Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

BigInsight

Hvordan åpne den svarte boksen?

Forklaring av prediksjoner

Martin Jullum

med Kjersti Aas, Anders Løland og Nikolai Sellereite

NRs styremøte, 10.oktober 2018

Hva slags situasjoner skal vi jobbe med?

- Beskriver sammenhengen mellom en responsvariabel Y basert på en mengde forklaringsvariable, $\boldsymbol{x} = x_1, \ldots, x_p$ ved hjelp av en statistisk modell eller maskinlæringsmodell
- Bruke modellen til å predikere Y for nye $\boldsymbol{x}$-er
- Eksempler $\boldsymbol{x} \rightarrow Y$

→ Salgspris bolig

→ Neste film hun vil se

→ Hund/ katt

Transaction history

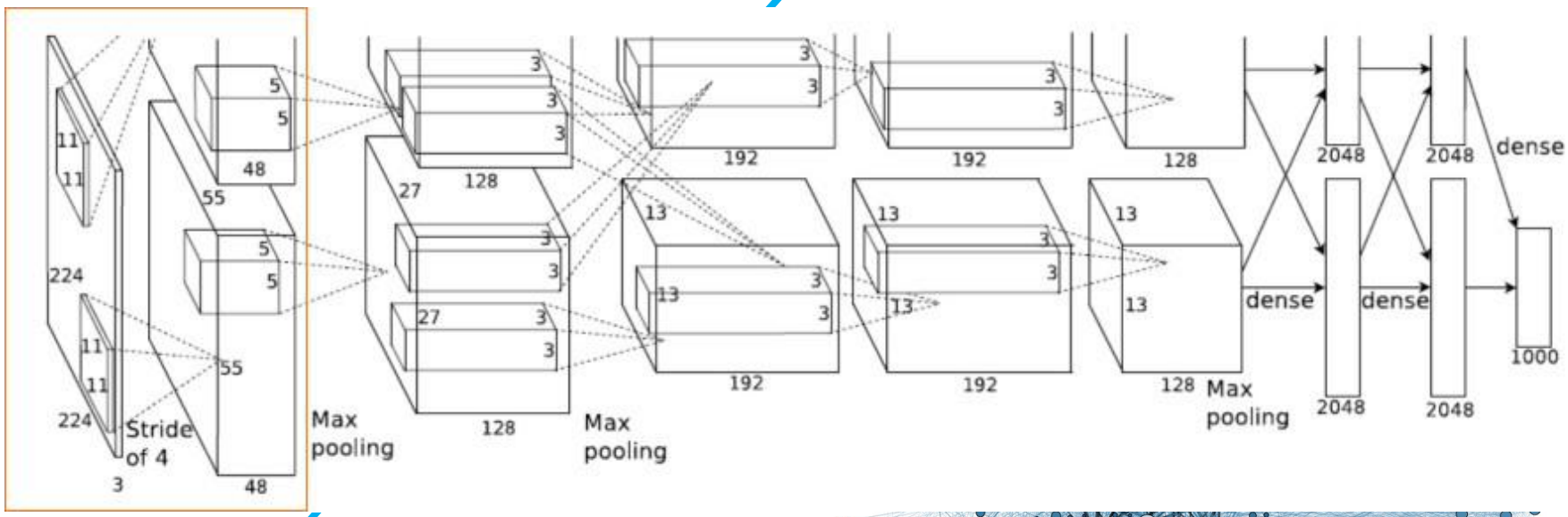
Commercial Card - 456480111111111: 08/04/2004 - 14/04/2004

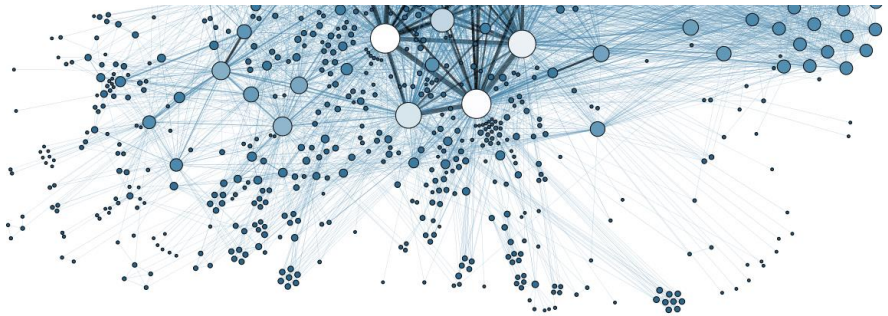
Date Processed	Description	Debit	Credit
08/04/2004	CASH ADVANCE FEE	$5.00-	
09/04/2004	SUNSHINE VILLAGE BANFF	$86.97-	
10/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	
10/04/2004	CARD MEMBERSHIP FEE	$19.00-	
10/04/2004	PHOTOCARD FEE	$3.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
11/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	
12/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	
12/04/2004	CARD MEMBERSHIP FEE	$19.00-	
12/04/2004	PHOTOCARD FEE	$3.00-	
12/04/2004	PRINCIPAL DEBIT ADJUSTMENT	$22.00-	

→ Misligholder lån?

Hva mener vi med svarte bokser?

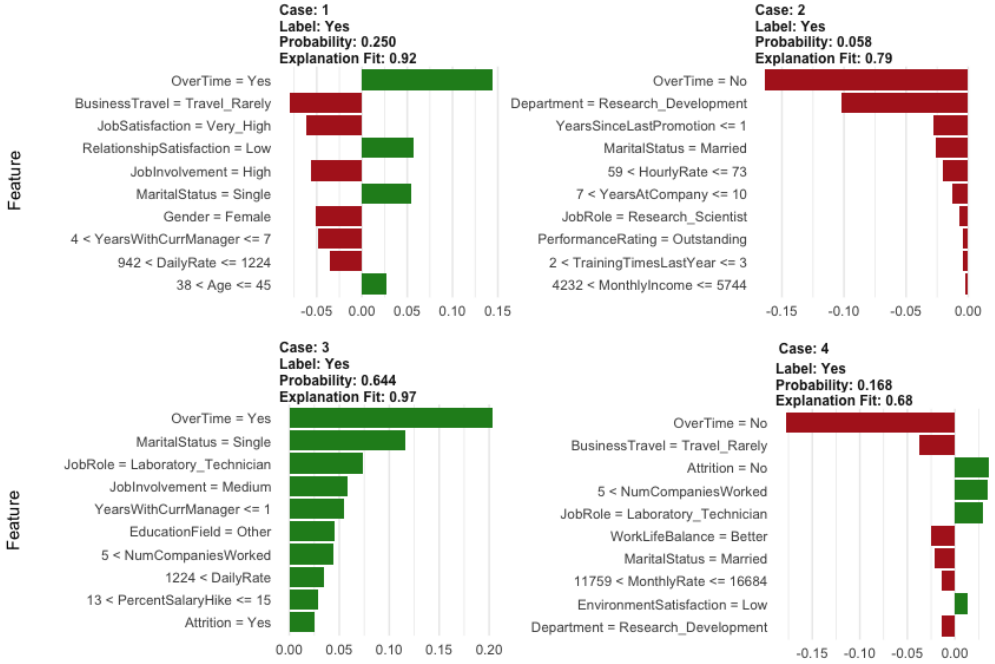
- Den svarte boksen er den statistiske modellen/ maskinlæringsmodellen: $Y \approx f(\boldsymbol{x})$





Hva mener vi med å forklare?

- Ønsker å forklare **individuelle prediksjoner** fra modellen – **ikke** modellen som helhet
- For et spesifikt sett med variable: $\boldsymbol{x} = x_1, \ldots, x_p$
  Hvilke variable bidro positivt/negativt til prediksjonen $\hat{Y} = f(\boldsymbol{x})$
  - Og (typisk) hvor mye?
- Representerer bidragene som **én** score for ***hver*** variabel: $\phi_1, \ldots, \phi_p$
- Indivduell forklaring for ***hver enkelt*** prediksjon

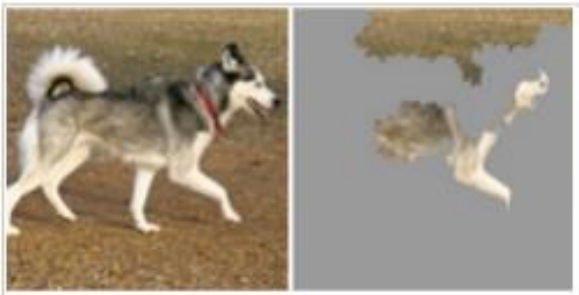


Motivasjon

- Klassifisering husky/ulv basert på bilde (deep learning)

Motivasjon

Predicted: wolf
True: wolf



Predicted: husky
True: husky

Predicted: wolf
True: wolf

Predicted: wolf
True: husky

Predicted: husky
True: husky

Predicted: wolf
True: wolf

GPDR – kan kreve forklaring

Official Journal of the European Union

L 119

Volume 59

English edition — Legislation — 4 May 2016

Contents

I *Legislative acts*

REGULATIONS

★ **Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)** (1) 1

DIRECTIVES

★ **Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA** 89

★ **Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime** 132

(1) Text with EEA relevance

EN

Acts whose titles are printed in light type are those relating to day-to-day management of agricultural matters, and are generally valid for a limited period.

The titles of all other acts are printed in bold type and preceded by an asterisk.

- GDPR = General Data Protection Act
  Gjelder også for Norge fra juli i år.
- "…in the existence of automated decision-making, including profiling, [the subjects have the right to be provided with] meaningful information about the logic involved."

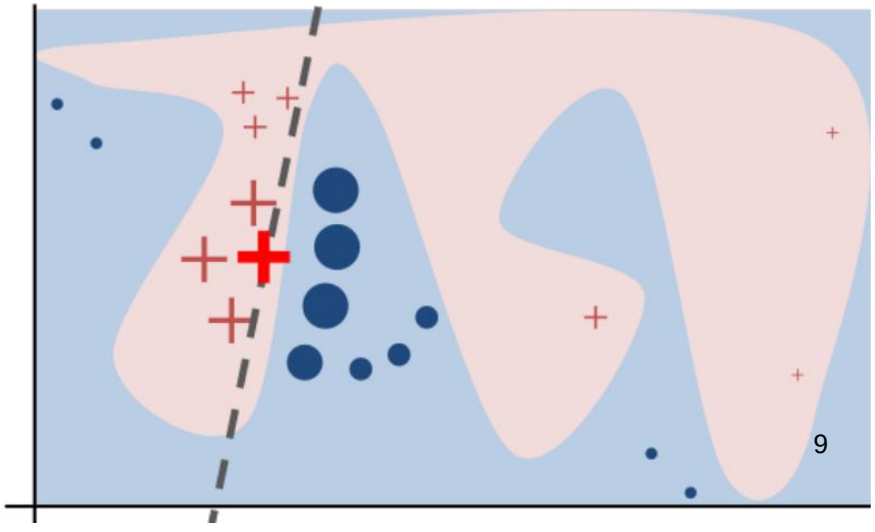
Eksisterende metodikk for prediksjonsforklaring

- Ferskt forskningsområde
  - Har oppstått med Big data-/maskinlæring-/data science-/AI-bølgen

- 2 eksisterende hovedmetoder
  - LIME (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanation)
  - SHAP (**Sh**apley **A**dditive ex**P**lanations)

LIME

- LIME (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanation)
  - Lager en enkel lokal (lineær) modell rundt hver prediksjon
  - Konseptet introdusert i Ribeiro et al. (2016), 720 siteringer
  - Konseptuelt enkelt
  - Ingen klar definisjon/algoritme => Mange varianter
  - Hvis lineær
    - $f(x) \approx \phi_0 + \phi_1 x_1 + \phi_2 x_2 + \cdots$
    - ϕ_j: **Hvordan endres prediksjonen når man endrer** x_j
  - Skaleringsproblematikk
  - Ingen matematisk begrunnelse
  - Ønsker man lokal forklaring?



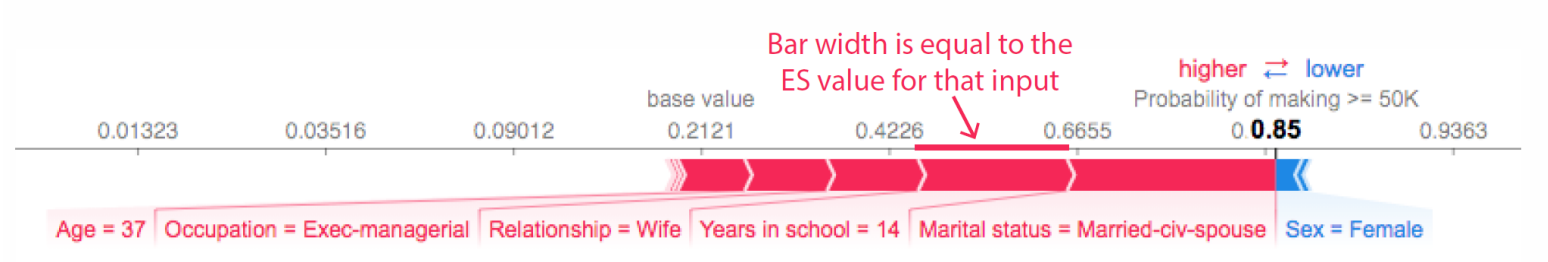
SHAP

- Shapley verdier
  - Konseptet stammer fra spillteori, der det lenge har blitt brukt til å fordele utbetaling til spillerne i et spill basert deres bidrag

- SHAP (**Sh**apley **A**dditive ex**P**lanations)
  - Spillerne = variablene ($x_1, \ldots, x_p$), utbetaling = prediksjonen ($f(\boldsymbol{x})$)
  - ϕ_j: **Hvordan endres prediksjonen om man ikke kjenner verdien av** x_j
  - Lundberg & Lee (2017), 60 siteringer
    Strumbelj & Kononenko (2014), 28 siteringer

SHAP

- Flere gode matematiske egenskaper:
  - Dekomponerer prediksjonen direkte: $\sum_{j=0}^{p} \phi_j = f(\boldsymbol{x})$

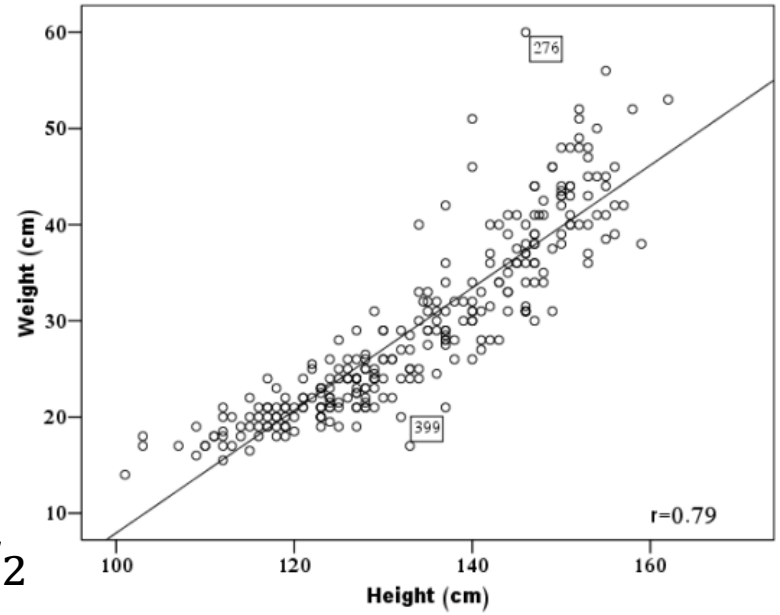


  - Konsistens ved endret modell, og likt bidrag => lik ϕ_j
- Vanskeligere å forstå hvordan bidrag fordeles:

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} (v(S \cup \{j\}) - v(S)) \,, \quad v(S) = E[f(\boldsymbol{x}) | \boldsymbol{x}_S]$$

Avhengige forklaringsvariable

- I så godt som all modellering ($Y \approx f(\boldsymbol{x})$) er det avhengighet mellom variablene $\boldsymbol{x} = (x_1, \ldots, x_p)$
- Eksempel
  - $x_1 =$ høyde (cm)
  - $x_2 =$ vekt (kg)
  - Y = Rekord i høydehopp (cm)
- Modell 1: $Y = 100 + 2x_1 - 2x_2$
- Modell 2: $Y = 100 + 0.5x_1 - 0.5x_2$
- Modell 3: $Y = 100 - 2x_1 + 2x_2$
- Selv ikke en enkel lineær modell kan forklares uten å ta hensyn til avhengighet



Avhengighet i forklaringsmetoder

- Både LIME og SHAP ignorerer avhengighet
- Kan gi helt feil forklaring hvis sterk avhengighet
- LIME
  - Trekker variable til lokal modell helt uavhengig + lokalt modell "velger" typisk en av variablene ved sterk avhengighet
  - Kan basere modell på umulige variabelkombinasjoner
- SHAP
  - Gjør approksimasjonen: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S] \approx E[f(\boldsymbol{x})]$
  - Gir typisk avhengige variable for lite bidrag

Vårt arbeid

- Forstå litteraturen!
- Reparere SHAP ved å ta hensyn til avhengigheten mellom variablene
  - Forsøke å estimere alle betingede forventninger

$$E[f(\boldsymbol{x})|\boldsymbol{x}_S] = E[f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S)|\boldsymbol{x}_S] = \int f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S) p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S) d\boldsymbol{x}_{\bar{S}}$$

- 3 retninger
  - Anta at $p(\boldsymbol{x})$ er Gaussisk => Analytisk $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S)$ + Monte Carlo integrasjon
  - Bruke en empirisk fordelingsfunksjon for $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S)$ + Monte Carlo integrasjon
  - Bruke en maskinlæringsmetode til å estimere $E[f(\boldsymbol{x})|\boldsymbol{x}_S]$ direkte

Utfordringer med vårt arbeid

- Vanskelig å estimere betingede forventninger
- Beregningsmessig tyngre enn å ignorere uavhengighet
- Finne gode defaultverdier for “tuningparametere”

- Publisere en god og forståelig artikkel før noen andre gjør det
- Få folk til å ta i bruk metoden vår

Oppsummering

- Prediksjonsforklaring er viktig
- Vi gjør individuell forklaring, ikke forklaring av hele modellen
- NR kan være i forskningsfronten

- LIME er en populær forklaringsmetode
- SHAP er godt rammeverk for forklaring

- Begge metoder ignorerer avhengighet mellom x-ene
- Vi reparerer SHAP!