Political Data Science

Lektion 9

Estimation & prædiktion

Opgave 1

Angiv for datasættene a) til d) herunder, om vi generelt vil forvente performance af en meget fleksibel model at være bedre end en meget lidt fleksibel model. Beggrund dit svar.

a) Sample størrelsen n er meget stor, og antallet af prædiktorer p er lille.

b) Antallet af prædiktorer p er meget stort, og sample størrelsen n er lille.

c) Relationen ml. prædiktorerne og outcome er i høj grad ikke-lineær.

d) Variansen i fejlleddet, dvs. $Var(\epsilon) = \sigma^2$ er ekstremt stor.

Løsning

a) Vi vil forvente bedst performance af en fleksibel model, fordi den vil fitte data mest præcist samtidig med den store sample størrelse n sammenlignet med p mindsker sandsynligheden for overfitting idet der er nok datapunkter til at skille signalet fra støjen.

b) Vi vil forvente bedst performance af en mere restriktiv model. Det omvendte af ovenstående vil være tilfældet - en fleksibel model vil risikere at overfitte det lille antal observationer. Der er så at sige ikke nok information til rådighed til at isolere og fitte til de "sande" mønstre i data.

c) Vi vil forvente bedst performance af en mere fleksibel model, fordi den bedre vil "indfange" de ikke-lineære relationer mellem prædiktorer og outcome.

d) Vi vil forvente bedst performance af en mere restriktiv model, fordi vi ellers risikerer at overfitte til den store varians (støj).

Opgave 2

Beskriv de mest centrale forskelle på estimation og prædiktion.

Løsning

1. Prædiktion hviler på korrelationer (forskelligt fra antagelser) => vi kan måle performance
2. Vi kan måle performance => vi kan specificere vores model empirisk
3. Fokus på $\widehat{y}$ frem for f => vi kan behandle f som en black box
4. Vi interesserer os først og fremmest for out-of-sample performance

Opgave 3

Forklar om scenarierne herunder er klassifikations- eller regressionsproblemer, og indiker om vi er mest interesserede i estimation eller prædiktion. Angiv antal observationer n og prædiktorer p for hvert scenarium.

a) Vi har samlet data om de 500 største virksomheder i Europa. For hvert firma vi kender deres profit, antal ansatte, industri og direktørløn. Vi er interesserede i at forstå, hvilke faktorer, som har betydning for direktør-lønninger.

b) Vi overvejer at lancere et nyt produkt, og vi er interesserede i at vide, om det bliver en succes eller fiasko. Vi har samlet data om 20 andre lignende produkter, som blev lanceret tidligere. For hvert produkt har vi indsamlet data om, hvorvidt de blev en succes eller en fiasko, om prisen på produkterne, deres marketingbudget, konkurrenternes pris og ti andre variable.

c) Vi er interesserede i at forudsige den procentuelle forandring i Euroen som valuta i relation til ugens forandringer på aktiemarkederne. Vi har samlet ugentlige data for alle uger i 2017, og for hver uge har vi målt den procentuelle forandring i Euroen, den procentuelle forandring på det amerikanske, britiske, franske og tyske aktiemarked.

Løsning

a) Estimation. Regressionsproblem. n = 500 firmaer, p = profit, antal ansatte og industri.

b) Prædiktion. Klassifikationsproblem. n = 20 lignende produkter, p = pris, marketingbudget, konkurrenters pris og 10 andre variable.

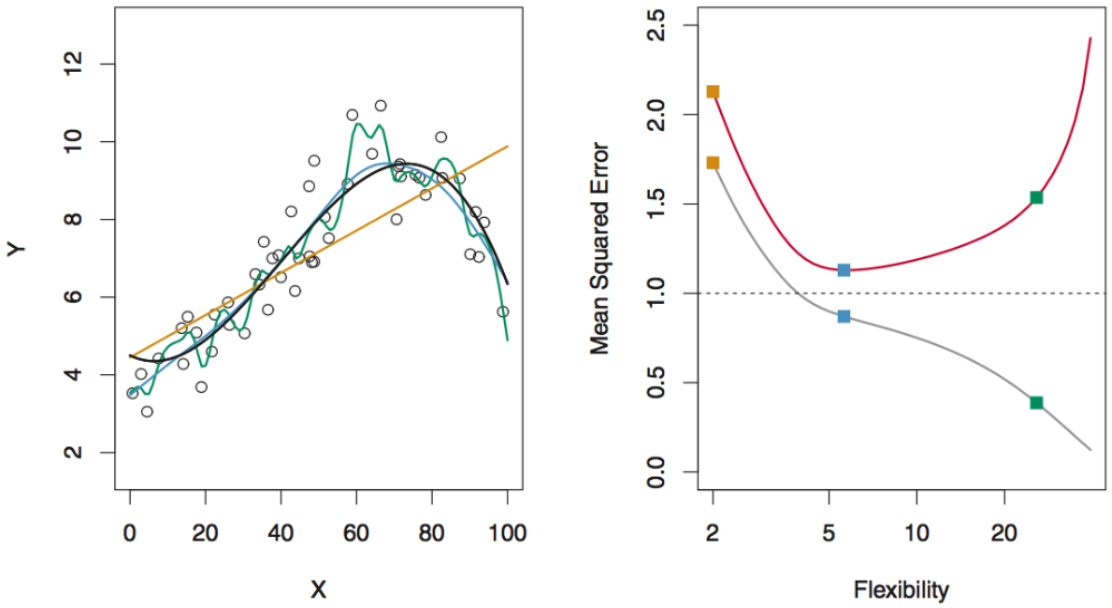
c) Prædiktion. Regressionsproblem. n = 52 uger med aktiedata, p = den procentuelle forandring på de 4 andre aktiemarkeder.

Opgave 4

Tegn det typiske tradeoff mellem bias og varians ved at tegne et plot, hvor du angiver kurver for en models bias, varians, fejl i træningssættet og fejl i testsættet. X-aksen angiver graden af fleksibilitet for din model, og y-aksen angiver værdierne for hver af dine kurver. Tegn også en vandret kurve, som viser modellens ikke-reducerbare fejl (variationen i fejlleddet). Der skal være 5 kurver i alt. Giv dem labels. Tegn fx dit plot i hånden, i paint eller på *youidraw.com*. Så er der også styr på årets julegave.

Løsning

Tegningen skulle gerne ligne:



{1}

Opgave 5

Tænk på eksempler, hvor forskellige typer af statistisk læring er relevant:

a) Beskriv 3 idéer til anvendelse af *klassifikation* i den virkelige verden. Beskriv outcome såvel som prædiktorer. Er målet estimation eller prædiktion? Beggrund dit svar.

b) Beskriv 3 idéer til anvendelse af *regression* i den virkelige verden. Beskriv outcome såvel som prædiktorer. Er målet estimation eller prædiktion? Beggrund dit svar.

Opgave 6

Hvorfor bruger vi ikke altid bare den mest fleksible model til at fitte vores data så præcist som muligt, når vi laver estimation? Og hvorfor ikke, når vi laver prædiktion?

Løsning

I estimationssammenhæng er den primære årsag, at vi interesserer os for selve funktionen f, og derfor gerne vil kunne forstå og fortolke denne. Meget fleksible funktionelle former giver ikke nødvendigvis mening for os, selvom de potentielt modellerer data meget præcist. Desuden skal vi kunne begrunde teoretisk, hvorfor vi har specificeret funktionen, som vi har, og hvorfor vi tror, at denne er en retvisende model for en datagenererende proces.

Med prædiktionssammenhæng er den primære årsag, at vi med en meget fleksibel model risikerer overfitting - at vi fitter til støjen i data. Det vil give os en ringere performance out-of-sample, fordi støjen vil betyde, at punkterne i det nye datasæt vil variere ift. det datasæt, som vi har trænet vores model på.

Opgave 7

Load pakken `MASS` i R Studio. Du har nu datasættet `Boston` til rådighed, som indeholder information housing og residential zones i Boston's forstader.

Læs om datasættet med `?Boston`.

Opstil en lineær model (hint: `lm()`), hvor estimerer sammenhængen ml. crime rate per capita (`crim`) som outcome og de øvrige variable som prædiktorer. Hvilke variable hænger sammen med outcome, og hvordan kan det fortolkes? (hint: `summary()`).

Løsning

Opgaven har til formål at prøve en helt almindelig lineær model af i R. Det kan gøres som herunder.

```
# Loader pakker
library(MASS)
library(dplyr)

# Tjekker data ud
?Boston
glimpse(Boston)

# Fitter model
m <- lm(crim ~ ., data = Boston)

# Tjekker resultater ud
```

```r
summary(m)

    # kortere afstand til employment centres korrelerer med mere kriminalitet (dis)
    # adgang til motorveje korrelerer med mere kriminalitet (rad)
    # højere andel sorte korrelerer med mindre kriminalitet (black)
    # højere huspriser korrelerer med mindre kriminalitet (medv)
```

Opgave 8

Opdel datasættet `Boston` i et testsæt og et træningssæt, fx med kode som:

```r
index <- sample(nrow(Boston), nrow(Boston)*0.8)

boston_train <- Boston[index, ]
boston_test <- Boston[-index, ]
```

Træn en lineær model på træningssættet (hint: lm()) med crime rate per capita (`crim`) som outcome og de øvrige variable som prædiktorer. Test modellen ved at prædiktere outcome, crime rate per capita, i testsættet (hint: predict()). Ser det ved øjemål ud til, at der er en sammenhæng mellem det faktiske outcome og dit prædikterede outcome?

Løsning

```r
# Opdeler datasættet i trænings- og testsæt
index <- sample(nrow(Boston), nrow(Boston)*0.8)
boston_train <- Boston[index, ]
boston_test <- Boston[-index, ]

# Træner en model på træningssættet
m <- lm(crim ~ ., data = boston_train)

# Prædikterer om testsættet
boston_test$y_hat <- predict(m, newdata = boston_test)

# Ser der ud til, der umiddelbart er en sammenhæng mellem crim og prædikteret outcome?
View(boston_test)  # ja - de ser ud til at korrelere
cor(boston_test$crim, boston_test$y_hat)  # de korrelerer positivt

# Bonus: R^2
rss <- sum((boston_test$crim - boston_test$y_hat)^2)
tss <- sum((boston_test$crim - mean(boston_test$crim))^2)
rsq <- 1 - (rss/tss)
rsq  # R^2 = 0.295, dvs vi kan forklare 29.5 % af variationen i outcome-variablen crim
```