MNIST klassificering

Natalie Dobrovolska

EC Utbildning

Examensarbete- Byt namn

202403

Abstract

En kort sammanfattning över ditt arbete och de viktigaste resultaten skrivet på engelska, cirka 5 meningar totalt.

The objective of this thesis is to develop a model which can predict/classify with at least 90% accurace the handwritten digits from 0-9.

Three models – Stochastic Gradient Descent Classifier, Logistic Regression and Support Vector Machine were used. The best result in accuracy got Support Vector Machine with 94%.

Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.

Innehållsförteckning

1 Inledning

Maskininlärning är ett kraftfullt redskap som har breda användningsområden. Den använder algoritmer för att identifiera mönster i data som sedan används för att skapa datamodeller för att kunna göra förutsägelser.[1] Med hjälp av ML kan man lösa komplexa problem, effektivisera och automatisera processerna, prediktera mönster mm.

1.1 Syfte

I detta arbete har jag använt maskininlärning med syftet att kunna prediktera/klassificera handskrivna siffror 0-9. Målet är att kunna göra det med minst 90% precision.

För att uppfylla syftet kommer följande frågeställning besvaras:

1. Vilka av dessa modeller kommer kunna prediktera med högsta precision?
   - Stochastic Gradient Descent Classifier
   - Logistic Regression
   - Support Vector Machines

[1] https://azure.microsoft.com/sv-se/resources/cloud-computing-dictionary/what-is-machine-learning-platform

2 Teori

Nedan beskrivs de modellerna som användes i detta projekt.

2.1 Stochastic Gradient Descent Classifier

Stochastic Gradient Descent Classifier är en linjär klassificerings algoritm som hjälper att hitta den optimala beslutsgränsen för att separera datapunkter som tillhör olika klasser. Denna kan med fördel effektivt hantera mycket stora datamängder. (Géron, 2019, s.88, 100)

2.2 Logistic Regression

Logistisk regression används för att uppskatta sannolikheten att en instans tillhör en viss klass. Om den uppskattade sannolikheten är större än 50 %, då förutsäger modellen att instansen tillhör den klassen annars förutsäger den att den inte gör det. Denna algoritm klassas som binär klassificerare. (Géron, 2019, s.142)

2.3 Support Vector Machine

Support Vector Machine (SVM) är en kraftfull och mångsidig maskininlärning modell, kapabel att utföra linjär eller olinjär klassificering, regression och jämnt avvikande upptäckt. Det är en av de mest populära modellerna inom maskininlärning. SVM är särskilt väl lämpad för klassificering av komplexa små eller medelstora datamängder. (Géron, 2019, s.153)

3 Metod

Nedan beskrivs detaljerat hur projektet utfördes.

3.1 MNIST dataset

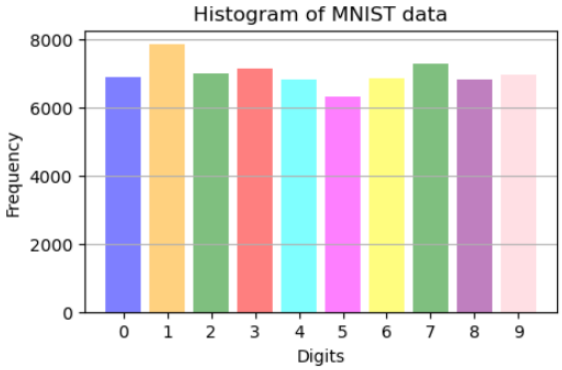
Figur 1. En siffra från datasetet

För att kunna genomföra detta arbete har MNIST data används.

MNIST databas står för Modified National Institute of Standards and Technology database.

Det en uppsättning av 70 000 små bilder av handskrivna siffror av gymnasieelever och anställda vid US Census Byrå. Varje bild är märkt med den siffra den representerar. Denna uppsättning har studerats så mycket att det ofta kallas maskininlärningens "Hello World". (Géron, 2019, s.85)

3.2 Utforskning av datan



Figur 2. Histogram av MNIST dataset

Efter att MNIST datan laddades har jag försökt att utforska den. För att få en bättre uppfattning har jag bl a tagit fram en av siffrorna för att se hur bilderna i datasetet kan se ut. (Figur 1)

Man kan se att fördelningen i antalet av varje siffra i datasetet inte är jämn. (Figur 2)

3.3 Modellering

Som nästa steg delades datasetet upp i tränings- och testdatan för att kunna börja modellera. Jag valde att dela upp datan i mindre delar för att det skulle gå mer tidseffektivt.

Därefter påbörjades modellering med Stochastic Gradient Descent Classifier, Logistisk regression och Support Vector Machine modeller. För varje modell hittades med hjälp av Grid Search de hyperparametrarna för den bästa prestandan och generaliseringsförmågan.

4 Resultat och Diskussion

Efter genomförandet av modellering fick jag fram dessa prestandaresultat.

Det var viktigt att leta efter de bästa parametrarna för varje modell. Det förbättrade prestandan avsevärt.

Accuracy score för de olika modellerna	
Stochastic Gradient Descent Classifier	0.89
Logistisk regression	0.91
Support Vector Machine	0.94

Tabell 3: Accuracy score för de olika modellerna.

5 Slutsatser

Den bäst presterande modellen blev Support Vector Machine som predikterade rätt med 94%.

Det går säkert att uppnå något bättre resultat, skulle vara kul att lyckas med det. Det skulle även vara spännande att testa andra modeller, t ex Random Forest.

6 Teoretiska frågor

1. Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test”, vad används respektive del för?

”Träning” används för att träna den/de valda modellerna med träningsdatan. Man brukar för detta ändamål använda 20% av den totala datan och den ska väljas slumpmässigt.

”Validering” används för att utvärdera modellerna som användes under ”Träning”. Här brukar man justera hyperparametrarna (parameter som används för att styra inlärningen) för att förbättra modellens prestanda. Under denna del väljer man den modell som presterar bäst.

Under ”Test” testas den valda bästa modellen på datan (som inte används under träning eller validering).

2. Julia delar upp sin data i träning och test. På träningsdatan så tränar hon tre modeller; ”Linjär Regression”, ”Lasso regression” och en ”Random Forest modell”. Hur skall hon välja vilken av de tre modellerna hon skall fortsätta använda när hon inte skapat ett explicit ”validerings-dataset”?

Då kan man använda sig av K-Fold Cross Validation. Det innebär att träningsdatan delas randomiserat upp i flera tränings-set bestående av 5 eller 10 delar (k-fold) som sedan utvärderar modeller genom att välja en k-fold för utvärdering och resterande k-folds för att träna de modellerna som valts att användas.

3. Vad är ”regressionsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden?

I regressionsproblem försöker man hitta relation mellan den beroende och oberoende variabler.

Målet med det är att kunna prediktera den beroende variabeln utifrån oberoende variabler.

Exempel på det är kunna prediktera priset på bostad (beroende variabel) utifrån information på området, antalet rum, våning (alla dessa är oberoende variabler).

Ett annat exempel är att kunna prediktera efterfrågan på en viss produkt/sortiment utifrån priset, erbjudandet.

Här är några modeller som kan användas:

- Linjär Regression
- Support Vector Machines (SVM)
- Beslutsträd

4. Hur kan du tolka RMSE och vad används det till: $\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

RMSE står för Root Mean Squere Error. Med hjälp av RMSE kan man tolka prediktionernas medelavstånd till de sanna värdena**.** Med andra ord är det medelavståndet mellan de förutsagda värden som vi fått från modellen till de verkliga värden som vi har i vår testdata.

Om man ska förklara formeln, så beräknar den medelvärdet mellan de predikterade och verkliga testdata värden. För att ”neutralisera” positiv eller negativ skillnad mellan dessa värden, kvadreras skillnaden mellan predikterade och de verkliga värden och sedan man tar roten ur det resultatet.

5. Vad är klassificeringsproblem? Kan du ge några exempel på modeller som används och potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

I klassificieringsproblem är målet att kategorisera observationer i klasser/kategorier utifrån observationens egenskaper.

Det finns binära (två klasser - ja/nej, Kvinna/Man) och flerklassklassificeringar (observationer med fler egenskaper än två).

Några modeller som kan användas vid klassificeringsproblem är Support Vector Machines, Random Forest, Regression.

Confusion Matrix är en tabell som används för att illustrera prestandamått på modellens gjorda kvalificering. Med hjälp av tabellen kan man också räkna ut Precision, Recall och F1 Score.

6. Vad är K-means modellen för något? Ge ett exempel på vad det kan tillämpas på.

K-Means är en Unsupervised Learning algoritm som används för klusteranalys. K-Means delar upp datan i kluster utifrån likheter och separerar från den datan som skiljer sig (som tillhör ett annat kluster). Den kan användas på stora datamängder.

Den kan tillämpas för att t.ex. identifiera kundgrupper, segmentera bilder, analys av användarbeteende inom digitala plattformar.

7. Förklara (gärna med ett exempel): Ordinal encoding, one-hot encoding, dummy variable encoding. Se mappen ”l8” på GitHub om du behöver repetition.

Det är sätt att hantera kategoriska variabler genom att tilldela dem siffror. Det gör man för att kunna använda dessa variabler, då maskininlärningsmodeller kräver att all data är numerisk.

Ordinal encodning innebär att man tilldelar varje unik kategorisk variabel en ordningsförhållande i form av en numerisk kod. Ett exempel på det är skolans betygsskala – ”IG” tilldelas 0, ”G” 1, ”VG” 2.

One-hot encoding används på kategoriska variabler som inte har någon ordningsförhållande. Man använder binära variabler för varje kategori som kod, oftast 0 och 1, som representerar förekomsten eller frånvaro av en visst kategorisk variabel. Ett exempel på det är att om en elev har fått en VG så tilldelas en 1 för denna variabel och till övriga variabler (IG och G) tilldelas en 0.

Dummy variable encoding liknar One-hot encoding förutom att man tar bort en ”dubblett” kodning på en kategori. Som i exemplet ovan skulle man slippa tilldela en 1 till en av kategorierna när de övriga kategorier har fått en 0, för man vet att om övriga har fått en 0, så är den som inte fått en 0 får en 1.

8. Göran påstår att datan antingen är ”ordinal” eller ”nominal”. Julia säger att detta måste tolkas. Hon ger ett exempel med att färger såsom {röd, grön, blå} generellt sett inte har någon inbördes ordning (nominal) men om du har en röd skjorta så är du vackrast på festen (ordinal) – vem har rätt?

Båda har rätt. Om man tänker på ordinal data, så är det den data som har en viss ordningsförhållande naturligt, t ex betyg från bästa till sämsta, storlek från största till minsta osv.

Nominal data till skillnad från ordinal har inte någon ordningsförhållande/rangordning från grunden. Men man kan tilldela även här ett ordningsförhållande, precis som Julia gör när hon säger att en röd skjorta gör en vackrast.

9. Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDa-RzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12

Och besvara följande fråga:

- Vad är Streamlit för något och vad kan det användas till?

Streamit är en open-source framework i Python som möjliggör snabbt skapandet och delning av appar för Machine inlärning. Med hjälp av den kan man bl a bygga och testa ML modeller, göra visualiseringar.

4. Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**
   Denna kurs var den svåraste för mig hittills. Det var utmanande med all information som man skulle ta till sig. Jag försökte läsa i boken ganska mycket samt googla, men kände ändå att det var väldigt tufft. Det var svårt att förstå det jag läste/gjorde.
   En annan sak är att jag har fortfarande lite utmaningar med Python, så det gjorde inte saken lättare.
   Planen är att gå igenom allt igen, så att jag känner mig mer trygg med ML.

2. **Vilket betyg du anser att du skall ha och varför.**
   Jag hoppas att jag kan få G. Ärligt talat så är jag väldigt besviken att jag inte kunde/hann fixa VG uppgiften. Mitt mål är att kunna görs det på fritiden så att jag känner att jag kan den biten.

3. **Något du vill lyfta fram till Antonio?**
   Det var väldigt intressant och viktig kurs.

Källförteckning

Machine Learning. (2024). https://azure.microsoft.com/sv-se/resources/cloud-computing-dictionary/what-is-machine-learning-platform/.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. (R. R. Tache,

Ed.) Canada: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Retrieved from hΣp://oreilly.com