MNIST Dataset Report

Marina Tarasova

EC Utbildning

Projekt i Data Science

202411

Abstract

This project explores the classification of handwritten digits from the MNIST dataset using machine learning and deep learning techniques. Two machine learning models (Logistic Regression and Random Forest) were implemented and evaluated using Cross-Validation, Classification Reports and Confusion Matrices. A deep learning approach was implemented using a Sequential Neural Network with EarlyStopping and L2 regularization to reduce overfitting. This report compares the performance of these techniques, as well as highlighting their strengths and limitations.

Innehållsförteckning

1 Inledning..... 3

2 Maskininlärningsmetod..... 4

2.1 Workflow..... 4

2.2 Resultat..... 4

2.3 Confusion Matrix..... 6

3 Djupinlärningsmetod..... 7

3.1 Workflow..... 7

3.2 Resultat..... 7

4 Självutvärdering..... 8

1 Inledning

MNIST-datasetet, en samling med 70.000 handskrivna sifferbilder, används ofta för maskininlärnings- och djupinlärningsalgoritmer.

Syftet med denna rapport är att klassificera siffror med hjälp av maskininlärningsmodeller och en djupinlärningsmetod, för att uppfylla syftet så kommer följande projekt att jämföras:

1. Maskininlärning med Logistisk Regression och Random Forest.
2. Djupinlärning med Sequential Neural Network.

2 Maskininlärningsmetod

2.1 Workflow

1. Standardisering av data med hjälp av StandardScaler.
2. Implementering av två modeller:
    - Logistisk Regression.
    - Random Forest.
3. Utvärdering av modellerna med hjälp av Cross-Validation och prestandamätningar.
4. Visualisering av klassificeringsresultaten med hjälp av Confusion Matrix.

2.2 Resultat

Logistisk Regression

- Cross-Validation Accuracy: 91%
- Classification Report:

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	1343
1	0.95	0.97	0.96	1600
2	0.90	0.89	0.90	1380
3	0.90	0.89	0.90	1433
4	0.92	0.92	0.92	1295
5	0.88	0.88	0.88	1273
6	0.93	0.94	0.94	1396
7	0.92	0.94	0.93	1503
8	0.90	0.86	0.88	1357
9	0.89	0.90	0.90	1420
accuracy			0.92	14000
macro avg	0.92	0.92	0.92	14000
weighted avg	0.92	0.92	0.92	14000

Logistisk Regression uppnådde goda resultat för siffror som “0” och “1”, men hade problem med mer komplexa siffror som “5” och “9”.

Random Forest

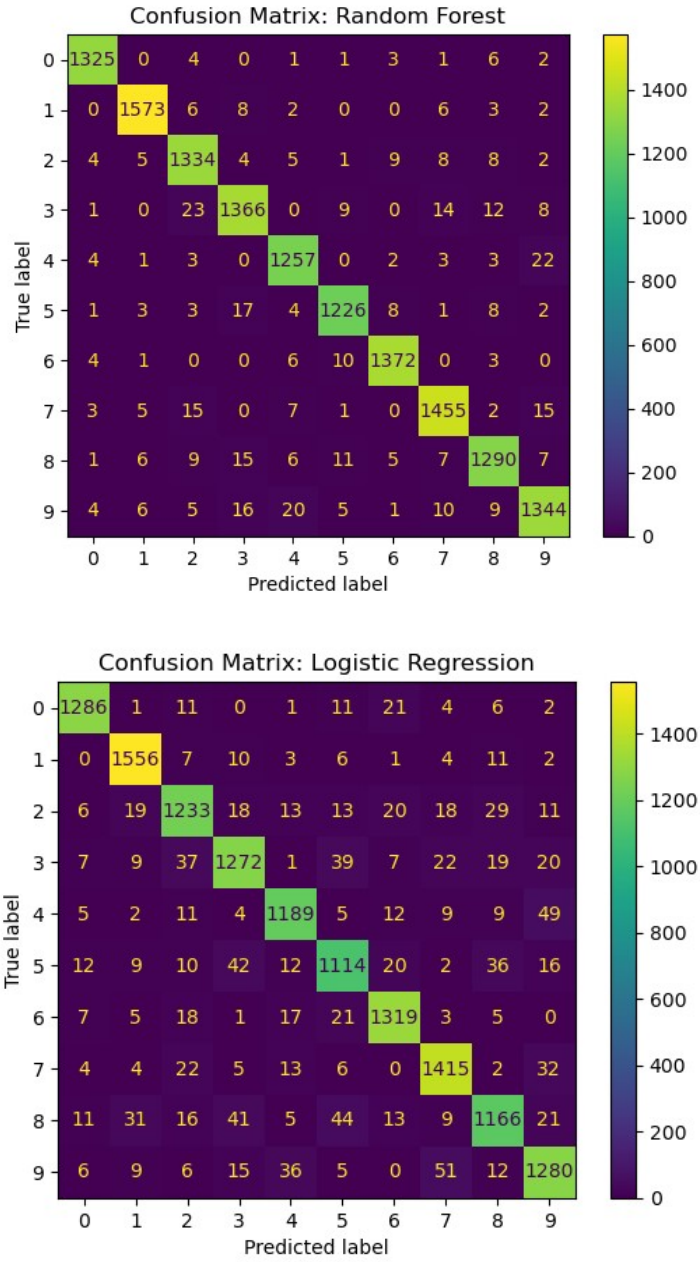
- Cross-Validation Accuracy: 96-97%
- Classification Report:

Random Forest Report:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	1343
1	0.98	0.98	0.98	1600
2	0.95	0.97	0.96	1380
3	0.96	0.95	0.96	1433
4	0.96	0.97	0.97	1295
5	0.97	0.96	0.97	1273
6	0.98	0.98	0.98	1396
7	0.97	0.97	0.97	1503
8	0.96	0.95	0.96	1357
9	0.96	0.95	0.95	1420
accuracy			0.97	14000
macro avg	0.97	0.97	0.97	14000
weighted avg	0.97	0.97	0.97	14000

Random Forest presterade betydligt bättre än Logistisk Regression, särskilt när det gällde komplexa sifferklassificeringar.

2.3 Confusion Matrix



Random Forest hade färre felklassificeringar än Logistisk Regression, men hade fortfarande problem med siffror som “8” och ”3”.

3 Djupinlärningsmetod

3.1 Workflow

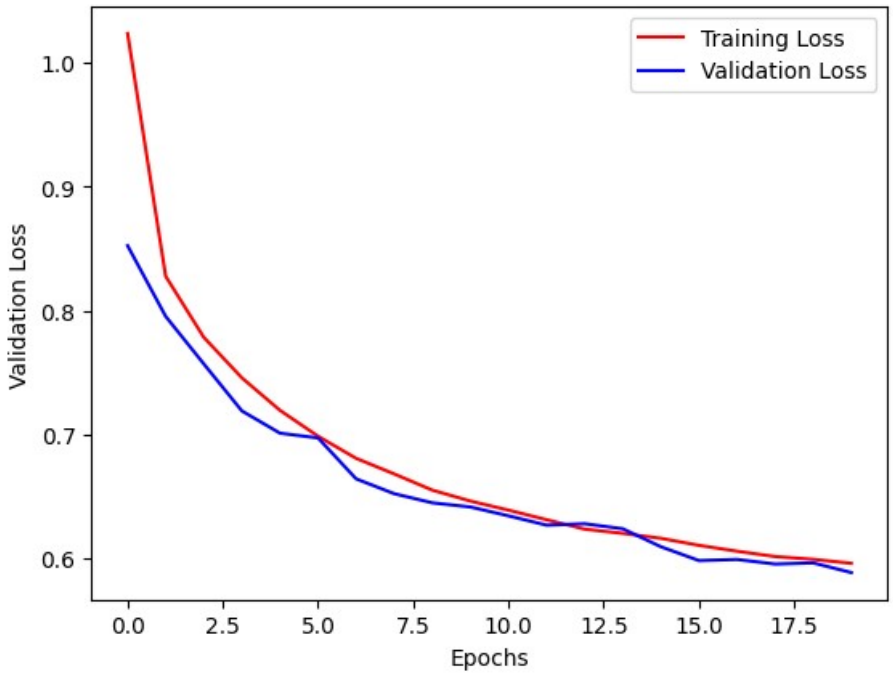
1. Omformning av datasetet.
2. Bygga ett Sequential Neural Network.
3. Förhindra överanpassning med regulariseringsmetoder (EarlyStopping och L2).
4. Träning och utvärdering av modellen.

3.2 Resultat

- Training Accuracy: 94-95%
- Validation Accuracy: 94-95%
- Test Accuracy: 94-95%

Djupinlärningsmodellen uppnådde ganska hög noggrannhet och regulariseringsmetoder mildrade effektivt överanpassning.

Även om regulariseringsmetoderna förbättrade generaliseringen minskade de också noggrannheten.



4 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

   En av utmaningarna var att lära mig begreppen maskininlärning och djupinlärning på egen hand, men jag hade mycket bra resurser för det.

   En annan utmaning var att träna djupinlärningsmodellen och förhindra överanpassning.

2. Vilket betyg du anser att du skall ha och varför.

   Jag tycker att jag borde få godkänd eftersom jag förstår begreppen maskin- och djupinlärning rätt bra.

3. Något du vill lyfta fram till Antonio?

   Tack för de goda föreläsningarna på Youtube!