Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

BigInsight

Hvordan åpne den svarte boksen?

Forklaring av prediksjoner

Martin Jullum

med Kjersti Aas, Anders Løland og Nikolai Sellereite

Lunsjpresentasjon NR, 23.november 2018

Hva slags situasjoner skal vi jobbe med?

- Statistisk modell eller maskinlæringsmodell trent opp å beskrive responsvariabel Y basert på forklaringsvariable, $\boldsymbol{x} = (x_1, \ldots, x_p)$
- Bruker modellen til å predikere Y for nye $\boldsymbol{x}$-er
- Eksempler $\boldsymbol{x} \rightarrow Y$

→ Salgspris bolig

→ Hund eller katt?

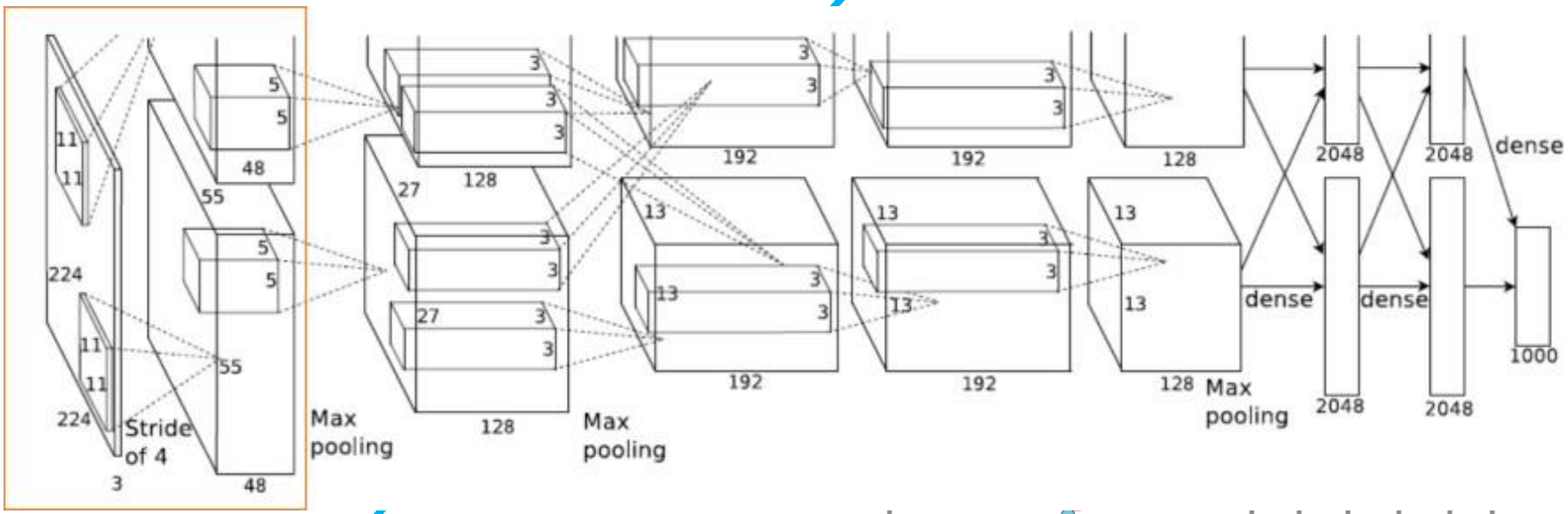
Transaction history

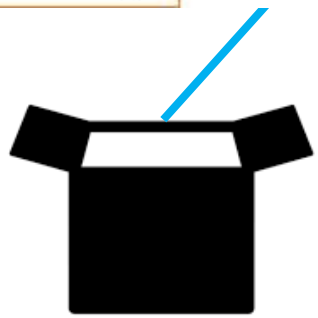
Commercial Card - 456480111111111: 08/04/2004 - 14/04/2004			
Date Processed	Description	Debit	Credit
08/04/2004	CASH ADVANCE FEE	$5.00-	
09/04/2004	SUNSHINE VILLAGE BANFF	$86.97-	
10/04/2004	PRINCIPAL CREDIT ADJUSTMENT	$22.00-	
10/04/2004	CARD MEMBERSHIP FEE	$19.00-	
10/04/2004	PHOTOCARD FEE	$3.00-	
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-	

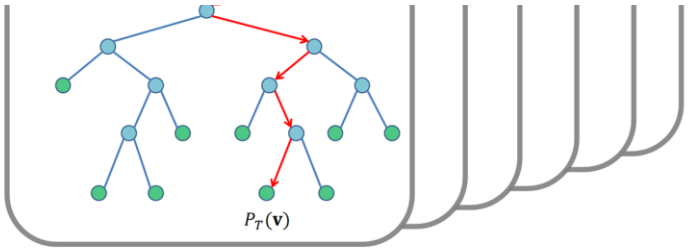
→ Misligholder lån

Hva mener vi med svarte bokser?

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

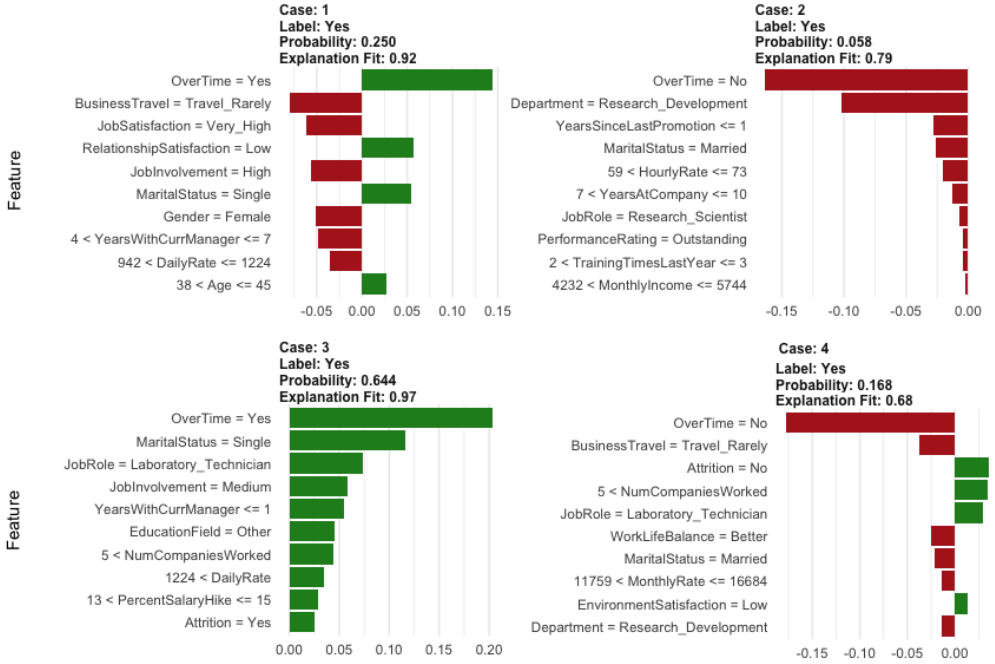






Hva mener vi med å forklare?

- **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 $\widehat{Y^*} = f(\boldsymbol{x}^*)$
  - Og (typisk) hvor mye?
- Angir forklaringen som **én** score for ***hver*** variabel: $\phi_1, \ldots, \phi_p$
- Individuell forklaring for ***hver enkelt*** prediksjon

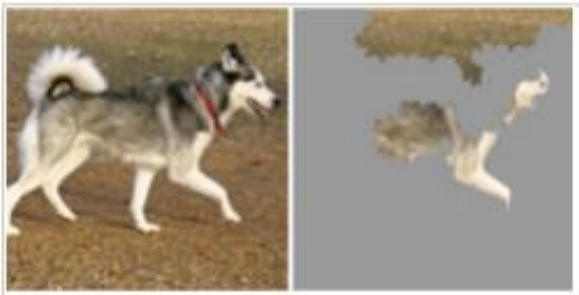


Motivasjon

- 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

GDPR – kan kreve forklaring

Official Journal of the European Union

L 119

English edition — Legislation — Volume 59, 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

- "…in the existence of automated decision-making, including profiling, [the subjects have the right to be provided with] **<u>meaningful information about the logic involved</u>**."

Eksisterende metodikk for prediksjonsforklaring

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

- Forsøker egentlig å svare på to ulike spørsmål

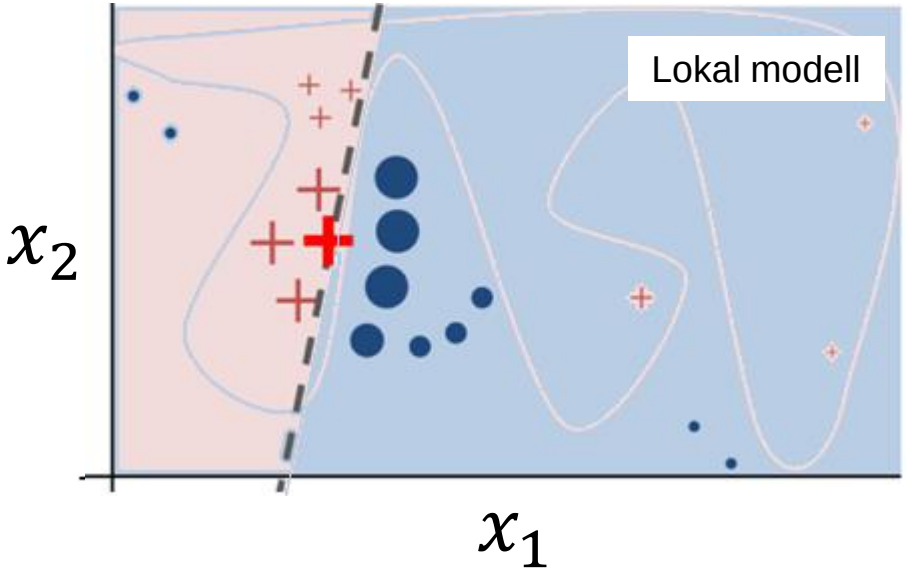
(Ribeiro et al. (2016), 720 siteringer)

LIME

(**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanation)

- Lager en enkel lokal (lineær) modell rundt hver prediksjon

EKSEMPEL: Binær klassifisering



$f(\boldsymbol{x}^*) \approx \phi_0 + \phi_1 x_1 + \phi_2\, x_2$

Tolkning ϕ_j: Hvordan endres prediksjonen **lokalt** når man endrer x_j

Rødt kryss: $\boldsymbol{x}^*$ som skal forklares

Symbolstørrelse angir nærhet til $\boldsymbol{x}^*$

Rosa område: Klasse Y=0
Lyseblått område: Klasse Y=1

Grensen mellom rosa/lyseblått angir "decision boundary" for den svarte boksen $f(\boldsymbol{x})$

(Ribeiro et al. (2016), 720 siteringer)

LIME

(**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanation)

- Lager en enkel lokal (lineær) modell rundt hver prediksjon

- Individuell forklaring med **lokalt** referansenivå
- Ingen klar definisjon/algoritme => Mange varianter
- Ingen matematisk begrunnelse
- God idé, men mange praktiske utfordringer

(Shapley (1953), 7532 siteringer
Lundberg & Lee (2017), 60 siteringer)

SHAP

- Shapley verdier
  - Konseptet stammer fra (lag-)spillteori, der det brukes til å fordele utbetaling til spillerne i et spill basert deres bidrag
- SHAP (**Sh**apley **A**dditive ex**P**lanations)
  - Spillerne = variablene ($x_1, \dots, x_p$)
  - Utbetaling = prediksjonen ($f(\boldsymbol{x})$)
  - Eksplisitt formel for ϕ_j
  - Matematiske egenskaper
    - Dekomponerer prediksjonen direkte: $\sum_{j=0}^{p} \phi_j = f(\boldsymbol{x})$
    - Konsistens ved endret modell, og likt bidrag => lik ϕ_j

NR

SHAP

(Shapley (1953), 7532 siteringer
Lundberg & Lee (2017), 60 siteringer)

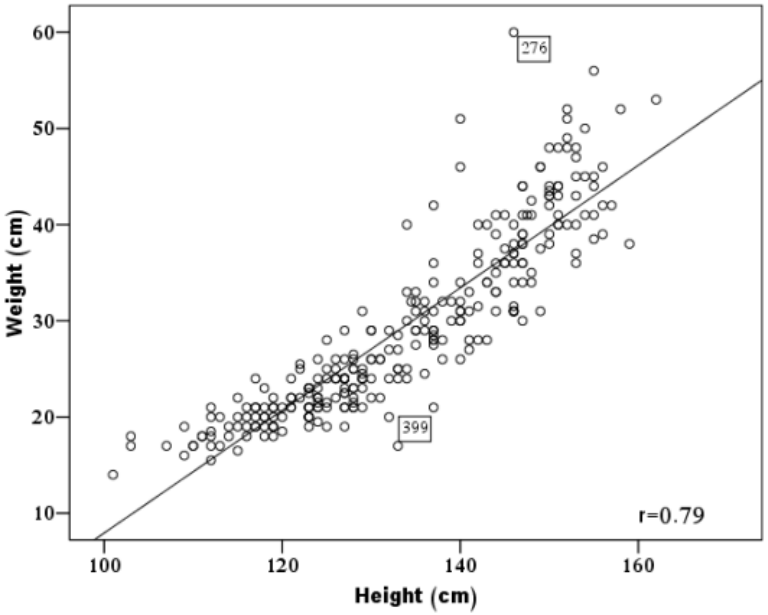
- Individuell forklaring med **globalt** referansenivå
- Løs tolkning ϕ_j: **Hvordan endres prediksjonen om man ikke kjenner verdien av** x_j
- Komplisert å forstå hvordan bidrag fordeles:

$$\phi_j = \sum_{S \subseteq M \setminus \{j\}} w(S)\left(v(S \cup \{j\}) - v(S)\right), \qquad w(S) = \frac{|S|!\,(|M| - |S| - 1)!}{|M|!}$$

- Hovedingrediens: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*]$, må approksimeres

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: $Y = 100 - 2x_1 + 2x_2$



- Selv ikke en enkel lineær modell kan forklares uten å ta hensyn til avhengighet

LIME og SHAP ignorerer avhengighet

- Kan gi helt feil forklaring hvis sterk avhengighet
- LIME: Trekker variable til lokal modell uavhengig
- SHAP: Approksimerer betinget forventning med marginal forventning: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*] \approx E[f(\boldsymbol{x}_{\bar{S}}, \boldsymbol{x}_S^*)]$
- Begge metoder krever evaluering av ulike $f(\boldsymbol{x})$-er
  - Kan basere forklaring på prediksjon av umulige variabelkombinasjoner
    - Eksempel
      - Alder = 17
      - Sivilstatus = Enke
      - Yrke = Professeor

Vårt arbeid: Reparere SHAP

- Vi vil ta hensyn til avhengigheten mellom variablene
- SHAP sier $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*) \approx p(\boldsymbol{x}_{\bar{S}})$ som del av approksimeringen av $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*]$
- Vi forsøker å estimere $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$ skikkelig
- 3 retninger
  - Anta $p(\boldsymbol{x})$ er Gaussisk => Analytisk $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$
  - Anta copula med Gaussisk avhengighetsstruktur
  - Bruke en type empirisk fordelingsfunksjon for $p(\boldsymbol{x}_{\bar{S}}|\boldsymbol{x}_S = \boldsymbol{x}_S^*)$

Oppsummering

- Forklaring av individuelle prediksjoner – ikke hele modellen!
- Prediksjonsforklaring er viktig!

- LIME forklarer med **lokalt** referansenivå
- SHAP forklarer med **globalt** referansenivå

- Begge metoder ignorerer avhengighet mellom x-ene
- Vi reparerer SHAP ved å ta hensyn til avhengigheten