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.

Opgave 2

Beskriv de mest centrale forskelle på estimation og prædiktion.

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 foradring på det amerikanske, britiske, franske og tyske aktiemarked.

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.

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?

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()`).

Opgave 8

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

```
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?