CBS Februar 2023

Opgavesæt til øvelsesgang 1

Predictive Modeling og Machine Learning

I Del 1 og Del 2 af dette opgavesæt kan der med fordel kigges i koden, der blev brugt til forelæsningerne i undervisningsuge 2 og 3. At få metoderne til at køre vil i vid udstrækning være et spørgsmål om at ændre variabelnavne :-)

Del 1: Lineære metoder til regression

Denne opgave tager udgangspunkt i datasættet `Salgspriser.txt`, som omhandler salgsprisen i forbindelse med i alt 657 hushandler i de tre udvalgte områder Vanløse, Rødovre og Ballerup. Datasættet er indsamlet af Peter Dalgaard via siden www.boliga.dk i perioden fra 1. januar 2011 til slutningen af oktober 2012. I datasættet er opført variablene

- `Salgspris`: I danske kroner
- `Dato`: Den endelige salgsdato
- `Kvadratmeterpris`: Prisen pr kvadratmeter i danske kroner
- `AntalVarelser`: Antallet af værelser i huset
- `Postnummer`: Med 3 mulige værdier: `2610` (svarende til Rødovre), `2720` (svarende til Vanløse) og `2750` (svarende til Ballerup)
- `Hustype`: To mulige værdier: `Hus` og `Rakkehus` (altså rækkehus)
- `Kvadratmeter`: Husets størrelse målt i kvadratmeter
- `Byggeår`: Giver sig selv, men året, hvor huset er bygget
- `Prisreduktion`: Procentvis nedslag i pris i forhold til den oprindeligt udbudte pris

Vi vil i det følgende være interesserede i, at prædiktere værdien af outputvariablen `Kvadratmeterpris` som resultat af inputvariablene

```
AntalVarelser  Postnummer  Hustype  Kvadratmeter  Byggeaar
```

og desuden vekselvirkningerne mellem

- `Kvadratmeter` og `Postnummer`
- `Hustype` og `Postnummer`

At der er variable i datasættet, som ikke skal bruges, og der desuden skal indgå vekselvirkninger, betyder, at højresiden i modelligninger i `R`–kaldene vil skulle præciseres mere, end de er blevet det ved forelæsningerne. Her skulle kaldet

```
Kvadratmeter*Postnummer+Hustype*Postnummer+AntalVarelser+Byggeaar
```

indsat efter $\sim$ kunne klare opgaven

(a) Indlæs datasættet i `R`. Det vil være fornuftigt at starte med at overbevise `R` om, at `Postnummer` er en kategorisk variabel. Det kan klares med (hvis ellers datasættet er indlæst som `Salgspriser`) kodestumpen

```
Salgspriser$Postnummer<-as.factor(Salgspriser$Postnummer)
```

(b) Opdel nu datasættet (tilfældigt) i et træningsdatasæt og et testdatasæt – gerne med mindst 200 observationer i testsdatasættet.

(c) Benyt træningsdatasættet og hver af følgende machine–learning–metoder

- Almindelig lineær regression (med alle de nævnte variable og vekselvirkninger som inputvariable
- Best subset
- Ridge regression
- Lasso
- (Eventuelt også PCR og PLS)

til at forudsige **Kvadratmeterpris** som en lineær funktion af de ovennævnte inputvariable (inkl. vekselvirkningerne). Afprøv metoderne ved at prædiktere kvadratmeterpriserne i testdatasættet. Sammenlign med de faktiske værdier ved at udregne MSE (det kan muligvis være rarere at udregne kvadratroden af MSE, da tallene ellers ser meget uoverskuelige ud).

Til den almindelige lineære regression (som faktisk ikke blev illustreret til forelæsningen) kan noget i stil med følgende eventuelt være nyttigt (prikkerne skal erstattes af modelligningen)

```
lin.model<-lm(...,data=Salgspriser[train,])
koeff<-lin.model$coefficients
Xmat.test<-model.matrix(...,data=Salgspriser[test,])
```

```
predictions<-Xmat.test%*%koeff
MSE.lin<-mean((Salgspriser$Kvadratmeterpris[test]-predictions)^2)
MSE.lin
sqrt(MSE.lin)
```

(d) Overvej og sammenlign de forskellige metoders performance.

(e) Overvej desuden, om der er overensstemmelse mellem de variable, som "best subset" hhv. "lasso" udvælger. Ville det give samme prædiktionsresultat, hvis de to metoder udvalgte de samme variable?

Del 2: Lineære metoder til klassifikation

Her skal vi i stedet benytte datasættet `Tilsalg.txt`, som er meget stærkt beslægtet med `Salgspriser.txt`, men som indeholder oplysninger om 524 huse, der var til salg i Vanløse, Ballerup eller Rødovre i perioden fra 1. januar 2011 til slutningen af oktober 2012. I datasættet er opført mange af de variable, der kendes fra `Salgsprisert.txt` og så nogle nye. De variable, vi får brug for, er

- `AntalVarelser`: Antallet af værelser i huset
- `NettoPrMd`: Annoncerede udgifter til huset pr måned (med skattefradrag)
- `Kvadratmeter`: Husets størrelse målt i kvadratmeter
- `KvadratmeterGrund`: Grundens størrelse målt i kvadratmeter
- `Byggeår`: Giver sig selv, men året, hvor huset er bygget
- `Kvadratmeterpris`: Prisen pr kvadratmeter i danske kroner
- `Postnummer`: Med 3 mulige værdier: `2610` (svarende til Rødovre), `2720` (svarende til Vanløse) og `2750` (svarende til Ballerup)
- `Hustype`: To mulige værdier: `Hus` og `Rakkehus` (altså rækkehus)
- `Solgt180`: To mulige værdier: `TRUE` (hvis huset er solgt inden for 180 dage) og `FALSE` (hvis det ikke er)

Vi vil i det følgende være interesserede i at kunne klassificere outputvariablen `Solgt180` som resultat af de ovennævnte inputvariable.

(a) Indlæs datasættet i `R`. Det vil endnu en gang være fornuftigt at starte med at overbevise `R` om, at `Postnummer` er en kategorisk variabel (se del 1).

(b) Opdel nu datasættet (tilfældigt) i et træningsdatasæt og et testdatasæt – gerne med mindst 200 observationer i testsdatasættet.

(c) Benyt træningsdatasættet og hver af følgende machine–learning–metoder

- Almindelig logistisk regression
- Regulariseret logistisk regression (ridge og lasso)
- LDA
- QDA

til at klassificere `Solgt180` på baggrund af de ovennævnte inputvariable. Afprøv metoderne ved at klassificere `Solgt180` i testdatasættet. Sammenlign med de faktiske værdier ved at regne fejlraten og/eller AUC.

(d) Overvej og sammenlign de forskellige metoders performance.

(e) Prøv at gentage almindelig og regulariseret logistisk regression, når **alle** vekselvirkninger inddrages. Dvs, at modelligningen skal have `*` i stedet for `+` imellem variablene.
Bemærk: Det er ikke sikkert, at estimationen i den almindelige logistiske regression vil konvergere. I så fald er det ikke en fejl.

Del 3: Illustration af virkningen af ridge og lasso mht. bias og varians

Denne delopgave tager udgangspunkt i den vedlagte `R`–kode, `Øvelsesgang1.del3.R`. Grundideen er at simulere data fra den lineære model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5, \beta_6 X_6 + \xi,$$

hvor $\xi \sim N(0,1)$, og hvor

$$\beta_0 = 2, \quad \beta_1 = 0.5, \quad \beta_2 = -0.6, \quad \beta_3 = 2, \quad \beta_4 = 0.001, \quad \beta_5 = 0, \quad \beta_6 = -1.$$

Og så benytte almindelig lineær regression samt ridge– og lasso–regression til at estimere β–parametrene og prædiktere nye Y–værdier ud fra nye X–værdier. Mere præcist sker der følgende i koden:

1. Der simuleres et datasæt med $N = 25$ efter ovenstående model. Hertil skal det siges, at X–variablene alle simuleres som uafhængige $N(0,1)$–fordelte variable.
2. Parametrene estimeres med almindelig lineær regression samt med ridge– og lasso–regression på baggrund af datasættet. For både ridge og lasso benyttes λ–parametrene 0.1 og 0.3. Der benyttes således ikke krydsvalidering til at finde et optimalt λ
3. Ud fra en ny X–vektor (kaldet `xtest`) prædikteres den tilhørende Y–værdi med de i alt 5 metoder: Lineær regression og så to udgaver af ridge hhv. lasso med de to λ–værdier.
4. Punkt 1–3 gentages `nsim=1000` gange. Dog bruges den samme værdi af `xtest` i alle 1000 gentagelser. Dette giver altså 1000 estimater af β–parametrene og 1000 prædikterede værdier af Y.

5. Der tegnes histogrammer af de estimerede værdier af β_3 for hver af de 5 metoder.

6. Der tegnes histogrammer af de estimerede værdier af β_5 for hver af de 5 metoder.

7. Der tegnes histogrammer af de prædikterede værdier af Y for hver af de 5 metoder. Den sande forventede værdi er indtegnet.

8. Der udregnes gennemsnit og varians for de prædikterede værdier af Y.

Prøv nu at køre hele koden! Hvis det er meget beregningstungt, kan `nsim` eventuelt sænkes.

(a) Gå lige igennem koden, så du er nogenlunde sikker på, hvad der foregår.

(b) Kig på de 5 histogrammer over estimerede β_3–værdier. Husk, at den sande værdi er 2. Passer det med, hvordan de 5 metoder fungerer? Overvej hvilke metoder, der giver centrale estimater? Hvilke metoder giver mindst varians på esimaterne?

(c) Kig på de 5 histogrammer over estimerede β_5–værdier. Husk, at den sande værdi er 0. Passer det med, hvordan de 5 metoder fungerer? Overvej hvilke metoder, der giver centrale estimater? Hvilke metoder giver mindst varians på esimaterne?

(d) Kig på de 5 histogrammer over prædikterede Y–værdier og inddrag også de udregnede gennemsnit og varianser for prædiktionerne. Hvordan virker de forskellige metoder mht. varians og bias på prædiktionerne?