Kunskapskontroll 2

Machine Learning

Boris Lovrenovic

EC Utbildning

Kunskapskontroll 2 - Machine Learning

202503

Abstract

This report investigates classification of handwritten numbers of images from the MNIST dataset, with the help of Logistic Regression and Random forest models.

The models were trained on 60,000 images and then tests were made on a testset consisting of 10,000 images where the pixels were normalized to the interval [0,1].

Logistic Regression showed an accuracy of roughly 92% while the Random Forest model showed 97%.

Random forest performed better thanks to its ability to manage complex patterns in the data.

The work conducted shows how two simple models can be used effectively for image classification.

Table of contents

1 Introduction

Machine Learning is an area in Data Science which can be used to teach computers how to learn from data to do different predictions and classifications. An example of this is to identify handwritten numbers, which is what I've done in my work here with the MNIST dataset.
It contains more than 70,000 images of numbers from 0 to 9. In this work I've decided to focus on training, test and compare two models, Logistic Regression and Random Forest, to classify the numbers from the dataset using the models and to see which models performs the best.

1.1 Purpose and question statement

Purpose is to classify handwritten numbers from the MNIST dataset with the help of two different machine learning models, Logistic regression and Random forest, as well as to compare their results.

To fulfill this purpose, I will answer the question;
Which of the models, Logistic regression or Random forest gives the most accurate result?

2 Theory

2.1 Logistic Regression

Logistic regression is a simple model used for classification. It uses a linear function to predict probability for a class.

2.2 Random Forest

Random forest is a model that combines a lot of decision trees. Each decision tree votes on a class, which eventually leads to and gives a final result. This makes the model stronger as compared to the logistic regression model.

3 Method

The work was done in Jupyter Notebooks using Python with the Scikit-Learn module.

First the dataset was loaded using fetch_openml.

The data was then prepared by normalizing the pixels of the images to [0,1], by dividing it by 255, X = mnist["data"] / 255.0.

I trained the logistic regression model with max_iter=1000 and Random forest with n_estimators=100.

I then ran tests for each model on the test set and measured the accuracy.

4 Result and discussion

Accuracy	
Logistic Regression	0.9259
Random Forest	0.9704

Logistic regression got a 92% accuracy result, which is decent for a linear model, while Random forest got 97% thanks to its ability to manage complex patterns through several decision trees.

Logistic regression had a harder time to differentiate between numbers with similar shapes.

5 Conclusion

Random forest has better accuracy than Logistic regression för the MNIST dataset.
This shows that an ensemble model is better suited for an image classification compared to a linear model.

The work shows how simple methods can be used and compared with, with a pretty decent accuracy, in a simple and clear way.

6 Teoretiska frågor

Träning används för att lära upp modellen, genom att låta den lära sig mönster från datan.
Validering används för att förbättra modellen under träning.
Test används för att testa den färdiga modellen för att se hur exakt/precis den är.

Då Julia inte har har ett set för validering så kan hon använda korsvalidering. Vilket betyder att splitta datan i mindre delar och träna varje modell i en mindre del och sedan testa den på de andra delarna.

Linear Regression, kan användas för att förutspå huspriser.
Lasso Regression, liknande linear regression men man kan ignorera mindre viktiga variabler.
Random forest, använder många decision trees - kan användas för att rekommendera produkter till kunder eller hantera lager (ecommerce exempelvis).

RMSE visar hur långt ifrån en modells prediction är från dess riktiga värden.
Lägre RMSE betyder att det är en bättre modell, och används för att jämföra modeller eller kolla hur väl en modell fungerar.

Klassificeringsproblem = Problem där mål kolumnen är en kategorisk kolumn.
Logistic Regression, kan användas för att förutspå hur sannolikt det är att en kund genomför ett köp.
Decision Trees, kan användas att exempelvis förutspå kundförlust i marknadsföring, baserat på tidigare köp och engagemang.
Confusion Matrix är en tabell som visar hur många predictions som var rätt eller fel för varje kategori.

K-means är en modell som grupperar data i “K-clusters” baserat på hur lika data points man har för vad det nu man använder modellen till. Exempelvis gruppera kunder i deras köpesbeteende för riktad marknadsföring mot en målgrupp.

Ordinal encoding - Nummer per kategori i ordning. A=1, B=2, C=3
One-hot encoding - Skapar ny kolumn för varje kategori med 1or och 0or. A=[1,0,0], B=[0,1,0], C=[0,0,1]
Dummy variable encoding - Liknande one hot encoding men droppar en kategori. A=[1,0], B=[0,1], C[0,0]

Normalt sätt är färger nominal(ingen ordning). Men om man säger att om man bär en röd skjorta till en fest så är man vackrast på festen, så ger man det en ordning, vilket gör det till ordinal data.
Så Julia har rätt i det här fallet.

Streamlit är ett open source framework för att skapa data applikationer i Python för ML och Data science teams.
Det används för att bygga datadrivna webbapplikationer som exempelvis interaktiva dashboards, visualiseringar och rapporter. Vilket gör det enkelt att mäta och presentera olika datadrivna insikter.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   Inga jättestora utmaningar, men en jag minns var när jag tränade LR modellen så fick jag ett error pga jag satt max iterations till 100. Löste problemet med att öka till 1000.

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

3. Något du vill lyfta fram till Antonio?

   Nej.

References

Géron, A. SECOND EDITION - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems