George Glor

EC Utbildning

2024-03-22

Abstract

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

Koden utför maskininlärningsuppgifter med MNIST-datasetet. Den importerar nödvändiga bibliotek och laddar sedan MNIST-datasetet. Därefter delar den upp

datasetet i tränings- och testuppsättningar och sparar testuppsättningarna för senare användning. Två olika klassificerare tränas - en Random Forest Classifier och en Logistic Regression Classifier. De tränade modellerna sparas och laddas sedan för att utvärderas med hjälp av testdata. Slutsatserna baseras på noggrannhetsberäkningar och visualisering av förvirringsmatriser. Slutligen jämförs de två modellerna och den bästa modellen väljs utifrån dess prestanda.

Förkortningar och Begrepp

Detta avsnitt behövs oftast inte.

- MNIST: Förkortning för "Modified National Institute of Standards and Technology". Det är en vanligt använd dataset inom maskininlärning och omfattar handskrivna siffror.
- Random Forest: En typ av ensemble-algoritm för klassificering och regression. Den består av flera beslutsträd och använder majoritetsröstning för att fatta beslut.
- Logistic Regression: Trots namnet är detta en klassificeringsalgoritm som används för att modellera sannolikheten för ett givet utfall.
- Accuracy Score: Ett mått på hur väl en klassificerare presterar genom att beräkna antalet korrekt klassificerade instanser i förhållande till det totala antalet instanser.
- Confusion Matrix: En tabell som används för att utvärdera prestanda för en klassificeringsmodell. Den visar antalet korrekta och felaktiga klassificeringar som gjorts för varje klass.
- Visualisering av Confusion Matrix: En grafisk representation av en förvirringsmatris som ger en översiktlig bild av en klassificeringsmodells prestanda.
- Joblib: Ett bibliotek i Python som används för att spara och ladda maskininlärningsmodeller och andra Python-objekt.

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

Innehållsförteckning

1 Inledning

Maskininlärning vi behandlar data och automatiserar beslutsfattande processer över olika branscher. Ett sådant exempel är hanteringen av handskrivna siffror, där exaktheten i klassificeringen av dessa siffror är av stor betydelse i många tillämpningar såsom optisk teckenigenkänning (OCR) och postsortering.

Denna rapport syftar till att undersöka och jämföra prestandan hos två olika klassificeringsalgoritmer, nämligen Random Forest och Logistic Regression, när de tillämpas på MNIST-datasetet. Genom att analyzera och utvärdera resultaten från dessa algoritmer kan vi få insikt i deras styrkor, svagheter och lämplighet för hantering av handskrivna siffror.

Syfte och Frågeställningar

Syftet med denna rapport är att jämföra prestandan hos Random Forest och Logistic Regression-algoritmer för klassificering av handskrivna siffror representerade i MINST-datasetet.

För att uppfylla syftet kommer följande frågeställningar att besvaras:

1- Vilken algoritm, Random Forest eller Logistic Regression, uppnår högre noggrannhet vid klassificeringen av handskrivna siffror enligt MINST-datasetet?

2- Hur skiljer sig dessa algoritmers prestanda när det gäller tränings- och testdata? Vilken är mer benägen att överanpassa modellen?

3- Vilken av dessa algoritmer är mer lämpad för att hantera handskrivna siffror med tanke på komplexiteten hos MINST-datasetet

1.1 Underrubrik – Exempel

I denna uppgift ska jag förtydliga koncepten genom att ge ett konkret example hur rankdom forest och logistic regression algoritmerna ska klassificera hanskrivna i minst datasetet

2 Teori

2.1 Exempel: Regressionsmodeller

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

> Träningsdata används för att träna modellen, valideringsdata används för att finjustera modellparametrar och förhindra överanpassning, och testdata används för att utvärdera modellens slutliga prestanda.

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 använda korsvalidering på träningsdatan för att jämföra modellernas prestanda eller använda en del av träningsdatan som ett pseudo-valideringsdataset för att välja den bästa modellen.

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

Ett regressionsproblem är en typ av maskininlärningsproblem där målet är att förutsäga kontinuerliga värden. Exempel på modeller inkluderar Linjär Regression, Polynomisk Regression, och Support Vector Regression. Tillämpningsområden kan vara fastighetsprissättning eller aktiemarknadsförutsägelser.

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

RMSE (Root Mean Square Error) är ett mått på modellens förutsägelsefel. Det används för att kvantifiera hur väl en modell kan förutsäga ett resultat. Matematiskt uttrycks det som:

$$RMSE = \sqrt{\sum_i \left(y_i - \hat{y}_i\right)^2}$$

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

Ett klassificieringsproblem handlar om att förutsäga kategoriska etiketter. Modeller som används inkluderar Logistisk Regression, Beslutsträd, och Neurala Nätverk. Tillämpningsområden kan vara e-postspamdetektering eller sjukdomsdiagnostik. En Confusion Matrix är en tabell som används för att beskriva prestandan för en klassificeringsmodell.

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

K-means är en klusteranalysmetod som delar in data i K antal kluster. Den kan tillämpas på marknadssegmentering för att identifiera olika kundgrupper

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 omvandlar kategoriska data med en ordning till numeriska koder. One-hot encoding skapar en binär kolumn för varje kategori. Dummy variable encoding liknar one-hot encoding men undviker “dummy variable trap” genom att ha en kategori mindre än det totala antalet kategorier.

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; det beror på sammanhanget. Om ingen inbördes ordning finns är datan nominal. Om en ordning appliceras baserat på någon kriterium, blir datan ordinarie.

9. Kolla följande video om Streamlit: https://www.youtube.com/watch?v=ggDaRzPP7A&list=PLgzaMbMPEHEx9Als3F3sKKXexWnyEKH45&index=12 Och besvara följande fråga: - Vad är Streamlit för något och vad kan det användas till?

Streamlight är ett öppen källkods verktyg för att snabbt skapa och dela datadrivna webbapplikationer. Det används ofta för att visualisera data och bygga interaktiva prototyper för maskininlärning modeller.

2.1.1 Exempel: Lasso

Lesso är en regulariseringsteknik som används flr att linör regression och logisk regression.

2.1.1.1 Regularisering i Lasso

Regularisring i Lasso innebär att en L1 regeringstrem läggs till förlustfinktionen för att visa komplexitet, Detta främjar variabelselektion och resulterar i en sparsam modell, vilket motverkar överanpassning.

2.1.1.2 Välja Hyperparameter

För att välja hyperparametrar, inklusive regleringsstyrkan (lambda) för Lasso och Ridge regression, används korsvalidering för att hitta de bästa värdena som optimerar modellernas prestanda och undviker över- eller underanpassning.

2.1.2 Exempel: Ridge

Ridge regression är en linjär som använer L2 reglering för att minska överpassning med att lägga till kvadraten på modellens koefficienter

2.1.3 Exempel: Elastic Net

Elastic Net kombinerar Lasso och Ridge regression genom att använda både L1- och L2-reglering för att hantera variabelselektion och kollinearitet, vilket gör den mer flexibel än enskilda metoder.

2.2 Exempel: Neurala Nätverk

det är djupa maskinlärningmodller som kan identifiera komplexa mönster i data, de använder prediktiva uppgifterm bildkänning

3 Metod

Hur har du genomfört ditt arbete? Exempelvis, hur har datan erhållits?

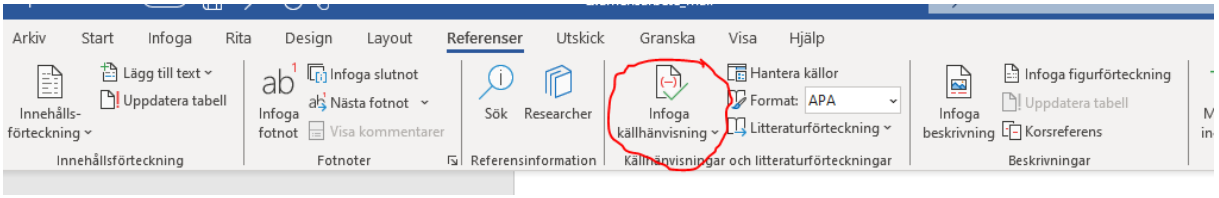
Det är att hämta data från MNIST-Datasete genot att använda denna funktion fetch_openml, från scikit learn biblotek, sen datan kommer upp och vi kan träna och testa med train_test_split-funktionen. För varje modell (Logistic Regression och Random Forest).

Slutligen användes testdata för att bedöma modellernas prestanda.

4 Resultat och Diskussion

RMSE för olika modeller	
Enkel Linjär Regression	1.8338237576746719
Lasso	`Lasso(alpha=0.1)`
Ridge	`Ridge(alpha=0.1)`

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.



Figur 1: Hur man lägger in tabell eller figur nummer samt beskrivning.

5 Slutsatser

I slutstasen sammanfattas studiens att Random forest uppvisade högre prestabde logistic regerssion flr att klassificera hanskrivning o Mnist dataset, så jag skulle rekommenderar Radnom forest som losa moell problem.

6 Teoretiska frågor

Kalle delar upp sin data i ”Träning”, ”Validering” och ”Test” för att:

- Träning: Används för att träna modellen.
- Validering: Används för att justera modellens hyperparametrar och för att bedöma dess prestanda under träningen.
- Test: Används för att bedöma modellens prestanda oberoende av de data den tränats på och validerats mot.

Julia kan använda en korsvalideringsmetod, som t.ex. k-folds cross-validation, för att utvärdera modellernas prestanda på träningsdatan. Sedan kan hon välja den modell som ger bäst prestanda på träningsdatan.
Ett regressionsproblem innebär att man försöker förutsäga kontinuerliga värden. Exempel på modeller inkluderar linjär regression, lasso regression, och random forest regression. Potentiella tillämpningsområden inkluderar prognos av aktiepriser, prissättning av fastigheter och förutsägelse av försäljningsvolymer.
RMSE (Root Mean Square Error) används för att mäta avvikelsen mellan förutsagda och faktiska värden i regressionsanalys. Det ger en indikation på hur väl modellen presterar. Ju lägre värdet på RMSE, desto bättre passar modellen data.
Ett klassificeringsproblem innebär att man försöker förutsäga en diskret klass eller kategori för varje observation. Exempel på modeller inkluderar logistisk regression, beslutsträd och support vector machines. Potentiella tillämpningsområden inkluderar spam-filtrering, medicinsk diagnos och kundsegmentering. En confusion matrix används för att visualisera prestandan hos en klassificeringsmodell genom att visa antalet korrekta och inkorrekta förutsägelser för varje klass.
K-means-modellen är en osuperviserad maskininlärningsalgoritm som används för klusteranalys och partitionering av data. Den delar in observationer i k grupper baserat på likheter i egenskaperna. Ett exempel på tillämpning är segmentering av kunder baserat på köpbeteende.

- Ordinal encoding: Omvandlar kategoriska variabler till en ordnad numerisk representation baserat på deras rangordning. Exempelvis kan {låg, medel, hög} kodas som {1, 2, 3}.
- One-hot encoding: Skapar binära variabler för varje unikt värde i den kategoriska variabeln. Varje variabel representerar en kategori och antar värdet 1 om observationen tillhör den kategorin och 0 annars.

- Dummy variable encoding: En form av one-hot encoding där en av kategorierna tas bort för att undvika multicollinearitet. Exempelvis om vi har kategorin {röd, grön, blå}, så skulle vi skapa två dummyvariabler, där en kategori inte kodas.

Julia har rätt. Begreppen "ordinal" och "nominal" beskriver olika typer av data, men tolkningen av data kan vara kontextuell. I fallet med färger är det en bra illustration: medan färgerna i sig själva är nominala (inget intrinsiskt ordningsförhållande), kan de i vissa sammanhang tolkas som ordinala (t.ex. i ett estetiskt sammanhang där en färg anses vara "vackrare" än en annan). Således är det viktigt att överväga både den inhemska naturen hos datan och dess kontext vid tolkning.

Träning Validering och Testdata: Kalle använder en uppdelning i träning, validering och test för olika mål, träning för att träna medellen, validering för att justera hyperparmaterar och bedöma prestanda under träning, testade används för att bedöma modellens.

Korsvalidering: Julia kan använda korsvalideringsmetoder som k-folds cross-validation för att utvärdera modellernas prestanda på träningsdata. hon kan välja efter modell som get bästa reslutat.

Regressionsproblem: Den involverar förutsägelse ab konttinuerliga värden, lasso regression och random forest regerssion.

RMSE (Root Mean Square Error): den är för att mäta avvikelsen mellan förutsagda och faktiska värden i regressionsnalys, lägra RMSE värden som visar idikerar en bättre passning av modell till data

Klassificeringsproblem: det handlar om att prediktera en diskret klass eller kategori för varje observation och kan lösa med modeller som logisk beslutsträd. Ex medcinsk diagnos.

K-means-modell: En maskininlärningsalgoritm för klusteranalys och partitionering av data. Den delar in observationer i k grupper baserat på likheter i egenskaperna, Ex segmentering.

Ordinal encoding: Omvandlar kategoriska variabler till en ordnad numerisk representation baserat på deras rangordning.

One-hot encoding: skapar binära variabler för att visar unikt värde där variablen repesenterar en kategori.

Dummy variable encoding: den beskriva one hot encoding där en av kategoriena ta bort för undvika multikollinearitet.

Begreppen "ordinal" och "nominal": Den visar olika typer av data, men tolkningen kan vara konterxtuell, det är viktigt att överväga båda den inhemska naturen hos data ock dess konttext vid tolkning.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
   En utmaning jag mötte var att förstå att varje träning av modellerna behövde köras separat. Detta resulterade i att jag blev fast i analysen tills jag hittade en lösning genom att lära mig mer om hur modellerna fungerar och hur de ska tränas korrekt. Detta betonar vikten av noggrannhet och förståelse för varje steg i maskininlärningsprocessen för att undvika liknande hinder i framtiden.

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

3. Något du vill lyfta fram till Antonio?

   tack för din tid och din förklaring var bäst

Appendix A

Källförteckning

Logistic Regression:
https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

https://www.kaggle.com/code/prashant111/logistic-regression-classifier-tutorial

Random Forest Classifier:

https://www.kaggle.com/code/prashant111/random-forest-classifier-tutorial

https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/