Farhana Moshammat
Kurs: Machine learning 30yhp
Lärare: Antonio Prgomet

MNIST RAPPORT

Introduktion

Bakgrund

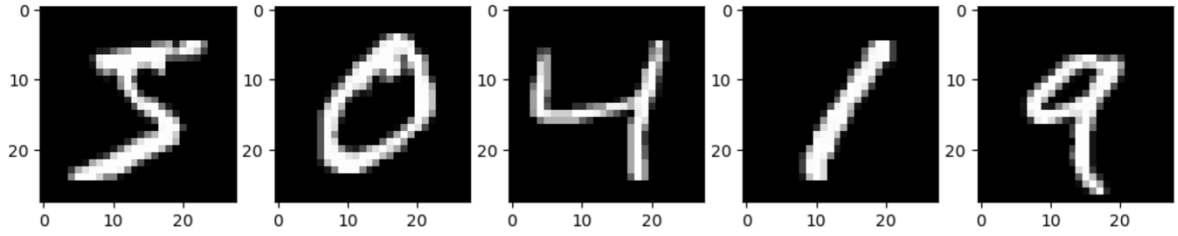
Maskininlärning är en gren inom artificiell intelligens som fokuserar på att utveckla datoralgoritmer och modeller som kan lära sig från data och skapa prediktioner utan explicit programmering. Alltså tränar maskiner att analysera, identifiera mönster och fatta beslut grundat utifrån den givna information och data. MNIST dataset är en väl använt och känd dataset i pedagogisk syfte inom maskininlärning och denna dataset kommer rapporten huvudsakligen ligga fokus på.

Frågeställning

I denna rapport ska frågeställning besvaras:
”Kan en accuracy score på minst 95 % överträffas på MNIST dataset? ”

Databeskrivning – EDA

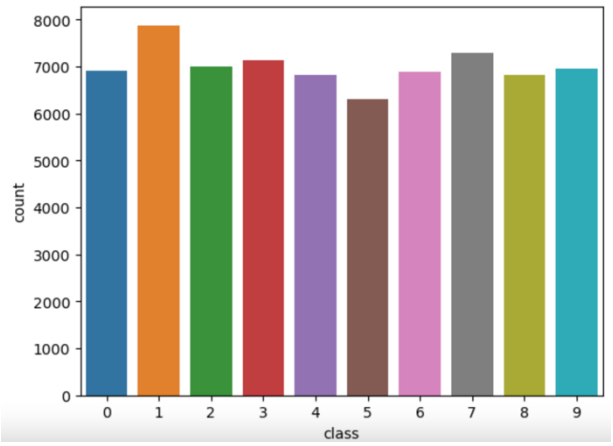
Rapporten kommer utgår från MNIST dataset som består av en stor samling på handskrivna siffror från 0 till 9, alltså har den 10 klasser. Innehåller 60 000 träningsdata och 10 000 testdata. En siffra är skriven av olika individer. MNIST dataset innehåller även information som indikerar den korrekta siffran för varje bild. Varje bild är 28x28 pixlar, vilket innebär att dataset har 784 features och därmed 784 kolumner. Kolumnerna heter pixel1 till pixel 784. MNIST dataset tillhör bildklassificering och fungerar för att utvärdera prestanda för olika modeller och algoritmer inom mönsterigenkänning och datorseende.



Figur 1 visar ett urval av hur de gråskalade handskrivna siffror kan se ut.

```
0    6903
1    7877
2    6990
3    7141
4    6824
5    6313
6    6876
7    7293
8    6825
9    6958
Name: class, dtype: int64
```

Figur 2 frekvens för varje siffra, från 0-9. Siffran 1 har flest siffra och siffran 5 har lägst.



Figur 3 visar stapeldiagram för de tio klasserna från 0 till 9. Fördelningen är jämn, alltså ingen siffran skiljer drastisk från övriga siffror.

Metod & Modeller – Teori

För att modellera MNIST datasetet och får en accuracy score på minst 95% kommer maskin learnings modeller att tränas för att känna igen handskrivna siffror. Det går att använda olika algoritmeter och olika modeller. Denna rapport har utnyttjat träningsdata och testdata. Träningsdata och testdata används i maskininlärning för att utvärdera prestanda och generaliseringsförmåga hos en tränad modell. Hela träningsdata på 60 000 handskrivna siffror har används för att lära ut modellen att göra korrekta prediktioner och därefter använts testdata på 10 000 för att utvärdera och dess prestanda samt säkerställa generalisering och upptäcka overfittning. Genom att dela upp test och träningsdata gör det möjligt att bygga tillförlitliga och effektiva maskin inlärningsmodeller.

Modeller som har tillämpats för att få en accuracy score på minst 95 % är SVM, logistic regression och Randomized search. Accuracy score används eftersom det är en vanlig och

enkel mått för att utvärdera prestandan för en klassificeringsmodell, även lått att förstå och tolka. Den ger svar på hur ofta modellens prediktioner matchar de faktiska observationer.

SVM (support vector machines) kan hanteras på både regressionsproblem och klassificieringsproblem. SVM passar utmärkt för att klassificera dessa bilder i respektive sifferkategori (0 till 9). SVM:er är särskilt effektiva när de hanterar data som i detta fall bilder. Iden med SVM klassificering är att försöka skapa en så bred väg som möjligt mellan observationer som i detta fall är handskrivna siffror i respektive klass. Modellen kommer automatisk att tillämpa one vesus the rest, OvR, och det innebär att den skapar en binär klassificerare för varje sifferkategori som kan svara på frågan ifall det är en nolla eller inte och därefter gör den samma sak för resterande siffra från 1 till 9.

Logistic regression är en binär klassificerare som uppskattar sannolikheten att en observation tillhör en klass och kan används på MNIST dataset för klassificering i flera klasser med hjälp av OvR algoritmen, vilket Scikit-learn gör automtisk. Det går alltså att använda logistisk regression för klassificering i flera klasser.

Randomized Search är en metod där slumpmässiga kombinationer av hyperparameter väljs och därefter används för att träna modell. Randomized Search är snabbt och detta på grund av den inte testar alla parameter utan sökningen görs slumpmässigt. I denna rapport har 10 slumpmässiga urval utförs i en randomiserad search.

Resultat & Analys

Resultat

Resultaten kan sammanfattas på tabellen nedan. Alla modeller ger högt accuracy score över 90%. Högst accuracy score ges av Randomized Search på 97% och det innebär att modellen prediktioner var 97 % korrekta av alla handskrivna siffor (observationer).

Modell	**Accuracy score**
SVM	**0,9121**
Logistic regression	**0,9255**
Randomized search	**0,9709**

Analys

En högt accuracy score indikerar en bättre prestanda eftersom modellen gör korrekta prediktioner för en stor andel av datamängden. Alla modeller har gett ett högt accuracy score på över 90 %, men enbart Randomized Search kan nå en accuracy score på över 95 %,

närmare bestämt 97%. Detta tyder på att Randomized search prestanda är bra att prediktera handskrivna siffror på MNIST dataset. Att Randomized Search har hittat och valt ut ”korretka” kombinationer av hyperparameter som valt ut slumpmässig och utvärderats för att hitta den optimala dataanpassningen och därmed resulterat till ett högt accuracy score.

Denna rapport har enbart använt sig av oss accuracy score, vilket kanske inte ger en komplett bild av modellens prestanda speciell i de fall där klasserna är obalanserade. Men i EDA avsnittet har vi sätt att siffrorna är ganska jämnfördelat, alltså det skiljer sig inte drastisk mellan de olika siffrorna. Men det kan ändå vara bra att överväga andra utvärderingsmått som precision, recall och confusion matrix. Dessutom hade man kunnat testa flera modeller som Random Forest. Detta skulle kunna vara potentiell vidareutveckling.