

Datorlaboration 4

Josef Wilzén

September 9, 2024

732G12 Data Mining HT2024

Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- Exempelkod för neurala nätverk: [här](#) (används i del 2)
- Kodmanual: [länk](#)
- **ISL**: An Introduction to Statistical Learning,
 - Boken: [länk](#)
 - R-kod till labbar: [länk](#)
 - Dataset: [länk](#) och [länk](#)
- **IDM**: Introduction to Data Mining
 - Kod till boken finns [här](#)
 - Sample chapters
- Dataset till vissa uppgifter finns [här](#). mnist-data finns på [Lisam](#).

Notera att ni inte behöver göra alla delar på alla uppgifter. Det viktiga är att ni får en förståelse för de olika principerna och modellerna som avhandlats. Dessa uppgifter ska inte lämnas in, utan är till för er övning.

Datauppdelning

För att motverka överanpassning bör ni dela upp data till träning-, validering-, (och testmängd). Detta kan göras med `createDataPartition()` från `caret`-paketet. Argument till den funktionen som är av vikt här är p som hur stor andel av observationerna som ska användas till träningsmängden. Ni kan också använda `subset()` för att göra detta också, men det blir svårare att tydligt ange de observationer som ska tilldelas till valideringsmängden. Denna uppdelning ska ske slumpmässigt. Notera att om en testmängd ska skapas måste uppdelningen ske en gång till från valideringsmängden.

Del 1: Installera Keras

- Installation se här. (**ISL**) och här (posit).
- Är ni i SU-salarna så ska det finnas installerat, se här (finns även på kurshemsidan) för hur ni läser in pakten.
- Se CHEAT SHEET för Keras. Dokumentation för keras finns här:
 - R interface to Keras
 - TensorFlow/Keras
 - Lista över funktioner i Keras
- Utvärdering vid klassificering kan göras med funktionen `class_evaluation_keras()`.

Del 2: Neurala Nätverk: klassificering

Målet med denna övning är att se hur bra neurala nätverk är på att känna igen bilder med siffror.

1. Läs in materialet “mnist_train.csv”¹ som är träningsmängden från MNIST databasen. Datamaterialet finns under Kursdokument på Lisam. Här lämpar sig `read.csv2()`. Läs också in testmängden “mnist_test.csv”. Neurala nätverk är känsliga för skalan på de förklarande variablerna. Dataseten som ni laddar ner från Lisam är skalat så att alla x-värden ligger mellan 0 och 1, vilket är lämpligt här². Tänk på hur keras-paketet vill ha data strukturerat och följ instruktionerna i denna kod ³.
2. Skatta ett neuralt nätverk med ett gömt lager och 10 gömda neuroner. Aktiveringsfunktionen ska vara Relu i det gömda lagret och Softmax i outputlagret. I träningsfasen ska ni ta bort den så kallade interna validering, alltså att träningsmängden delas upp till en valideringsmängd inuti algoritmen. Detta görs genom att ange `validation_split = 0`. Ange inlärningstakten till 0.1. Använd de övriga standardvärden som är angivna i koden för anpassning. Ta reda på hur många parametrar er modell har.
3. Utvärdera modellen som skattats genom att kontrollera nätverkets anpassningshistorik. Hur många epoker behövdes för att hitta en bra modell? Verkar modellen vara nog bra för att modellera datamaterialet?
4. Ta fram träffsäkerhet för träningsdata. Är modellen bra överlag? Ta fram förväxlingsmatrisen för träningsdata. Vilka siffror verkar modellen ha lyckas sämst med att prediktera? Vilka siffror har modellen predikerat de till istället? Kolla på klassspecifika utvärderingsmått. Tips: `class_evaluation_keras()`.
5. Utvärdera även modellen på er testmängd. Hur ser resultatet ut där? Ta fram förväxlingsmatrisen för testdata och jämför med träningsdata.
6. Testa nu att anpassa modellen med intern validering (30 procent) och repetera stegen ovan. Kod finns här. Hur ser resultatet ut nu? Kan man teoretiskt förvänta sig en bättre modell med denna förändring?
7. Testa nu att skatta modeller med intern validering där ni försöker ändra arkitekturen av nätverket, t.ex. aktiveringsfunktionerna, antalet gömda

¹mnist-data finns på Lisam. Detta data är “tillplattat” så värdet på pixlarna i bilderna ligger som värden i kolumner, dvs att varje rad är en observation/bild.

²När vi arbetar med neurala nätverk så är det vanligt att x skalas till att ligga mellan 0 och 1, eller att ha medelvärde 0 och standardavvikelse 1.

³Se även lab 10.9.2 A Multilayer Network on the MNIST Digit Data i **ISL**

neuroner, antalet gömda lager osv. Testa några olika modeller. Exempelpkod finns här. Vad verkar producera bättre resultat på valideringsdata?

8. Utgå från den modell som ni tyckte fungerade bäst i 7). Testa nu att skatta en modell där ni försöker ändra optimeringen av nätverket, t.ex. antalet epoker, batchstorleken, learning rate osv. Vad verkar producera bättre resultat? Förslag på upplägg:
 - (a) learning rate: 0.1, 0.05, 0.001
 - (b) batchstorleken: 32, 64, 128
 - (c) epoker: 20, 30, 40
9. Sammanfatta era modeller som ni skattat och dra slutsatser om huruvida neurala nätverk är lämpliga för denna sorts data och vad som krävs för att få bra resultat.
10. Gå igenom denna tutorial: [länk](#). I sidomenyn till höger: Klicka på “View source” ni enkelt vill ladda ner filen och öppna den i R. Här används också MNIST-data, men data läses in med funktionen `dataset_mnist()`, notera att data har annat format än när ni läste in den i 1), nu är det en array. Detta gör att inputlagret behöver vara annorlunda. Anpassa den föreslagna modellen⁴. Hur presterar den jämfört med er tidigare? Hur många parameterar skattar ni? Förutom själva arkitekturen på nätverket, hittar ni några andra skillnader jämfört med era tidigare modeller?
11. Lab 10.9.2 “A Multilayer Network on the MNIST Digit Data” i **ISL** skattar en liknande modell som ni har gjort i 2) och 10). Läs igenom labben, jämför deras precision för testdata med vad ni fått tidigare.

⁴Den modellen använder `layer_dropout()`, vi kommer att gå igenom dropout i kursvecka 4.

Del 3: Mer klassificering

Kör denna tutorial: [länk](#).

1. Vad är det för sorts data ni jobbar med här? Hur många obs? Hur stora bilder?
2. Skatta den föreslagna modellen. Hur många parameterar har den?
3. Hur presterar modellen på testdata? Beräkna förväxlingsmatrisen för test-data.
 - (a) Vilken klass var lättast att klassificera?
 - (b) Vilken klass var svårast?
4. Testa att ändra arkitekturen på modellen.
 - (a) Ändra antalet neuroner i lagren
 - (b) Testa att lägga till fler lager.
5. Testa att ändra optimeringen
 - (a) inlärningstakten
 - (b) antalet epoker
 - (c) batchstorlekek
6. Hur presterar den bästa modellen som ni lyckas hitta?
7. Se denna video: Parameters vs Hyperparameters

Del 4: Neurala Nätverk: Regression

1. Linjär regression: Om vi har ett neuralt nätverk med inga gömda noder, har identitetsfunktionen som aktiveringsfunktionen i sista lagret⁵ och tränas med MSE, då har vi en “vanlig” linjär modell. Testa följande kod, där olika regressionsmodeller skattas med keras.
2. Gå igenom labben 10.9.1 “A Single Layer Network on the Hitters Data” i **ISL**.
3. Ni ska nu analysera datasetet diamonds från ggplot2 paketet. Ni ska försöka prediktera priset baserat på de andra variablerna. Kod finns här. Skatta följande modeller:
 - (a) Modell 1: Neuralt nätverk: ett lager med 100 noder, använd relu
 - (b) Modell 2: Neuralt nätverk: två lager med 50 noder i varje lager, använd relu
 - (c) Modell 3: Neuralt nätverk: ni väljer arkitekturer på modellen
 - (d) Modell 4: En vanlig trädmodell med maxdjup på 15 och minsta antal observationer för en uppdelning är 50
 - (e) Modell 5: Random forest: ha 100 träd i modellen (om skattningen går långsamt så kan ni minska mängden träd)
4. Utvärdera modellerna på tränings- och valideringsdata med MSE. Sammanställ en tabell med dessa värden. Jämför och analysera modellerna.
5. Genomför residualanalys på träningsdata för de olika modellerna. Kod finns här (använd ‘).
6. Vilken av modellerna i 3) är ni mest nöjd med när det gäller prediktioner?

⁵I keras anger vi “linear” för att sätta identitetsfunktionen som aktiveringsfunktionen i ett lager.

Del 5: Mer regression

Ni ska nu modellera datasetet det simulerade datasetet “NN_reg_data.csv” med neurala nätverk, finns här. Datasetet har en prediktor, x och y har ett icke-linjärt samband med y .

1. Läs in datamaterialet i R. Standardisera x -variabeln till att ha medelvärde 0 och standardavvikelse 1.
2. Det finns tre olika y -variabler:
 - `y_true`: det sanna värdet på y
 - `y_less_noise/y_more_noise`: `y_true` + olika nivåer av brus.
3. Plotta de olika y mot x . Hur ser sambandet ut?
4. Börja med att ha `y_less_noise` som er responsvariabel.
5. Skatta modellerna nedan på alla observationer.
 - (a) Skatta en ploynomregression med `lm()` av ordning 20. Tips: `poly()`
 - (b) Skatta en modell med smoothing splines som skattar λ med korsvalidering.
 - (c) Skatta⁶ neurala nätverk med följande arkitekturer:
 - i. Ett gömt lager med 20 noder + Relu
 - ii. Ett gömt lager med 200 noder + Relu
 - iii. Två gömda lager med 50 + 50 noder + Relu
 - iv. Två gömda lager med 100 + 100 noder + Relu
 - (d) Gör c), men lägg till polynom av x upp till ordning 5 som förklarande variabler till neurala nätverken. Tips: `as.matrix(poly(train_x, degree = 5))`. Standardisera de nya x -variablerna innan ni använder dem.
 - (e) Plotta anpassade värden tillsammans med observerad data för de olika modellerna. Hur ser anpassningen ut?
 - (f) Ta fram MSE för träning och MSE där ni jämför de anpassade värdena med `y_true`. Hur funkar modellerna?
 - (g) Om ni vill: Skatta en trädmodell med cost complexity pruning och/eller k-nearest neighbors och jämför med modellerna ovan.
6. Dela upp data i träning/validering, där de 5% sista observationerna är valideringsmängden (alltså de längst åt höger i era plottar) och resterande i träningsdata.
 - (a) Gör om stegen i 5), men med den nya uppdelningen.

⁶Välj valfria inställningar på optimeringen. Ex: `batch_size=32, epochs = 30, learning_rate = 0.01`

7. I 5) försöker ni interpolera en funktion, i 6) försöker ni extrapolera en funktion. Vilken uppgift verkar lättast? Hur fungerade era modeller i de båda fallen?
8. Upprepa 5) och 6) fast använd `y_more_noise`. Hur presterar de olika modellerna när det finns mer burs i data?
9. Skatta några modeller på `y_true` och hela datasetet som träning.
 - (a) Skatta modeller med ett gömt lager och:
 - i. 5 noder + Relu
 - ii. 10 noder + Relu
 - iii. Så många noder som krävs för att få en mycket bra anpassning.
 - (b) Plotta observerade värden och anpassade värden i samma plot. Hur bra är neurala nätverk på att anpassa en godtycklig funktion?