train) into training and validation sets.

(b) Experiments and results:

- **Initial Experiment:** Implement the K-NN algorithm for classification and utilize a **grid search technique** to determine the optimal hyper parameters.
- **Subsequent Experiment:** Construct and train two different architectures of Artificial Neural Network (ANN) for classification, exploring variations in the number of hidden neurons, learning rate, and batch size.
- Compare the outcomes of the first and second experiments, discerning which approach yields the highest accuracy on the validation dataset.
- Get the confusion matrix of the best model.
- Save the best model, then reload it in a separate file, and use it on the testing data loaded from mnist_test.csv.

You can download dataset from:
https://www.kaggle.com/datasets/oddrationale/mnist-in-csv.

Instructions:
1. The number of students in a team is 5
2. No late submission is allowed
3. Cheating students will take ZERO and no excuses will be accepted
4. You can use python ML libraries e.g sklearn, keras, etc..

Deliverables:

- You are required to submit ONE zip file containing the following:
  - Your code (.py) file. If you have a (.ipynb) file, you have to save/download it as (.py) before submitting.
  - A report (.pdf) containing the team members' names and IDs, and the code with screenshots of the output of each part. If you have a (.ipynb) file, you can just convert it to pdf.
- The zip file must follow this naming convention: ID1_ID2_ID3_ID4_ID5_Group

Grading Critera:

Data Exploration and preparation	2
KNN with with grid search	3
ANN	3
Comparison	2
Confusion Matrix	1
Save, Load and Use Best Model	1
Total = 12 marks	