MACHINE LEARNING

PHÂN LOẠI ẢNH CHỮ SỐ VIẾT TAY

Khoa Toán - Cơ - Tin học
Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia Hà Nội

Ngày 12 tháng 5 năm 2023

Sinh viên thực hiện

VŨ NHẬT TÂN - 20000583
NGUYỄN XUÂN TRƯỜNG - 20000591
LƯU VĂN VIỆT - 20000598

Nội dung

Giới thiệu đề tài

Giới thiệu đề tài

Giới thiệu đề tài

Giới thiệu đề tài

Các bài toán phân loại (hay phân lớp) là một trong những lớp bài toán quan trọng nhất của Machine Learning.

Giới thiệu đề tài

Giới thiệu đề tài

Các bài toán phân loại (hay phân lớp) là một trong những lớp bài toán quan trọng nhất của Machine Learning.
Một bài toán phân loại nổi tiếng thuộc lĩnh vực xử lý ảnh là nhận dạng chữ số viết tay, trong đó mỗi chữ số được gán một trong 10 nhãn tương ứng với các số từ 0 đến 9.

Giới thiệu đề tài

Mục tiêu đề tài

Giới thiệu đề tài

Mục tiêu đề tài

Trong bài báo cáo này, chúng em sẽ xem xét 1 bộ dữ liệu ảnh các chữ số viết tay và thực hiện:

- Giảm số chiều dữ liệu bằng phương pháp PCA và trực quan hóa bộ dữ liệu.
- Thực hiện phương pháp phân loại Naive Bayes để huấn luyện mô hình phân loại ảnh chữ số viết tay, sau đó chạy kiểm thử và đánh giá mô hình.

Giới thiệu về bộ dữ liệu

Giới thiệu về bộ dữ liệu

Tập dữ liệu MNIST có nguồn gốc từ tập NIST do tổ chức National Institute of Standards and Technology (NIST) cung cấp.
Đây là tập dữ liệu thường dùng để đánh giá hiệu quả của các mô hình nhận dạng chữ số viết tay.
Dữ liệu gồm có 4 tệp tin nén:

Giới thiệu về bộ dữ liệu

Tập dữ liệu MNIST có nguồn gốc từ tập NIST do tổ chức National Institute of Standards and Technology (NIST) cung cấp.
Đây là tập dữ liệu thường dùng để đánh giá hiệu quả của các mô hình nhận dạng chữ số viết tay.
Dữ liệu gồm có 4 tệp tin nén:

- **train-images-idx3-ubyte**: Chứa 60000 mẫu dữ liệu ảnh train là ảnh các chữ số viết tay.
- **train-labels-idx1-ubyte**: Chứa 60000 mẫu dữ liệu nhãn ứng với tập ảnh train. Nhãn của từng ảnh là các số nguyên từ 0 đến 9 ứng với chữ số được ghi trong ảnh.
- **t10k-images-idx3-ubyte**: Chứa 10000 mẫu dữ liệu ảnh test.
- **t10k-labels-idx1-ubyte**: Chứa 10000 mẫu dữ liệu nhãn ứng với tập ảnh test.

Giới thiệu về bộ dữ liệu

Cấu trúc của file **train-images-idx3-ubyte** chứa dữ liệu ảnh training

Giới thiệu về bộ dữ liệu

Cấu trúc của file **train-images-idx3-ubyte** chứa dữ liệu ảnh training

[offset]	[type]	[value]	[description]
0000	32 bit integer	0x00000803(2051)	magic number
0004	32 bit integer	60000	number of images
0008	32 bit integer	28	number of rows
0012	32 bit integer	28	number of columns
0016	unsigned byte	??	pixel
0017	unsigned byte	??	pixel
.....			
xxxx	unsigned byte	??	pixel

Giới thiệu về bộ dữ liệu

Cấu trúc file **train-labels-idx1-ubyte** chứa nhãn của dữ liệu ảnh training:

Giới thiệu về bộ dữ liệu

Cấu trúc file **train-labels-idx1-ubyte** chứa nhãn của dữ liệu ảnh training:

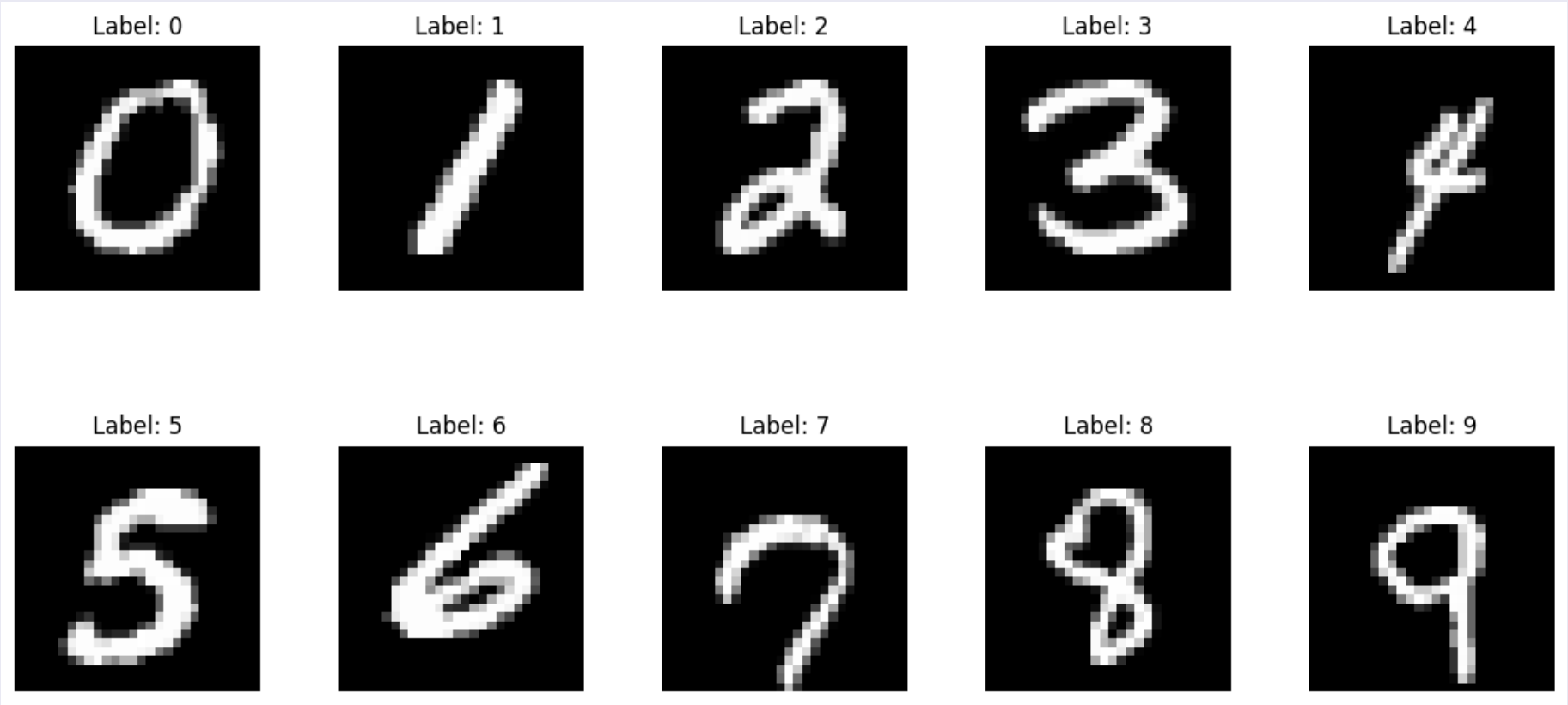
```
[offset] [type]          [value]          [description]
0000     32 bit integer  0x00000801(2049) magic number (MSB first)
0004     32 bit integer  60000            number of items
0008     unsigned byte   ??               label
0009     unsigned byte   ??               label
........
xxxx     unsigned byte   ??               label

The labels values are 0 to 9.
```

Giới thiệu về bộ dữ liệu

Một số dữ liệu ảnh

Giới thiệu về bộ dữ liệu

Một số dữ liệu ảnh



Phương pháp PCA

Giảm số chiều dữ liệu

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Ý tưởng đơn giản khi giảm số chiều dữ liệu từ d về k là ta sẽ chỉ giữ lại k chiều quan trọng nhất.

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Ý tưởng đơn giản khi giảm số chiều dữ liệu từ d về k là ta sẽ chỉ giữ lại k chiều quan trọng nhất.

Phương pháp phân tích thành phần chính - PCA thực hiện ý tưởng này bằng việc tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong k thành phần đầu tiên.

Phương pháp PCA

Giảm số chiều dữ liệu

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Ý tưởng đơn giản khi giảm số chiều dữ liệu từ d về k là ta sẽ chỉ giữ lại k chiều quan trọng nhất.

Phương pháp PCA

Giảm số chiều dữ liệu

Giảm số chiều dữ liệu có thể hiểu là việc tìm một hàm số đầu vào nhận tham số là một điểm $x \in \mathbb{R}^d$ với d rất lớn, và tạo ra kết quả là một điểm dữ liệu mới $z \in \mathbb{R}^k$ với $k < d$ sao cho thông tin về x không bị mất mát quá nhiều.

Ý tưởng đơn giản khi giảm số chiều dữ liệu từ d về k là ta sẽ chỉ giữ lại k chiều quan trọng nhất.

Phương pháp phân tích thành phần chính - PCA thực hiện ý tưởng này bằng việc tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong k thành phần đầu tiên.

Phương pháp PCA

Các bước thực hiện PCA

Phương pháp PCA

Các bước thực hiện PCA

1. Tính vector kỳ vọng của toàn bộ dữ liệu:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_n$$

2. Tính dữ liệu chuẩn hóa:

$$\hat{x}_n = x_n - \bar{x}$$

3. Tính ma trận hiệp phương sai:

$$S = \frac{1}{N}\hat{X}\hat{X}^T$$

4. Tính các trị riêng λ_i và vector riêng u_i có $||u_i||_2 = 1$ của ma trận này, sắp xếp chúng theo thứ tự giảm dần của trị riêng.

Phương pháp PCA

Các bước thực hiện PCA

Phương pháp PCA

Các bước thực hiện PCA

6. Chọn k giá trị riêng lớn nhất và k vector riêng trực chuẩn tương ứng. Xây dựng ma trận U_k .
7. Chiếu dữ liệu ban đầu đã chuẩn hoá $\hat{X}$ xuống không gian con nói trên. Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới:

$$Z = U_k^T \hat{X}$$

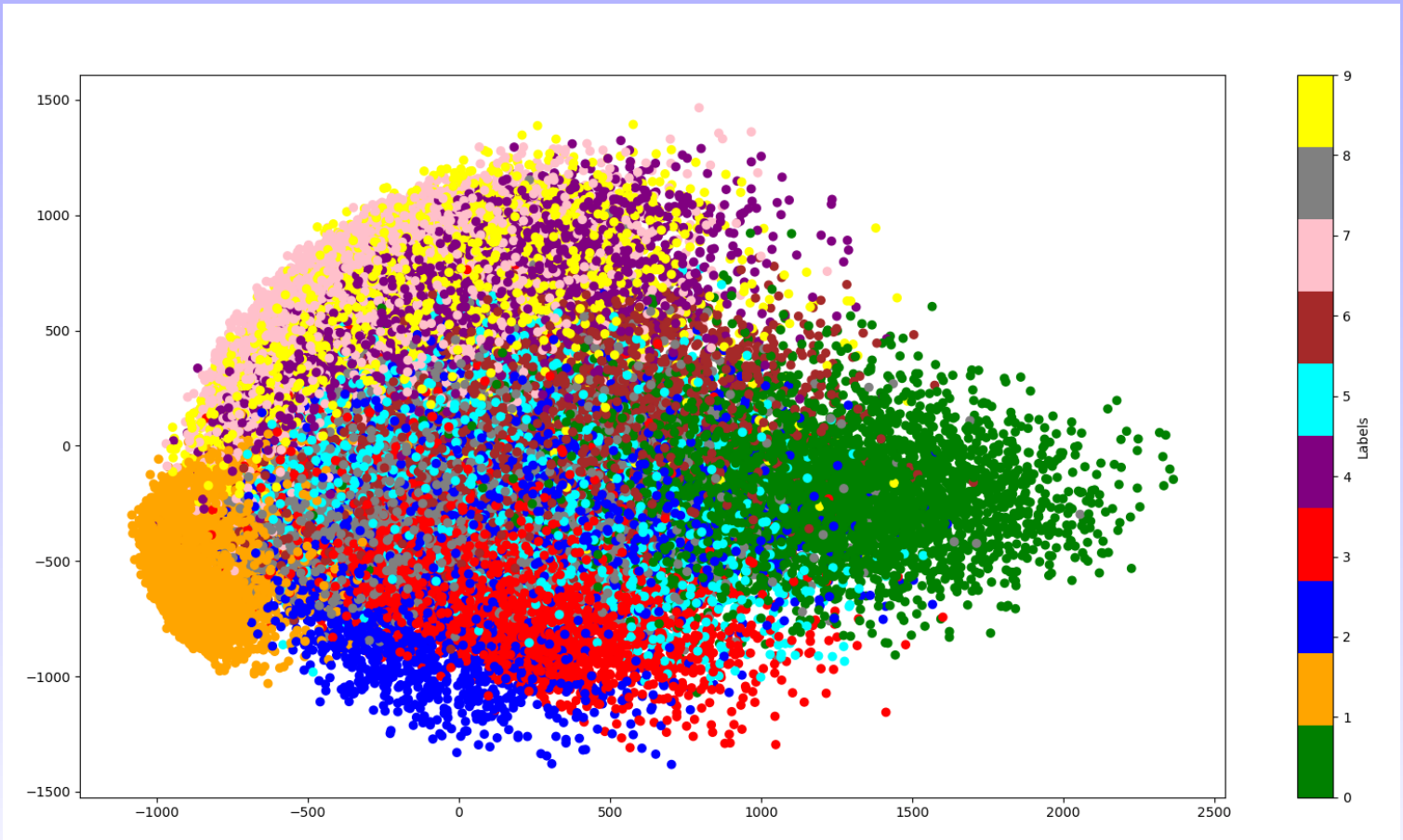
Dữ liệu ban đầu có thể được tính xấp xỉ theo dữ liệu mới qua công thức:

$$x \approx U_k Z + \bar{x}$$

Trực quan hóa dữ liệu

Trực quan hóa dữ liệu

Áp dụng phương pháp PCA giảm số chiều của bộ dữ liệu xuống còn 2 chiều và vẽ biểu đồ tán xạ của các điểm dữ liệu, ta thu được kết quả như sau:

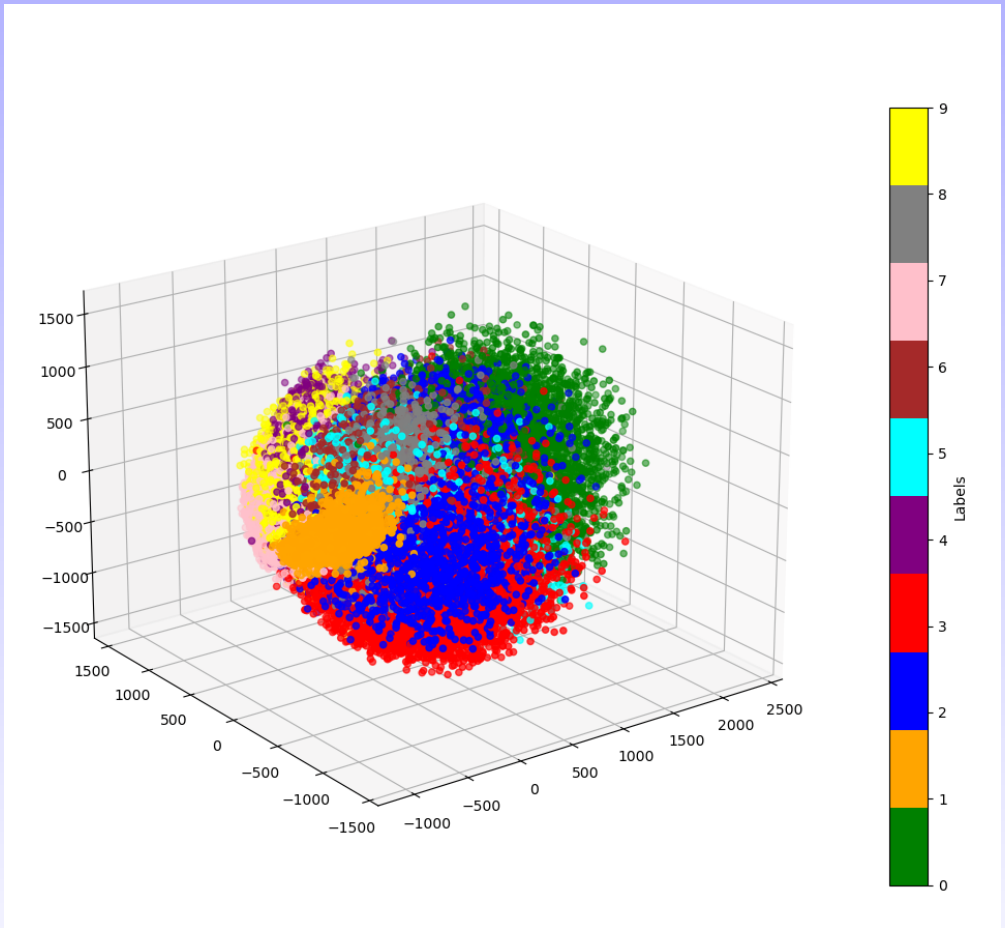


Hình 1: Dữ liệu khi giảm số chiều về 2

Trực quan hóa dữ liệu

Trực quan hóa dữ liệu

Tương tự áp dụng phương pháp PCA giảm số chiều của bộ dữ liệu xuống còn 3 chiều và vẽ biểu đồ tán xạ của các điểm dữ liệu, ta thu được kết quả như sau:



Hình 2: Dữ liệu khi giảm số chiều về 3

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Xét bài toán phân loại với C lớp khác nhau: $\{1, 2, 3, .., C\}$. Với mỗi điểm dữ liệu $x \in \mathbb{R}^d$, thay vì tìm ra nhãn chính xác của x, ta có thể tính xác suất để x thuộc về lớp $c \in \{1, 2, 3, .., C\}$:

$$p(y = c|x) \text{ hoặc viết gọn thành } p(c|x)$$

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Xét bài toán phân loại với C lớp khác nhau: $\{1, 2, 3, .., C\}$. Với mỗi điểm dữ liệu $x \in \mathbb{R}^d$, thay vì tìm ra nhãn chính xác của x, ta có thể tính xác suất để x thuộc về lớp $c \in \{1, 2, 3, .., C\}$:

$$p(y = c|x) \text{ hoặc viết gọn thành } p(c|x)$$

Biểu thức $p(c|x)$ có nghĩa là xác suất để đầu ra là lớp c nếu đầu vào là điểm dữ liệu x. Từ đó nếu như ta tính được $p(c|x)$ thì ta có thể xác định được lớp c của điểm dữ liệu x bằng việc chọn ra lớp mà có xác suất rơi vào cao nhất:

$$c = \underset{c \in \{1,2,..,C\}}{argmax} \; p(c|x)$$

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Áp dụng quy tắc Bayes và tính độc lập giữa dữ liệu quan sát x và phân lớp c thì bài toán tối ưu có thể được phát biểu lại như sau: Xác định c sao cho:

$$c = \underset{c \in \{1,2,..,C\}}{argmax} \; p(x|c)p(c)$$

Ý tưởng phương pháp phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Áp dụng quy tắc Bayes và tính độc lập giữa dữ liệu quan sát x và phân lớp c thì bài toán tối ưu có thể được phát biểu lại như sau: Xác định c sao cho:

$$c = \underset{c \in \{1,2,..,C\}}{argmax}\ p(x|c)p(c)$$

Giả thiết các thành phần của x là độc lập với nhau, với c đã cho. Khi đó:

$$p(x|c) = p(x_1, x_2, ..., x_d|c) = \prod_{i=1}^{d} p(x_i|c)$$

Ý tưởng phương pháp phân loại Naive Bayes

Các mô hình phân loại Naive Bayes

Ý tưởng phương pháp phân loại Naive Bayes

Các mô hình phân loại Naive Bayes

1. Gaussian Naive Bayes: Áp dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.

Ý tưởng phương pháp phân loại Naive Bayes

Các mô hình phân loại Naive Bayes

1. Gaussian Naive Bayes: Áp dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.
2. Multinomial Naive Bayes: Áp dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến rời rạc. Được sử phổ biến để phân loại văn bản mà vector đặc trưng được xây dựng dựa trên kỹ thuật Bags of Word.

Ý tưởng phương pháp phân loại Naive Bayes

Các mô hình phân loại Naive Bayes

1. Gaussian Naive Bayes: Áp dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến liên tục.
2. Multinomial Naive Bayes: Áp dụng chủ yếu trong loại dữ liệu mà các thành phần là các biến rời rạc. Được sử phổ biến để phân loại văn bản mà vector đặc trưng được xây dựng dựa trên kỹ thuật Bags of Word.
3. Bernoulli Naive Bayes: Được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bằng 0 hoặc 1.

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

Áp dụng mô hình Naive Bayes vào thực nghiệm

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

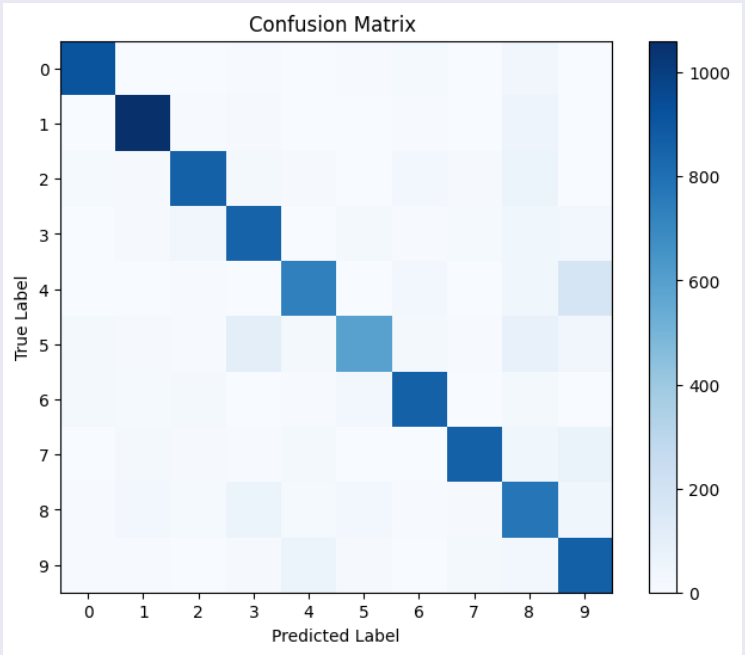
Áp dụng mô hình Naive Bayes vào thực nghiệm

Ta sử dụng mô hình Multinomial Naive Bayes để giải quyết bài toán phân loại chữ số viết tay. Huẩn luyện mô hình bằng phương pháp Multinomial Naive Bayes và chạy kiểm tra với tập test, chúng ta thu được kết quả đánh giá mô hình như sau:

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

Áp dụng mô hình Naive Bayes vào thực nghiệm

Ta sử dụng mô hình Multinomial Naive Bayes để giải quyết bài toán phân loại chữ số viết tay. Huẩn luyện mô hình bằng phương pháp Multinomial Naive Bayes và chạy kiểm tra với tập test, chúng ta thu được kết quả đánh giá mô hình như sau:



Hình 1: Confusion Matrix

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

Áp dụng mô hình Naive Bayes vào thực nghiệm

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

Áp dụng mô hình Naive Bayes vào thực nghiệm

Ta sử dụng mô hình Multinomial Naive Bayes để giải quyết bài toán phân loại chữ số viết tay. Huẩn luyện mô hình bằng phương pháp Multinomial Naive Bayes và chạy kiểm tra với tập test, chúng ta thu được kết quả đánh giá mô hình như sau:

Áp dụng mô hình Naive Bayes để phân loại ảnh chữ số viết tay

Áp dụng mô hình Naive Bayes vào thực nghiệm

Ta sử dụng mô hình Multinomial Naive Bayes để giải quyết bài toán phân loại chữ số viết tay. Huẩn luyện mô hình bằng phương pháp Multinomial Naive Bayes và chạy kiểm tra với tập test, chúng ta thu được kết quả đánh giá mô hình như sau:

Metric	Value
Accuracy	0.8365
Precision	0.843316299712613 2
Recall	0.8334531845906966

Hình 2: Một số metric đánh giá mô hình

Độ chính xác cũng như Precision và Recall của mô hình đạt được khá cao với $83,65\%; 84,33\%; 83,35\%$ tương ứng.

Tổng kết

Các kết quả của bài báo cáo:

- Mô tả về bộ dữ liệu ảnh chữ số viết tay. Xây dựng được phương pháp PCA và áp dụng để giảm số chiều dữ liệu sau đó trực quan hóa bộ dữ liệu.
- Xây dựng được các mô hình phân loại Naive Bayes, sau đó áp dụng mô hình phân loại Naive Bayes phù hợp với bộ dữ liệu để huấn luyện mô hình. Chạy kiểm thử mô hình trên tập dữ liệu Test sau đó tính toán một số metric đánh giá mức độ hiệu quả của mô hình. Có thể nói mô hình phân loại Multinomial Naive Bayes hoạt động khá tốt với giá trị của 3 metric đánh giá đều đạt trên 80%.

Tài liệu tham khảo

[1] Vũ Hữu Tiệp, *Machine Learning cơ bản*, https://machinelearningcoban.com/ebook/

[2] https://machinelearningcoban.com/2017/06/15/pca/

[3] http://yann.lecun.com/exdb/mnist/

THANK YOU FOR YOUR ATTENTION!