Maskininlärning – Modellering av MNIST-dataset

Erik Nordén
EC Utbildning
Kunskapskontroll 2 – Maskininlärning
2025-05

Abstract

This project explores two machine learning models applied to the MNIST dataset: a simple neural network and a Random Forest classifier. The neural network was implemented using Keras and achieved over 97% accuracy. The Random Forest classifier was trained on flattened pixel data and performed slightly below the neural network. The project includes code for model training, evaluation and saving. Results indicate the neural network is better suited for image recognition tasks.

Innehållsförteckning

Innehåll

1. Inledning

Artificiell intelligens har fått en allt större betydelse i samhället och används idag i allt från rekommendationssystem till självkörande bilar. Ett viktigt delområde inom AI är maskininlärning, där algoritmer tränas för att lära sig mönster från data. Denna rapport fokuserar på klassificeringsproblem, där målet är att förutsäga vilken siffra som visas i en bild.

Syftet med denna rapport är att jämföra två olika maskininlärningsmodeller för klassificering av handskrivna siffror. Följande frågeställningar besvaras:
1. Hur väl presterar ett enkelt neuralt nätverk jämfört med en Random Forest-modell?
2. Hur påverkar datans förbehandling resultaten?

Rapporten är disponerad enligt följande: först presenteras relevant teori och svar på teoretiska frågor, därefter metod, resultat och slutligen slutsatser.

2. Teoretiska frågor

1. Vad används träning, validering och test till?
Träning används för att lära modellen. Validering används under träningen för att se hur bra modellen funkar på ny data och hjälpa oss justera inställningar. Test används sist för att kolla hur bra modellen är på helt ny data som den aldrig sett förut.

2. Hur väljer Julia den bästa modellen utan ett valideringsdataset?
Hon kan använda något som heter korsvalidering (cross-validation). Det betyder att hon delar upp sin data i flera små delar och testar modellen flera gånger på olika delar.

3. Vad är ett regressionsproblem?
Ett regressionsproblem är när man försöker räkna ut ett tal, till exempel vad en lägenhet kostar. Exempel på modeller är linjär regression och random forest regression. Det används när vi vill förutspå ett värde, inte en kategori.

4. Vad är RMSE?
RMSE betyder Root Mean Squared Error. Det visar hur långt ifrån sanningen modellen gissar i genomsnitt. Ju lägre RMSE, desto bättre.

5. Vad är ett klassificeringsproblem?
Det är när man vill att modellen ska gissa vilken kategori något tillhör, som t.ex. om en bild är en hund eller katt. Vi använder t.ex. logistisk regression eller random forest. En confusion matrix visar vad modellen gissat rätt och fel på.

6. Vad är K-means?
K-means är en metod som grupperar saker i kluster. Den gör det utan att vi berättar i förväg vilka grupper det finns. Till exempel kan den hitta olika sorters kunder i en butik baserat på hur de handlar.

7. Vad är ordinal, one-hot och dummy encoding?
Ordinal encoding ger varje kategori ett nummer i ordning, som liten=1, mellan=2, stor=3. One-hot encoding gör en kolumn för varje kategori och sätter 1 på den som gäller. Dummy encoding är som one-hot, men man tar bort en kolumn för att undvika överflödig info.

8. Är färger ordinal eller nominal?
Det beror på. Vanligtvis är färger nominal (ingen inbördes ordning). Men om vi säger att färger betyder något i en viss ordning så kan det bli ordinal.

9. Vad är Streamlit och vad används det till?
Streamlit är ett verktyg i Python som man kan använda för att göra enkla webbsidor och appar, till exempel för att visa resultat från en maskininlärningsmodell.

3. Teori

De modeller som används i denna studie är ett enkelt neuralt nätverk och en Random Forest-modell.

Neurala nätverk består av lager av noder (neuroner) som försöker efterlikna hur hjärnan arbetar. De är särskilt bra på att identifiera mönster i bilder.

Random Forest är en ensemblemetod som kombinerar flera beslutsträd för att förbättra stabilitet och noggrannhet.

För att utvärdera modeller används testnoggrannhet, vilket är andelen korrekt klassificerade bilder i testdatan.

4. Metod

Datan som används är MNIST, ett dataset med 70 000 handskrivna siffror (0–9). Datan hämtas automatiskt via TensorFlow. Förbehandling sker genom att bilderna normaliseras så att pixelvärden ligger mellan 0 och 1.

Två modeller tränas:
- Ett neuralt nätverk med två lager
- En Random Forest-modell

Modellerna tränas på träningsdatan (60 000 bilder) och utvärderas på testdatan (10 000 bilder). Modellerna sparas till disk efter träning.

5. Resultat och Diskussion

Neurala nätverket nådde en testnoggrannhet på cirka 97.3%, medan Random Forest nådde ca 97.0%. Skillnaden kan förklaras av att neurala nätverk bättre kan hantera bilders tvådimensionella struktur.

Random Forest är dock enklare att förstå och kräver ingen GPU för att köras snabbt. Resultatet visar tydligt att val av modell påverkar prestandan och bör anpassas efter uppgiftens karaktär.

6. Slutsatser

Båda modeller presterar bra, men det neurala nätverket var överlägset på klassificeringsuppgiften. För framtida arbete vore det intressant att testa konvolutionella neurala nätverk (CNN) för ännu bättre resultat.

Förhandsbehandlingen, såsom normalisering och reshaping, är viktig för båda modellerna och påverkar resultatet markant.

7. Självutvärdering

Jag upplevde vissa svårigheter i att förstå hur datan skulle förberedas för olika modeller. Genom att läsa dokumentationen och öva med kodexempel kunde jag lösa det mesta. Jag anser att jag uppfyller kriterierna för betyget Godkänt. Jag har genomfört uppgiften enligt instruktion och lärt mig mycket om både neurala nätverk och Random Forest.

Jag vill lyfta att det var särskilt roligt att se hur olika modeller kan tolka bilder på så olika sätt.