Modellval för MNIST

Jämförelse av två klassificieringsmodeller på MNIST

ECUTBILDNING

Linniéa Truong

EC Utbildning

Machine Learning Rapport

202503

Abstract

This report examines which machine learning model performs best on the MNIST dataset, which contains images of handwritten digits. Two models were tested and compared: Logistic Regression and Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel. The models were evaluated based on their accuracy and training time. The results showed that SVM performed better at recognizing new images, which makes it a stronger choice for this task. However, Logistic Regression is still a good option, especially if you need a faster model or want better generalization across different training sets.

Förkortningar och Begrepp

RBF - Radial Basis Function
MNIST - Modified National Institute of Standards and Technology
CV - Cross Validation
dvs - det vill säga

Innehållsförteckning

1 Inledning

Maskininlärning är ett snabbt växande område inom datavetenskap och blir en alltmer integrerad del av vårt vardagsliv. Detta kan vi se från alla möjliga håll, alltifrån bilen bredvid oss på motorvägen som är helt självkörande, till rösten i våra mobiler som kan hjälpa oss att ringa en vän utan att vi själva behöver slå en signal. Men vad är egentligen maskininlärning?

Maskininlärning är ett område inom artificiell intelligens som samlar på sig en massa data vilket ger datorer möjligheten att träna och lära sig, utan behov av programmering. En av de vanligaste uppgifterna inom maskininlärning är klassificering, där algoritmer tränas för att känna igen och förutsäga olika kategorier baserat på data. Ett vanligt exempel är att identifiera handskrivna siffror – ett problem som länge använts som referens inom området.

I denna rapport har jag använt mig av datasetet MNIST som består av 70 000 bilder av handskrivna siffror (0–9). Med hjälp av logistisk regression och Support Vector Machine kunde jag jämföra och få en bättre förståelse för deras styrkor och svagheter för att slutligen avgöra vilken modell som passar bäst just för MINST-datasetet.

1.1 Syfte

Syftet med detta arbete är att undersöka och jämföra prestandan hos två olika maskininlärningsmodeller – logistisk regression och SVM med RBF-kernel – vid klassificering av handskrivna siffror i MINST-datasetet.

1. 2 Frågeställning

Vilken av modellerna, logistisk regression eller SVM med RBF-kernel, presterar bäst på MNIST-datasetet sett till noggrannhet och effektivitet?

2 Teori

Maskininlärning kan delas in i två huvudkategorier: väglett lärande (supervised learning) och icke- väglett lärande (unsupervised learning). Vid väglett lärande får programmet tillgång till både indata och utdata, vilket gör att den kan lära sig att efterlikna dessa exempel och göra liknande förutsägelser. I icke-väglett lärande får algoritmen däremot enbart indata och måste själv identifiera mönster eller strukturer i datan, utan att ha tillgång till några förutbestämda resultat. Men inom väglett lärande kan man dela in den i ytterligare två underkategorier: Regression och Klassificering.

Inom regression vill vi förutspå en kontinuerlig numerisk variabel för en given numerisk variabel. Det innebär att modellen försöker hitta ett samband mellan indata (features) och ett numeriskt utfall (målvariabel). Vanliga typer av regressioner är Linjär- och Polynomregression samt Ridge och Lasso.

Klassificiering däremot är en typ av maskininlärning som används för att förutsäga kategorier eller klasser. Baserat på indata så försöker modellen avgöra vilken grupp något tillhör. Detta kan man ta hjälp av bland annat Logistisk Regression, SVM och Decision Tree. Ett exempel på detta är att känna igen handskrivna siffror och bestämma vilket nummer det är.

2. 1 MNIST-dataset

MNIST (Modified National Institute of Standards and Technology) är ett standardiserat dataset som består utav handskrivna siffror för att och används oftast för att jämföra olika maskininlärningsmodeller. Det innehåller 60 000 träningsbilder och 10 000 testbilder varav varje bild är 28x28 pixlar och omfattar en handskriven siffra mellan 0-9.

2. 2 Logistisk regression

Logistisk regression är en linjär modell som fungerar på liknande sätt som linjär regression. Som linjär regression så beräknar den först en viktad summa, alltså att varje input eller funktion i en modell multipliceras med en vikt som styr hur mycket den ska påverka resultatet, för att sedan adderas ihop med en *bias-term.* Men till skillnad från linjär regression så skickar den inte ut summan direkt som resultat. Istället skickas den genom en sigmoidfunktion, som även kallas logistisk funktion.

Logistisk regression förutsäger sannolikheten att en datapunkt tillhör en viss klass. Modellen använder sigmoidfunktionen för att omvandla ett linjärt resultat till ett värde mellan 0 och 1, vilket tolkas som sannolikhet. Den matematiska formeln för sigmoidfunktionen är: $\sigma(t) = 1/(1 + exp(-t))$

Fördelar med logistisk regression är att den är enkel att tolka, snabb att träna och speciellt bra vid linjära problem. Nackdelar är att den har en begränsad kapacitet att hantera komplexa, icke-linjära mönster.

2. 3 SVM med RBF-kernel

Support Vector Machine är en kraftfull och flexibel modell inom maskininlärning som används främst för klassificering, men kan även användas för regression och för att upptäcka avvikande värden, även kallad outliers. Fördelen med SVM är att den bland annat klarar av att hantera både linjära och icke-linjära problem, beroende på hur modellen utformas.

SVM fungerar genom att försöka hitta en optimal gräns som delar upp olika klasser i datan med så stor marginal som möjligt. Det gör modellen stabil och mindre känslig för brus eller variationer i datan. Om datan inte är linjärt separerbar kan modellen använda sig av en kernel-funktion för att omvandla datan till ett högre dimensionellt rum där det blir lättare att skilja klasserna åt. En av de vanligaste kärnfunktionerna är RBF-kernel, även kallad för gaussisk kernel. Den mäter avståndet mellan datapunkter, och omvandlar sedan dessa till likheter. En fördel med SVM är att den ofta fungerar bra vid klassificering av komplexa men små eller medelstora dataset, vilket gör den lämplig i till exempel bildklassificering, som i det här fallet med MNIST-datasetet.

I denna rapport har jag använt SVM med en RBF-kernel (Radial Basis Function) för att testa hur bra modellen kan klassificera handskrivna siffror, och sedan jämfört resultatet med logistisk regression.

3 Metod

Arbetet har genomförts i Visual Studio Code med användning av Jupyter Notebook (.ipynb)-filer. Jag använde mig av programmeringsspråket Python, och flera viktiga bibliotek användes även under arbetets gång som scikit-learn, matplotlib, och numpy. Datasetet som användes var MINST, ett klassiskt dataset med handskrivna siffror från 0 till 9 och denna laddade jag ner utifrån beskrivningarna på GitHub som tillhandahölls av kursens lärare. Jag börjar med att visa en bild av en handskriven siffra från datasetet, med hjälp av matplotlib. Bilden symboliserar en trea och jag ger även bilden färg istället för att använda standardfärgen grey.

Jag delar sedan upp datasetet i 80% träning och 20% test. Sedan bryter jag ned de 20% av testdatan till 10% validering och 10% test data och skapar och tränar en logistisk regressionsmodell. Därefter plattar jag ut varje bild från en 2D-matris till 1D-vektor med 784 element så att bilderna kan matas in i SVM och jag skalar om pixelvärdena för att förbättra modellens prestanda och stabilitet. Jag skapar en pipeline och tränar sedan en Support Vector Machine med en RBF-kernel. Jag visualiserar sedan en Confusion Matrix.

Jag jämförde båda modellerna med CV. Slutligen räknade jag ut medelvärdet för båda modellerna och skrev ut resultaten på både test accuracy samt cross-validation.

4 Resultat och Diskussion

Jag testade två olika maskininlärningsmodeller: Logistisk Regression och Support Vector Machine. För att utvärdera hur bra modellerna presterade använde jag både test accuracy och cross-validation. Logistisk Regression uppnådde en test accuracy på 87.85%, medan den genomsnittliga cross-validation accuracy låg på 88.64%. Support Vector Machine fick ett test accuracy resultat på 88.15% och cross-validation poäng på 86.43%

Efter att ha testat båda modellerna ser vi att de båda gav bra resultat. Logistisk Regressions fungerade på efter att ha testats på ny data och när vi använde olika tränings- och testuppdelningar, dvs cross-validation. Även SVM fick ett bra resultat, däremot var modellen lite bättre på att klassificera nya bilder då den fick högre poäng på test accuracy, men presterade inte lika bra under cross-validation som Logistic Regression.

SVM använde en RBF-kernel som hjälper modellen hantera mer komplexa datamönster, medan Logistic Regression är en enklare modell som kanske inte klarar av att hitta lika komplicerade mönster i datan. Detta kan vara anledningen till att SVM hade en högre test accuracy men en lägre cross-validation poäng.

Utifrån resultaten kan vi dra slutsatsen att båda modellerna fungerar bra för att klassificera MNIST-datasetet, men att det ändå finns vissa delar som skiljer sig utifrån deras prestanda. SVM gav ett lite bättre resultat på test accuracy, alltså är den bättre på att känna igen nya bilder, medan Logistic Regression var bättre på cross-validation.

Sammanfattningsvis verkar SVM vara den bästa modellen när det kommer till MNIST då den lyckades bättre på att känna igen nya bilder. Däremot är Logistic Regression också ett bra val, speciellt om vi behöver en snabbare modell eller vill ha bättre generalisering på olika träningsuppsättningar.

5 Teoretiska frågor

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

När Kalle delar upp sin data i de tre olika processerna, det vill säga “Träning”, “Validering” och “ Test” så börjar han med att “Träna” sin modell med en **träningsdata**. Den används för att lära modellen genom att justera dess parametrar och modellen optimeras baserat på detta dataset.

När det är färdigt används **valideringsdatan** för att bedöma hur bra modellen går ifrån sig. Hyperparametrar justeras och man undviker även överträning. Med valideringsdatan hjälper den till att hitta den bästa versionen av modellen innan den slutligen testas i **testdatan**. Den används för att utvärdera modellens prestanda

på tidigare osedd datan. Den visar hur mycket vi kunde skapa en generaliserad modell som kan representera hur modellen fungerar i verkliga scenarier.

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 ”valideringsdataset”?

Julia kan utvärdera modellerna på testdatan genom att exempelvis räkna ut RMSE för regression. Den modell som har lägst fel för regression eller bäst prestanda för klassificering på testdatan är det bästa valet.

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

Ett regressionsproblem är en typ av prediktionsproblem där målet är att förutsäga ett kontinuerligt värde. Det innebär att vi försöker modellera sambandet mellan indata (oberoende variabler) och utdata (beroende variabler), där utdata kan anta vilket värde som helst inom ett visst intervall. När vi har hittat en sån funktion kan vi använda den för att förutsäga värden för nya indata som modellen inte har sett tidigare.

Exempel på modeller som är linjär regression, polynomial regression och Lasso regression. **Linjär regression** kan användas när det finns en relation mellan en eller flera oberoende variabler och en beroende variabel, t.ex. för att förutsäga bostadspriser baserat på kvadratmeter och läge. Polynomial regression: när det inte finns en linjär korrelation mellan två variabler. Lasso regression är en variant av en linjär regression som kan användas bland annat när den traditionella linjära regressionsmodellen visar tecken på överanpassning (overfitting) för att sedan välja ut de mest relevanta variablerna.

4. Hur kan du tolka RMSE och vad används det till:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2}$$

Den mäter den genomsnittliga skillnaden mellan en statistisk modells förutsedda värden jämfört med de faktiska värdena.

RMSE beräknas genom att ta roten ur medelvärdet av de kvadrerade felen.

yi = de faktiska värdena

y^i = de förutsagda värdena

n = antalet observationer

Ett lågt RMSE värde innebär att modellen gör små fel, alltså betyder det att dess förutsägelser ligger nära de faktiska värdena. Medan ett högt RMSE värde betyder att felen är större och därmed innebär det en sämre modellprestation.

5. Vad är ”klassificieringsproblem? Kan du ge några exempel på modeller som används och

potentiella tillämpningsområden? Vad är en ”Confusion Matrix”?

Ett klassificeringsproblem handlar om att förutsäga rätt kategori eller etikett för given data. Modellen tränas på en uppsättning träningsdata och utvärderas på testdata innan den används för att göra prediktioner på ny data.

Några exempel är Logistisk regression som används för binära klassificeringsproblem inom exempelvis medicinska områden. Random Forest används för både binär och fler kategorisk klassificering och fungerar bra med obalanserad data som att hantera dataset där vissa klasser är sällsynta. Exempelvis bedrägeri detektering.

SVM, även kallad Support Vector Machine, är ytterligare en modell som fungerar bra för små dataset och komplexa mönster som inte heller är linjära. Man kan använda det vid

exempelvis marknadsföring där man kollar sannolikheten för en kund att köpa en viss produkt.

Confusion Matrix är en tabell som visar hur bra en klassificeringsmodell presterar. Det består av två rader och två kolumner som visar antalet true positives TP, false negatives FN, false positives FP och true negatives TN.

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

K-means clustering är en icke-superviserande maskininlärning algoritm, dvs en typ av maskininlärning där modellen inte har tillgång till fördefinierade etiketter eller rätt svar under träningen. Med modellen kan man gruppera ihop data poäng för att hitta likheter. Efter att man har grupperat poängen till grupper så letar man efter gömda trender eller information mellan dem. Ett exempel som det kan tillämpas på är kundsegmentering. Alltså att företag analyserar kundbeteenden och därmed grupperar kunder i olika segment baserat på köpvanor.

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

Ordinal encoding, one-hot encoding och dummy variable encoding är metoder för att omvandla kategoriska data till numeriska värden. Till exempel kan kategorin "färg" med värdena röd, grön och blå representeras med ordinal encoding som 1, 2 och 3 med one-hot encoding som [1,0,0], [0,1,0], [0,0,1], och med dummy variable en coding som [0,0], [1,0], [0,1] där den första kategorin utesluts för att undvika multikollinaritet.

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?

Både Göran och Julia har rätt om man kollar från olika perspektiv. Data kan antingen vara ordinal eller nominal beroende på sammanhanget då färger som röd, grön, blå inte har en naturlig ordning som siffror eller bokstäver har, och de är därmed nominella. Men om man utgår från en åsikt, som att en röd skjorta anses vara snyggast, får färgerna

en rangordning och blir då ordinala, vilket visar att tolkning är en avgörande roll i klassificeringen av datatyper.

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?

Streamlit är ett ramverk där man kan skapa dataapplikationer inom Python för maskininlärning och för datavetenskap grupper. Det kan användas för att snabbt och enkelt bygga och dela interaktiva webbappar.

6 Självutvärdering

Efter detta arbete har jag fått en bra och grundläggande förståelse för maskininlärning. Jag har fått möjligheten att arbeta med olika modeller, specifikt Logistisk Regression och Support Vector Machine med en RBF-kernel, vilket har lett till att jag har fått bättre förståelse för modellernas fördelar, begränsningar och hur olika parametrar kan påverka prestandan.

Jag tycker att jag har fått en mycket bättre förståelse för hur man tränar, testar och validerar maskininlärningsmodeller, och hur viktigt det är att anpassa metoden efter datans egenskaper. I början av arbetet tyckte jag att det var svårt att veta vilken modell jag skulle välja, eftersom det fanns så många olika alternativ. Jag behövde läsa på, söka information och titta på flera YouTube-videor för att förstå skillnaderna mellan modellerna. Till slut valde jag att arbeta med logistisk regression och SVM, eftersom de båda är vanliga och bra att jämföra.

Något som förvånade mig var att SVM kunde användas med olika kernels, vilket jag inte visste innan. Jag var osäker på vilken jag skulle använda, och det kändes lite överväldigande i början. Men tack vare videorna och allt jag läste, lärde jag mig väldigt mycket under processen.

Det var också roligt att få skriva en rapport igen då jag skrivit det på tidigare studier och det gjorde mig mer bekväm med arbetet. Överlag har den här inlämningen hjälpt mig att utvecklas både tekniskt och i min förmåga att förstå maskininlärning.

7 Källor

Wikipedia. (2024, mars 26). *Maskininlärning*. Wikipedia. https://sv.wikipedia.org/wiki/Maskininl%C3%A4rning

Wamba, S. F. (2023, januari 4). *Machine learning, explained.* MIT Sloan School of Management. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Integritetsskyddsmyndigheten. (n.d.). *Vad är maskininlärning?* IMY. http://imy.se/verksamhet/dataskydd/innovationsportalen/vagledning-om-gdpr-och-ai/teknisk-beskrivning-av-ai/vad-ar-maskininlarning/

Mercur. (n.d.). *AI och maskininlärning – vad är det och varför är det viktigt för framtiden?* Mercur. https://www.mercur.se/Kunskapsbank/AI-och-Maskininlarning--vad-ar-det-och-varfor-ar-det-viktigt-for-framtiden_

StatQuest. (2020, september 21). *Support vector machines (SVMs) for beginners - Part 1: Introduction to SVMs* [Video]. YouTube. https://www.youtube.com/watch?v=E0Hmnixke2g

Wikipedia. (2024, mars 26). *MNIST database*. Wikipedia. https://en.wikipedia.org/wiki/MNIST_database

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

StatQuest. (2020, september 21). *Support vector machines (SVMs) for beginners - Part 2: How SVM works [Video]. YouTube.* https://www.youtube.com/watch?v=fwY9Qv96DJY

Matplotlib. (n.d.). *Colormap reference* [Webbplats]. Matplotlib. https://matplotlib.org/stable/users/explain/colors/colormaps.html

StatQuest. (2020, oktober 5). *Support vector machines (SVMs) for beginners - Part 3: Choosing the right kernel* [Video]. YouTube. https://www.youtube.com/watch?v=a0Vd8QEyAQI

StatQuest. (2020, oktober 26). *Support vector machines (SVMs) for beginners - Part 4: The math behind SVMs* [Video]. YouTube. https://www.youtube.com/watch?v=pDmRNHjCZRE

StatQuest. (2020, november 2). *Support vector machines* (SVMs) *for beginners - Part 5: How to train an SVM* [Video]. YouTube. https://www.youtube.com/watch?v=s8q_OQBJpwU