Báo cáo đồ án cuối kỳ Nhập môn CNTT

Đề 2: Nhận diện chữ viết tay

20120015 - Lý Hoàng Khải
20120027 - Lê Hải Duy
20120463 - Nguyễn Lê Duy
20120548 - Lương Thanh Hoàng Phú
20120572 - Nguyễn Kiều Minh Tâm

Mục lục

1 Tổng quan đề tài

Mục tiêu: Viết chương trình Python để nhận diện chữ viết tay trên tập dữ liệu MNIST[1]

2 Xử lý dữ liệu

2.1 Load dữ liệu

Trước tiên, để dễ xử lý ta lần lượt đổi tên 2 file **t10k-images-idx3-ubyte.gz** và **t10k-labels-idx1-ubyte.gz** thành **test-images-idx3-ubyte.gz** và **test-labels-idx1-ubyte.gz**

Để load dữ liệu ta dùng hàm `load_mnist()`.

Hàm nhận 2 tham số.

- `path`: Đường dẫn đến tập dữ liệu
- `kind`: Nhận giá trị là `'train'` hoặc `'test'` tương ứng với việc load tập train hay tập test

Hàm trả về hai mảng tương ứng là tập ảnh và tập nhãn với tập dữ liệu tương ứng

2.2 Hiển thị dữ liệu

Để vẽ tập dữ liệu tương ứng ra màn hình ta dùng hàm `display()`.

Hàm nhận 2 tham số là tập `X` và tập `y` lần lượt là 2 mảng biểu diễn tập ảnh và tập nhãn

Hàm sẽ tạo một cửa sổ mới và vẽ ra cửa sổ đó 10 ảnh ứng với các số từ 0 đến 9

3 Rút trích đặc trưng

3.1 Vector hóa

Hàm `myVectorize()` nhận vào một mảng 2 chiều `arr` và trả về vector hóa tương ứng với mảng đấy

3.2 Histogram

Hàm `hist()` nhận vào một mảng 2 chiều `arr` và trả về một mảng `f` gồm 256 phần tử. Trong đó phần tử `f[x]` cho biết giá trị x xuất hiện bao nhiêu lần trong mảng `arr`

3.3 Sampling

Hàm `sampling()`. Hàm này nhận vào 3 tham số:

- `arr`: Một mảng đầu vào
- `sz`: Kích thước của một ô vuông mà ta dùng để chia nhỏ mảng `arr`
- `func`: Hàm số để ta thực hiện trên từng ô vuông kích thước `sz` (Ví dụ: min, max, trung bình cộng, ...

Hàm trả về sampling tương ứng với các tham số trên

4 Phân lớp

4.1 kNN (kth-Nearest Neighbours)

Hàm `getAccuracyKNN()`. Hàm nhận vào 4 tham số:

- `Xtrain`: Tập ảnh của tập dữ liệu train đã qua rút trích đặc trưng
- `ytrain`: Tập nhãn của tập dữ liệu train
- `Xtest`: Tập ảnh của tập dữ liệu test đã qua rút trích đặc trưng
- `ytest`: Tập nhãn của tập dữ liệu test

Hàm này sẽ chạy thuật toán `kth-nearest neighbours` [2][3] từ những tham số được đưa vào. Ở đây, với mỗi ảnh trong bộ test, ta sẽ xét 250 'hàng xóm' gần nhất ($k = 250$).

Sau khi thực hiện xong, hàm sẽ trả về một số thực trong đoạn $[0, 1]$ để chỉ độ chính xác của thuật toán

4.2 Mẫu trung bình

Gồm 2 hàm chính.

Hàm `createFeature()` nhận vào 2 tham số `X` và `y` lần lượt là tập ảnh đã qua rút trích đặc trưng và tập nhãn tương ứng. Hàm này sẽ trả về 10 mảng lần lượt là phân lớp của các nhãn từ 0 đến 9

Hàm `createFeature()` cũng được gọi trong hàm `getAccuracyAvg()`

Hàm `getAccuracyAvg()` nhận các tham số tương tư như hàm `getAccuracyKNN()`. Với mỗi ảnh trong bộ dữ liệu test, hàm này sẽ gán nhãn cho ảnh hiện tại bằng cách tìm nhãn có khoảng cách Euclid gần với ảnh đó nhất. Hàm này cũng trả về một số thực trong đoạn $[0, 1]$ là độ chính xác của thuật toán

Ngoài ra để thuận tiện. Ta còn cài đặt hàm `euclidDist()` nhận vào 2 tập `X`, `Y` và tính khoảng cách Euclid giữa 2 tập này

5 Kết quả chạy thử

Ở đây, chương trình sẽ tiến hành chạy các bộ phân lớp trên tập dữ liệu MNIST. Với các phương pháp rút trích đặc trưng và bộ phân lớp khác nhau.

Ta thu được bảng kết quả sau:

Phương pháp rút trích đặc trưng	Bộ phân lớp	Độ chính xác	Thời gian chạy (Giây)
Vector hóa	kNN	0.92	34.19
Vector hóa	Mẫu trung bình	0.82	432.19
Sampling kích thước 2, hàm min	kNN	0.82	17.58
Sampling kích thước 2, hàm min	Mẫu trung bình	0.74	110
Sampling kích thước 4, hàm min	kNN	0.28	17.70
Sampling kích thước 4, hàm min	Mẫu trung bình	0.22	31.46
Sampling kích thước 2, hàm trung bình cộng	kNN	0.93	19.10
Sampling kích thước 2, hàm trung bình cộng	Mẫu trung bình	0.82	26.54
Sampling kích thước 4, hàm trung bình cộng	kNN	0.91	15.12
Sampling kích thước 4, hàm trung bình cộng	Mẫu trung bình	0.78	6.54
Histogram	kNN	0.33	17.22
Histogram	Mẫu trung bình	0.26	143.75

Tài liệu

[1] The MNIST Database of handwritten digits
`http://yann.lecun.com/exdb/mnist/`

[2] Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (Report). USAF School of Aviation Medicine, Randolph Field, Texas.

[3] Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.