ĐỒ ÁN 2 - LOGISTIC REGESSION

1 Tập dữ liệu MNIST

1.1 Giới thiệu

- MNIST là tập dữ liệu về các chữ số viết tay.
- Tập dữ liệu được chia làm 2 phần: training set và test set.
- Training set gồm có 60000 mẫu. Test set gồm có 10000 mẫu.
- Mỗi mẫu có đặc điểm như sau:
  + Mỗi mẫu gồm có một bức ảnh và nhãn tương ứng đi kèm.
  + Bức ảnh trong mẫu là ảnh đen trắng (grayscale).
  + Bức ảnh có kích thước 28×28 điểm ảnh (pixel).
  + Mỗi điểm ảnh là một con số có giá trị 0 đến 255.
  + Chính giữa bức ảnh sẽ có hình ảnh về một chữ số nào đó trong các chữ số từ 0 đến 9.
  + Nhãn của bức ảnh sẽ có giá trị từ 0 đến 9, giá trị đó sẽ tương ứng với chữ số viết tay xuất hiện trong bức ảnh.

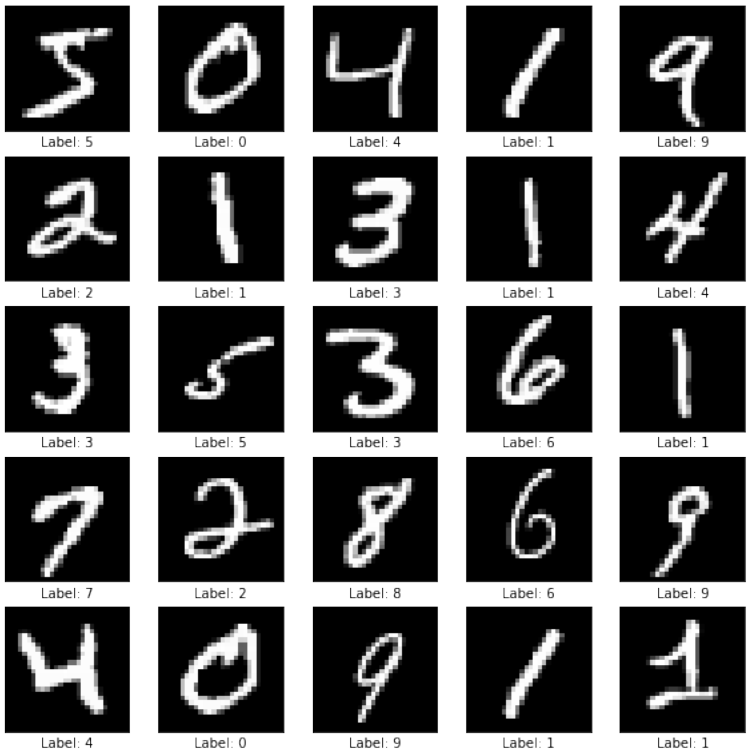
1.2 Tải tập dữ liệu

- Tải từ trang web gốc của tập dữ liệu:
  + Link: `http://yann.lecun.com/exdb/mnist/`
  + Bộ dữ liệu có tất cả 4 file, gồm các file ảnh và nhãn của tập training set và test set.
  + Các file được lưu bằng định dạng IDX.
  + Định dạng IDX tổ chức cấu trúc file như một mảng nhiều chiều.
  + Các file chứa ảnh có thể được xem là mảng 3 chiều, mỗi số có giá trị từ 0 đến 255.
  + Các file chứa nhãn có thể được xem là mảng 1 chiều, mỗi số có giá trị từ 0 đến 9.
- Tải thông qua thư viện sklearn:
  + Link: `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_openml.html`
  + Chọn bộ dữ liệu: `mnist_784`
  + Bộ dữ liệu này sẽ không chia thành training set và test set, mà sẽ không chung lại thành một bộ duy nhất gồm có 70000 mẫu.
  + Bộ dữ liệu được tổ chức thành một bảng gồm có 785 cột và 70000 dòng.
  + Các cột là `class`, `pixel1`, ..., `pixel784`, trong đó `class` là giá trị của nhãn, còn `pixel1`, ..., `pixel784` là các giá trị của các điểm ảnh trong bức ảnh.
- Tải thông qua thư viên tensorflow:
  + Link: `https://www.tensorflow.org/api_docs/python/tf/keras/datasets/mnist/load_data`
  + Bộ dữ liệu này sẽ không chia thành 4 phần là `x_train`, `y_train`, `x_test`, `y_test`.
  + `x_train` và `x_test` là mảng các mảng 2 chiều có kích thước 28×28 chứa các giá trị từ 0 đến 255 (mảng các bức ảnh).
  + `y_train` và `y_test` là mảng một chiều chứa các giá trị từ 0 đến 9 (mảng các nhãn).

1.3 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Tải xuống và đọc được toàn bộ tập dữ liệu MNIST bao gồm cả phần training set và test set.
- Lấy ngẫu nhiên một số mẫu bất kỳ trong tập dữ liệu, không phân biệt training set và test set. Hiển thị hình ảnh và nhãn của các mẫu đó giống như Hình 1.



Hình 1: Một số mẫu trong tập dữ liệu MNIST

2 Tiền xử lý dữ liệu

2.1 Về dữ liệu

- Đầu vào là một bức ảnh có kích thước 28×28 điểm ảnh. Như vậy, về bản chất, đó là mảng 2 chiều có kích thước 28×28 số nguyên có giá trị từ 0 đến 255.
- Đầu ra là một nhãn của bức ảnh. Như vậy, về bản chất, đó là số nguyên có giá trị từ 0 đến 9.

2.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Vì mô hình trong đồ án này là Logistic Regression nên ta sẽ không thể thực hiện được nếu giữ nguyên dữ liệu đầu ra như trên. Do đó, hãy nêu cách biến đổi dữ liệu đầu ra và giải thích cách làm.

- Khi quan sát một số mẫu, một số điểm ảnh có thể không có ý nghĩa trong việc phân loại. Chẳng hạn, các điểm ảnh ở biên luôn mang giá trị 0. Do đó, hãy nêu cách xác định các điểm ảnh vô nghĩa và giải thích cách làm. Khi đó, ta loại bỏ các điểm ảnh như vậy ra khỏi mô hình và cho biết số lượng điểm ảnh đã loại bỏ.
- Tìm hiểu về hiện tượng đa cộng tuyến. Nêu cách khắc phục và giải thích cách làm. Khi đó, ta loại bỏ các điểm ảnh gây ra hiện tượng đa cộng tuyến khỏi mô hình và cho biết số lượng điểm ảnh đã loại bỏ.
- Để gia tăng hiệu quả của mô hình, các nhóm được phép sử dụng thêm một số phương pháp khác như chuẩn hóa dữ liệu, giảm số chiều dữ liệu,... Nếu nhóm có thực hiện các phương pháp trên thì nhóm cần cho biết phương pháp mình áp dụng là gì và cho biết mức độ cải thiện cụ thể gia tăng bao nhiêu.

3 Logistic Regression

3.1 Mô hình

Đồ án này yêu cầu các nhóm sử dụng mô hình Logistic Regression để thực hiện nhiệm vụ phân loại và nhận dạng chữ số viết tay.

3.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Vì dữ liệu đầu vào và đầu ra đã bị biến đổi trong quá trình tiền xử lý nên nhóm phải nhắc lại dữ liệu đầu vào và đầu ra của mô hình là gì.
- Trình bày cấu trúc và cách thiết kế mô hình Logistic Regression một cách cụ thể, chi tiết từng bước tính toán từ đầu vào cho đến đầu ra.
- Trong mã nguồn, nếu nhóm sử dụng các tham số đặc biệt nào đó thì cần tìm hiểu và giải thích lý do tại sao chọn.
- Sau khi huấn luyện, cho biết độ chính xác của mô hình đối với toàn bộ tập dữ liệu, đối với từng lớp và lập ma trận Confusion Matrix.

4 Thử nghiệm thực tế

4.1 Mục đích

Mặc dù ta đã có tập dữ liệu test set để đánh giá mô hình, nhưng ta vẫn muốn thử khả năng của mô hình bằng cách viết chữ số trực tiếp.

4.2 Yêu cầu

Trong phần này, nhóm cần hoàn thành các yêu cầu sau:

- Viết chương trình cho phép sử dụng chuột để viết một chữ số bằng cách vẽ lên canvas.
- Chương trình có khả năng canh giữa chữ số vừa được vẽ ở trên và chuyển thành bức ảnh 28×28 điểm ảnh.
- Sau khi áp dụng các phương pháp tiền xử lý giống như đã làm với tập dữ liệu, chương trình chạy mô hình và trả ra kết quả dự đoán.

5 Các yêu cầu khác

- Ngôn ngữ sử dụng bắt buộc là Python, không được phép sử dụng ngôn ngữ khác.
- Không giới hạn thư viện được sử dụng trong Python.
- Các nhóm cần kiểm tra mã nguồn trước khi nộp. Nếu mã nguồn không chạy được mà không phải do nguyên nhân khách quan (thiếu thư viện, lỗi do thư viện gây ra, sử dụng thư viện sai phiên bản,...) thì sẽ bị 0 điểm đồ án.
- Bài nộp phải gồm có 3 phần:
  + Report: Chứa các file báo cáo.
  + Source: Chứa các file mã nguồn.
  + Presentation: Chứa các file dùng để thuyết trình.
- Trong các file nộp, nhóm cần ghi rõ thông tin về các thành viên gồm họ tên và MSSV. Riêng đối với mã nguồn, nhóm có thể ghi thông tin trên dưới dạng comment trong code của nhóm.
- Bài nộp sẽ được đặt trong thư mục có tên `MSSV01[_MSSV02[_MSSV03[...]]]` và được nén lại bằng định dạng ZIP với cùng tên như trên. Ví dụ đặt tên nhóm có 1 sinh viên là `MSSV01`, nhóm có 2 sinh viên là `MSSV01_MSSV02`.
- Nghiêm cấm các hành vi gian lận, không trung thực trong học tập như sao chép bài làm giữa các nhóm với nhau, sao chép bài làm của các nhóm khóa trước hoặc các nhóm lớp khác trường khác, nhờ người làm hộ. Nếu phát hiện các hành vi trên thì cả nhóm sẽ bị 0 điểm và xử lý theo quy định của Khoa và Trường.