Việc phân loại ảnh là một trong những ứng dụng quan trọng của học máy và học sâu. Bộ dữ liệu MNIST, bao gồm các hình ảnh chữ số viết tay từ 0 đến 9, là một tập hợp phổ biến để thử nghiệm các thuật toán phân loại ảnh. Trong bài tiểu luận này, chúng tôi trình bày một quy trình sử dụng mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) để phân loại các hình ảnh sau khi đã giảm số chiều bằng Phân tích Thành phần Chính (Principal Component Analysis - PCA). Chúng tôi sử dụng bộ dữ liệu MNIST để huấn luyện và kiểm tra mô hình ANN, đồng thời đánh giá hiệu suất bằng các chỉ số như độ chính xác, ma trận nhầm lẫn, recall, precision và F1-score.

**Phương pháp và dữ liệu**

Bộ dữ liệu MNIST bao gồm 60.000 hình ảnh cho tập huấn luyện và 10.000 hình ảnh cho tập kiểm tra. Mỗi hình ảnh là một ma trận 28x28, tương ứng với 784 giá trị pixel. Chúng tôi thực hiện các bước sau:

1. **Chuẩn hóa dữ liệu**: Chúng tôi chuẩn hóa các giá trị pixel bằng cách sử dụng **StandardScaler** để đảm bảo dữ liệu có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1.
2. **Giảm số chiều bằng PCA**: Sau khi chuẩn hóa, chúng tôi áp dụng PCA để giảm số chiều của dữ liệu từ 784 xuống còn 100. PCA là một kỹ thuật được sử dụng để giảm số chiều bằng cách chọn những thành phần chính (principal components) có phương sai lớn nhất.

Giảm số chiều về 100 bằng phương pháp phân tích thành phần chính (PCA) trước khi huấn luyện mô hình ANN (Artificial Neural Network) là một chiến lược phổ biến trong học máy và khoa học dữ liệu, nhất là khi làm việc với dữ liệu hình ảnh. Sau đây là các lý do chính tại sao việc giảm số chiều trước khi huấn luyện ANN có thể hữu ích:

1. Giảm độ phức tạp và tránh overfitting:

Dữ liệu hình ảnh thường có rất nhiều chiều (chẳng hạn, 28x28 pixels cho dữ liệu MNIST, tương đương 784 chiều). Với số lượng chiều lớn như vậy, mô hình ANN có thể trở nên quá phức tạp và dễ bị overfitting, nghĩa là mô hình học quá mức chi tiết của dữ liệu huấn luyện và hoạt động kém trên dữ liệu chưa thấy trước đó (validation).

PCA giúp giảm số chiều và giữ lại các thành phần quan trọng nhất của dữ liệu. Điều này làm cho mô hình ANN trở nên đơn giản hơn và giảm nguy cơ overfitting.

2. Tăng tốc độ huấn luyện và tính toán:

ANN là loại mô hình phức tạp với nhiều tham số. Khi số chiều dữ liệu đầu vào lớn, số lượng tham số của mô hình tăng lên, dẫn đến việc huấn luyện chậm và đòi hỏi nhiều tài nguyên tính toán.

Giảm số chiều giúp giảm khối lượng tính toán, tăng tốc độ huấn luyện và đòi hỏi ít tài nguyên hơn.

3. Giảm nhiễu (noise) và cải thiện hiệu suất:

Dữ liệu với nhiều chiều có thể chứa nhiều nhiễu (noise). PCA giúp loại bỏ các chiều ít thông tin, tập trung vào các chiều mang nhiều thông tin nhất.

Bằng cách tập trung vào các thành phần chính, mô hình ANN có khả năng hoạt động hiệu quả hơn trên dữ liệu với ít nhiễu hơn.

1. **Huấn luyện mô hình ANN**: Chúng tôi sử dụng MLPClassifier từ thư viện scikit-learn để xây dựng mô hình ANN với một lớp ẩn gồm 100 nơ-ron. Mô hình được huấn luyện với tập dữ liệu đã giảm số chiều.
2. **Kiểm tra và đánh giá hiệu suất**: Sau khi huấn luyện, chúng tôi sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình. Các chỉ số được sử dụng để đánh giá bao gồm độ chính xác (accuracy), ma trận nhầm lẫn (confusion matrix), recall, precision và F1-score.

**Kết quả và thảo luận**

Sau khi huấn luyện mô hình ANN, chúng tôi đã kiểm tra hiệu suất trên tập kiểm tra và thu được các kết quả sau:

* **Độ chính xác**: Mô hình đạt được độ chính xác 96.72% trên tập kiểm tra, chứng tỏ khả năng phân loại tốt của ANN.
* **Ma trận nhầm lẫn**: Ma trận nhầm lẫn cho thấy các lỗi phổ biến giữa các lớp. Ví dụ, số 3 và 5 thường bị nhầm lẫn với nhau, điều này có thể là do đặc điểm của các chữ số viết tay.
* **Recall, precision, F1-score**: Các chỉ số này đều đạt mức 97.60%, cho thấy sự cân bằng tốt giữa việc xác định đúng các lớp và việc tránh nhầm lẫn.

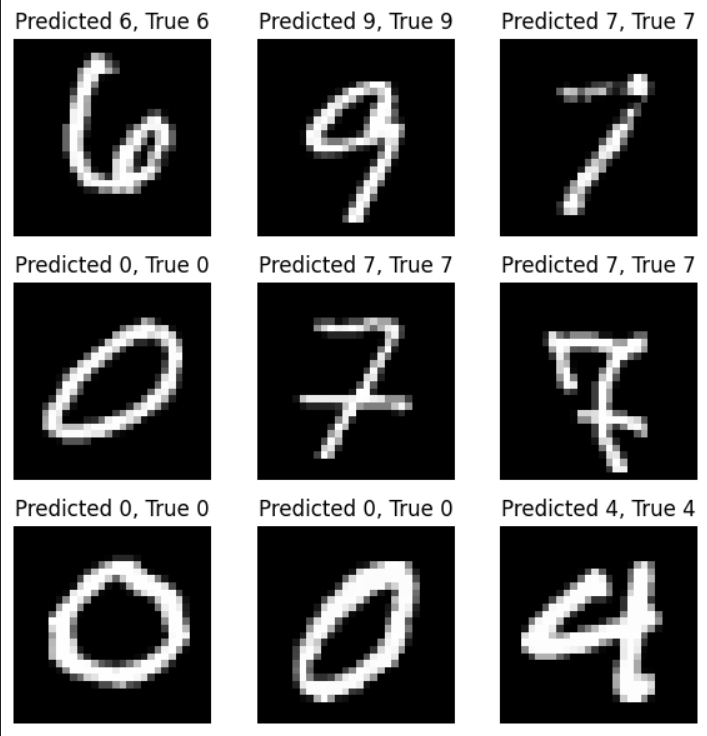
Nhìn chung, mô hình ANN đã thể hiện hiệu suất cao trong việc phân loại ảnh MNIST sau khi giảm số chiều bằng PCA. Việc giảm số chiều giúp tăng tốc quá trình huấn luyện và giảm nguy cơ overfitting, trong khi vẫn duy trì độ chính xác cao.

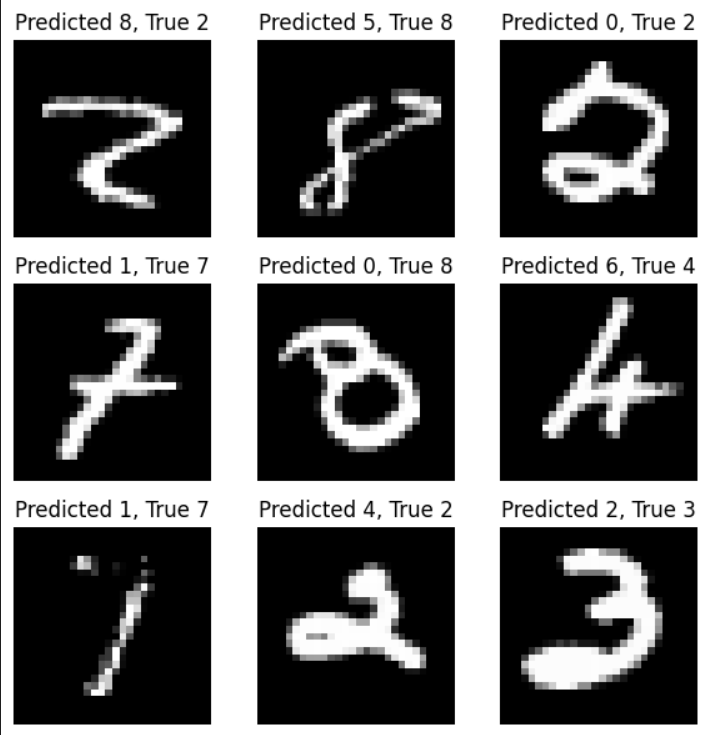
**Kết luận**

Việc sử dụng PCA để giảm số chiều kết hợp với ANN là một phương pháp hiệu quả để phân loại ảnh. Trong bối cảnh bộ dữ liệu MNIST, mô hình ANN đã đạt được độ chính xác và các chỉ số hiệu suất cao. Các kết quả này cho thấy tiềm năng của việc sử dụng các kỹ thuật học sâu và giảm số chiều trong các bài toán phân loại ảnh phức tạp hơn.

Một số trường hợp phân loại đúng và sai của mô hình ANN:

* Phân loại đúng:





Việc phân loại sai là do chữ viết tay ko rõ rang

MLPClassifier là một trong những thuật toán học máy quan trọng thuộc họ Mạng Nơ-ron Nhân tạo (Artificial Neural Networks - ANN) MLPClassifier là một mạng nơ-ron dạng feedforward (mạng truyền thẳng), có thể bao gồm một hoặc nhiều lớp ẩn (hidden layers). Mỗi lớp ẩn chứa một số lượng nơ-ron nhất định, và mỗi nơ-ron trong một lớp được kết nối với tất cả các nơ-ron trong lớp tiếp theo, một kiến trúc được gọi là "fully connected" hoặc "dense". MLPClassifier hoạt động dựa trên nguyên tắc lan truyền tiến (forward propagation) và lan truyền ngược (backward propagation). Trong lan truyền tiến, dữ liệu đầu vào được truyền qua các lớp, áp dụng các hàm kích hoạt (activation functions) để tạo ra đầu ra cuối cùng. Để huấn luyện mạng, lan truyền ngược được sử dụng để điều chỉnh các trọng số thông qua một quá trình gọi là "backpropagation", với mục tiêu giảm thiểu hàm mất mát (loss function). Quá trình này thường sử dụng các thuật toán tối ưu như Stochastic Gradient Descent (SGD) hoặc Adam.Top of Form