Political Data Science

Lektion 9: Estimation & prædiktion

Undervist af Jesper Svejgaard, foråret 2018
Institut for Statskundskab, Københavns Universitet
github.com/jespersvejgaard/PDS

I dag

1. Opsamling fra sidst
2. Dagens pensum
3. Mini-workshop
4. Næste gang

Overblik

1. Intro til kurset og R
2. R Workshop I: Explore
3. R Workshop II: Import, tidy, transform
4. R Workshop III: Programmering & Git
5. Web scraping & API
6. Tekst som data
7. Visualisering
8. GIS & spatiale data
9. Estimation & prædiktion
10. Superviseret læring I
11. Superviseret læring II
12. Usuperviseret læring
13. Refleksioner om data science
14. Opsamling og eksamen

Opsamling fra sidst

Opsamling fra sidst I

Undervisning

- **Ny lov**: "*Everything is related to everything else, but near things are more related than distant things.*" - Toblers første lov
- **Nyt mindset**: Spatial tænkning åbner mulighed for merging af data via en spatial relation, fx lys-emission og etniske gruppers organisering
- **Nye datatyper**: Raster + shapefiler (vektor)
- **Nye features**: Punkter, linjer, polygoner
- **Nye udfordringer**: Projektion af en rummelig kugle på et fladt plan
- **Nye pakker**: `sp`, `rgdal`, `raster`, `tmap` m.fl.

Opsamling fra sidst II

Workshop

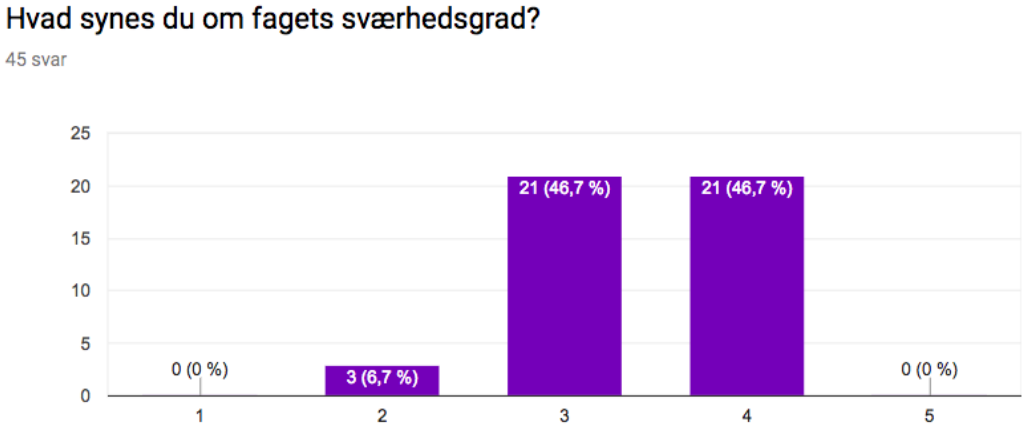
Find løsning på opgaverne fra sidste uge på Github:

`PDS/scripts/08_script.R`

Opsamling

Midtvejsevaluering

- Tak for kommentarer - mange rigtig gode inputs!
- Mere/mindre repetition, mere/mindre tid til opgaver, mere/mindre kode



Formalia & praktisk

1. Seminar-opgaver til inspiration på `PDS/seminaropgaver/`
   - Bemærk: De fra VKM, ikke PDS ⇒ forskelligt indhold/form
2. Flere datakilder til inspiration på `PDS/seminaropgaver/datakilder.txt`
3. Feedback på emne hvis tvivl om, hvorvidt man rammer inden for pladen:
   - Mulighed 1: Send mail med kort eksamensoplæg, så sender jeg et par korte kommentarer den anden vej. Muligheden er åben indtil sidste undervisningsgang, derefter lukker jeg for spørgsmål.
   - Mulighed 2: Spørg mig i undervisningen (fx i workshopdelen)

Estimation vs. prædiktion

Estimation vs. prædiktion

Prædiktion i samfundsvidenskaben

Estimation:

- Hvad kan forklare Y?

Prædiktion:

- Hvad er vores forudsigelser af Y?

Estimation vs. prædiktion

Prædiktion i samfundsvidenskaben

Breiman et al. (2001): The Two Cultures

- I samfundsvidenskaben har vi fokuseret på forklaring (≠ naturvidenskaben)

Kleinberg et al. (2015): Prediction Policy Problems

- Nogle samfundsproblemer kan bedst anskues som prædiktionsproblemer

Estimation vs. prædiktion

Umbrella problems

En policy-maker befinder sig i to forskellige situationer:

1. Der er udsigt til tørke - kan regndans betale sig?
2. Det trækker op til regn - kan det betale sig at tage sin paraply med på job?

(Kleinberg et al., 2015)

Estimation vs. prædiktion

Umbrella problems

Fordi regndans ikke har nogen direkte effekt på nytten:

$$\frac{\delta \pi}{\delta X} = 0$$

er beslutningen om regndans er et rent **kausal inferens-problem**:

$$\frac{d\pi(X, Y)}{dX} = \frac{\delta \pi}{\delta Y} \frac{\delta Y}{\delta X}$$

$\Rightarrow$ vi er kun interesserede i $\frac{\delta Y}{\delta X}$

Estimation vs. prædiktion

Umbrella problems

Fordi paraplyen ikke har nogen direkte effekt på, om det vil regne:

$$\frac{\delta Y}{\delta X} = 0$$

er beslutningen om at medbringe en paraply er et rent **prædiktions-problem**:

$$\frac{d\pi(X,Y)}{dX} = \frac{\delta \pi}{\delta X} * Y$$

$\Rightarrow$ vi er kun interesserede i $\widehat{Y}$

Estimation vs. prædiktion

Umbrella problems: Eksempel

Kleinberg et al. (2015):

- Målretning af kirurgiske indgreb for patienter med slidgigt
- 500.000 amerikanske patienter er berettiget til indgrebet årligt via Medicare
- Nytten af indgrebet er kendt ($\delta\pi/\delta X$)
- Nytten af betinget af den forventede levetid ($\widehat{Y}$)
- Hvis indgrebet 'omfordeles' fra de 10 % af patienterne med højst forudsagt dødelighed til øvrige patienter, der ellers har fået afslag, kan 10.512 nytteløse indgreb forhindres årligt

"*Predicion policy problems are, in sum, important, common, and interesting, and deserve much more attention from economists than they have received.*"

Estimation vs. prædiktion

Estimationsproblemer og prædiktionsproblemer

Estimationsproblemer:

- Når vi har adgang til Y og $\mathbf{X}$ og er interesserede i, hvordan Y ændrer sig afhængigt af $\mathbf{X}$
- Eksempel: Hvad kan forklare uddannelsesfrafald?
- Eksempel: $Y = f(\mathbf{X})$

Prædiktionsproblemer:

- Når vi har adgang til $\mathbf{X}$, men ikke til Y, som vi er interesserede i
- Eksempel: Hvilke studerende har den største risiko for at frafalde?
- Eksempel: $\widehat{Y} = f(\mathbf{X})$

Om vi skal lave estimation eller prædiktion afhænger af forskningsspørgsmålet

Estimation vs. prædiktion

Centrale forskelle

Estimation:

- Vi er interesserede i $f()$ => modelspecifikationen er vigtig
- Hviler på antagelser =>
  - performance kan ikke måles
  - teoridreven modelspecifikation

Prædiktion:

- Vi er interesserede i $\widehat{Y}$ => vi kan behandle $f()$ som en black box
- Hviler på korrelationer =>
  - performance kan måles
  - datadreven modelspecifikation

Måling af prædiktionsperformance

Måling af prædiktionsperformance

Klassifikations- og regressionsproblemer

Klassifikationsproblemer:

- Når outcome er kategorisk
- Eksempler: Stemmer/stemmer ikke, frafald/ikke frafald

Regressionsproblemer:

- Når outcome er kontinuert
- Eksempler: Ledighedsperiode, ejendomspriser, sandsynlighed for kriminalitet

Måling af prædiktionsperformance

Regressionsproblemer: MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

Måling af prædiktionsperformance

Klassifikationsproblemer: Accuracy

$$Accuracy = 1 - errorrate = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

Accuracy:

- Andelen af korrekte klassifikationer. Intuitivt, right?

Problem:

- Ubalanceret outcome => høj accuracy ved at forudsige majoritets-klassen
- Vi er ikke lige interesserede i $y = 0$ og $y = 1$

Løsning:

- Performance-mål som indeholder mere information, fx confusion matricer

Måling af prædiktionsperformance

Confusion-matrix I

		Prædikteret outcome	
		$\hat{y}_i = 1$	$\hat{y}_i = 0$
Faktisk outcome	$y_i = 1$	Sande positive (SP)	Falske negative (FN)
	$y_i = 0$	Falske positive (FP)	Sande negative (SN)

Måling af prædiktionsperformance

Confusion-matrix II

		Prædikteret outcome	
		$\hat{y}_i = 1$	$\hat{y}_i = 0$
Faktisk outcome	$y_i = 1$	228	331
	$y_i = 0$	518	3287

Måling af prædiktionsperformance

Fastsættelse af tærskelværdi

Vi prædikterer ikke $\widehat{y}_i = 0 \vee \widehat{y}_i = 1$

Vi prædikterer $\widehat{y}_i = \widehat{p}(y_i = 1|\mathbf{X}_i)$

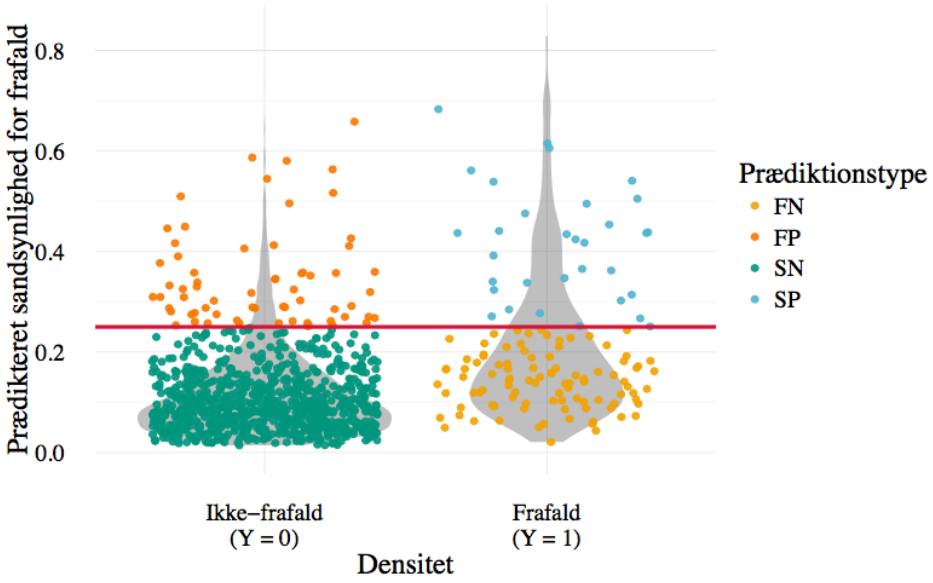
Derfor har vi behov for at sætte en tærskelværdi:

$$\widehat{y}_i < \delta \Rightarrow \widehat{y}_i = 0$$

$$\widehat{y}_i \geq \delta \Rightarrow \widehat{y}_i = 1$$

Måling af prædiktionsperformance

Tærskelværdi og prædiktionstyper



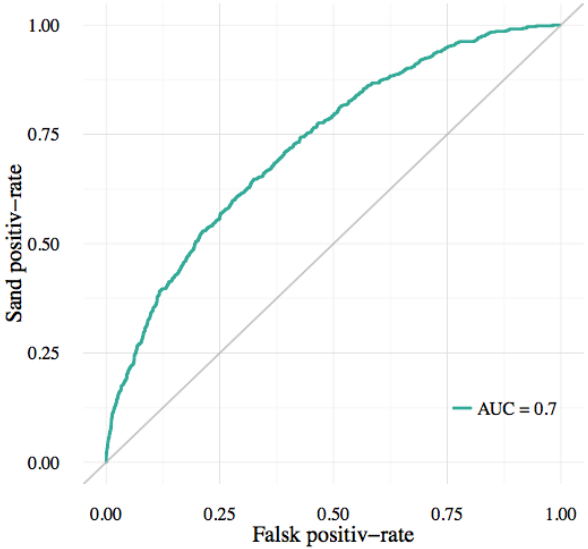
Måling af prædiktionsperformance

Prædiktionstyper og performancemål

		Prædikteret outcome		Performancemål
		$\hat{y}_i = 1$	$\hat{y}_i = 0$	
Faktisk outcome	$y_i = 1$	SP	FN	$Sand\ positiv\text{-}rate = \frac{SP}{SP+FN} = \frac{SP}{\sum y_i=1}$
	$y_i = 0$	FP	SN	$Falsk\ positiv\text{-}rate = \frac{FP}{FP+SN} = \frac{FP}{\sum y_i=0}$

Måling af prædiktionsperformance

ROC og AUC

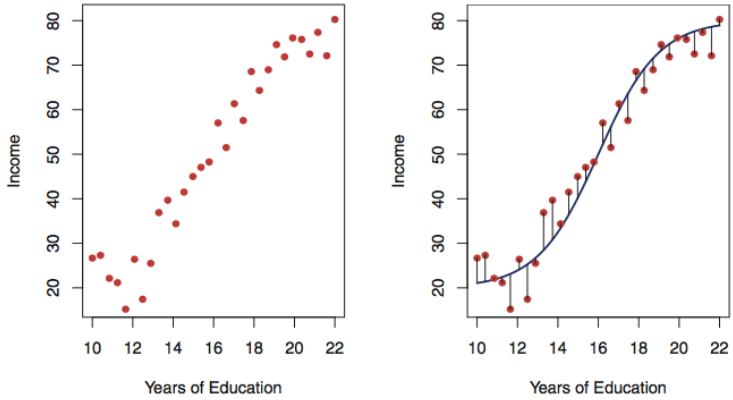


Statistisk læring

Statistisk læring

Hvad er det?

- Overlappende betegnelser: statistisk læring, maskinlæring, data mining m.fl.
- Tilgange til at estimere en funktion f, der beskriver smh. ml. et outcome og en række uafhængige variable, fx $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$



Statistisk læring

Superviserede og usuperviserede metoder

Superviseret læring

- Der eksisterer et outcome Y, vi vil forklare, prædiktere eller klassificere efter
- Eksempel: Forudsigelse af uddannelsesfrafald

Usuperviseret læring

- Der eksisterer ikke et outcome Y på forhånd
- Eksempel: Gruppere vælgere på baggrund af p antal holdningsspørgsmål

Statistisk læring

Hvordan estimerer vi f?

Vi estimerer f vha. algoritmer, dvs. et sæt af regler for, hvordan et problem løses, fx minimeringen af en loss-funktion.

Eksempel: OLS

$$\widehat{f}_{OLS} = \arg\min_{f} \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \sum_{j=1}^{J} \widehat{\beta}_j x_{ij})^2$$

Statistisk læring

Estimation $\Rightarrow$ prædiktion

So what's the fuss about?

- Vi kan behandle f som en black box => nye algoritmer bliver egnede
- Vi kan måle prædiktionsperformance => empirisk modelspecifikation + tuning

Statistisk læring

Den datagenererende proces

Vi kan antage en datagenererende proces:

$$y = f(\mathbf{X}) + \epsilon$$

Hvor y er et outcome givet af en matrix af uafhængige variable, $\mathbf{X}$, samt fejlledet ϵ, der af uafhængigt af $\mathbf{X}$ og varierer tilfældigt.

Pba. observationer om y og $\mathbf{X}$ ønsker vi at finde frem til en tilnærmelse på f ved:

$$\hat{y} = \hat{f}(\mathbf{X})$$

Målet er, at $f(\mathbf{X}) = \hat{f}(\mathbf{X})$.

Statistisk læring

Grænser for prædiktionsperformance I

Modellens fejl består af to led, en reducerbar og en ikke-reducerbar fejl:

$$E[y - \hat{y}]^2 = E[f(\mathbf{X}) + \epsilon - \hat{f}(\mathbf{X})]^2 = [f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2 + Var(\epsilon)$$

Fejlen er altså lig med **kvadreret bias** + **variansen i fejlleddet**.

Målet er at minimere fejlen ved at finde frem til den bedste tilnærmelse af f i form af $\hat{f}$.

Fejlen $Var(\epsilon)$ skyldes fænomenets natur og er ikke-reducerbar, og sætter derfor en grænse for vores prædiktionsmodels performance.

Statistisk læring

In-sample vs out-of-sample-performance

Estimation:

- Mål: at forklare sammenhæng i et datasæt ⇒ minimere loss in-sample

Prædiktion:

- Mål: at prædiktere sammenhænge i nyt data ⇒ minimere loss out-of-sample

Udfordring:

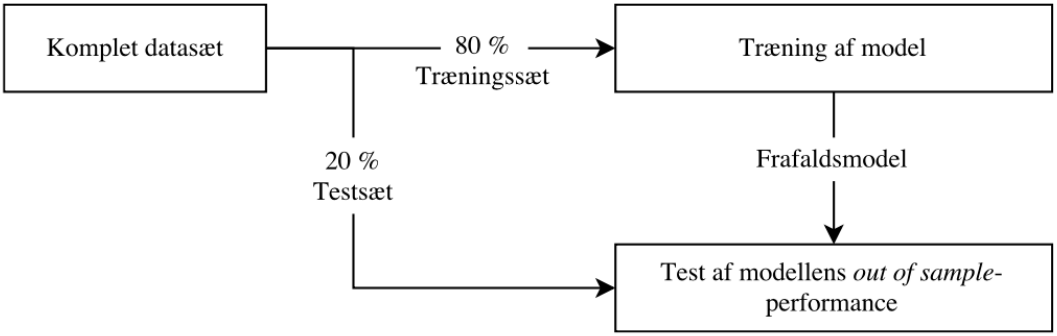
- Ingen garanti for at højt in-sample fit = højt out-of-sample fit (pga støj)

Løsning:

- Opsplitning i test- og træningssæt

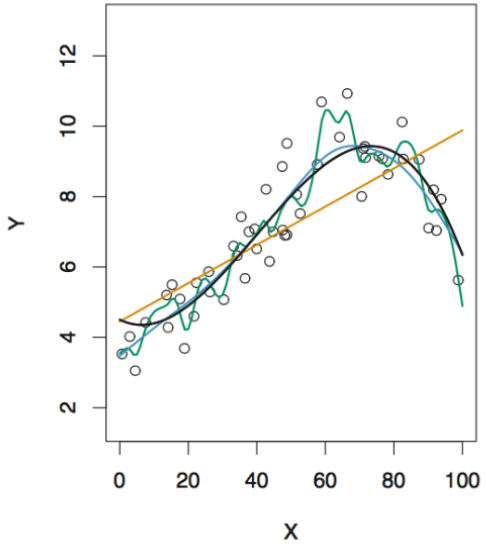
Måling af prædiktionsperformance

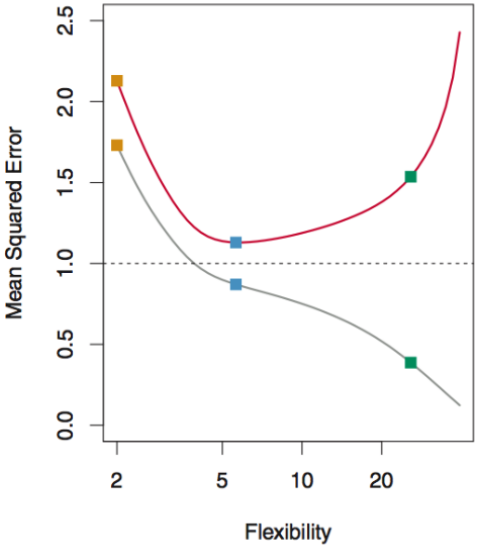
Trænings- og testsæt



Statistisk læring

Underfitting og overfitting





Statistisk læring

Bias-variance tradeoff

Den forventerede MSE i testsættet for en given observation x_0 kan nedbrydes til:

$$E\left[y_0 - \hat{f}(x_0)\right]^2 = \underbrace{E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]}_{Var(\hat{f}(x_0))} + \underbrace{\left(E\left[\hat{f}(x_0)\right] - f(x_0)\right)^2}_{Bias(\hat{f},f)^2} + Var(\epsilon)$$

Indsigter:

- Alle led er positive = den forventede MSE kan ikke blive mindre end $Var(\epsilon)$
- Vi ønsker at $\widehat{f}$ varierer mindst muligt (på tværs af træningssæt)
- Vi ønsker at $\widehat{f}$ er den bedst mulige tilnærmelse på f
- Tradeoff: Mere fleksible modeller har typisk højere varians, men mindre bias - og omvendt

Statistisk læring

De vigtigste pointer

- Sondring I: **Superviserede** / **usuperviserede** metoder
- Sondring II: **Regressionsproblemer** / **klassifikationsproblemer**
- Fokus i prædiktion: **Out-of-sample performance**
- Fokus på $\widehat{y} \Rightarrow f$ som **black box** $\Rightarrow$ **komplicerede algoritmer**
- Prædiktion hviler på **korrelationer** $\Rightarrow$ **test af performance** $\Rightarrow$ empirisk minimering af tradeoff ml. bias og varians ved **empirisk modelspeficikation** og **tuning**
- Fejlled indgår i DGP $\Rightarrow$ grænse for en models performance
- Større **fleksibilitet** $\Rightarrow$ mindre bias + større varians (risiko for **overfitting**)

Mini-workshop

Mini-workshop

Find opgaverne på Github under PDS/opgaver/:

`09_opgaver.R`

Næste gang

Næste gang

- Indhold:
  - Superviseret læring I
- Pensum:
  - ISL: kap 4 afs 4.1-4.3 (logistisk regression: skimmes)
  - ISL: kap 5 afs 5.1-5.2 (resampling metoder: læses)
  - ISL: kap 6 afs 6.2-6.2.1 (regularisering: skimmes)
  - Varian (2014): læses
- DataCamp:
  - Supervised Learning in R: Regression

Tak for i dag!