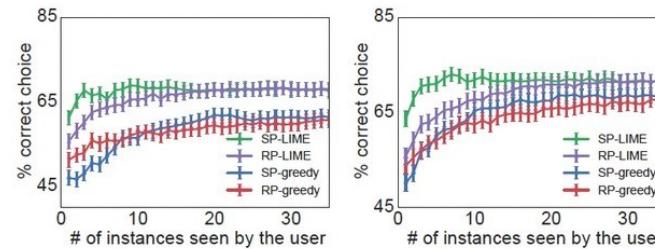


Forklarbarhet ved bruk av maskinlæringsmodeller i forvaltningsavgjørelser

Åse Haram, 23. september 2022

Problemstilling og tilnærming

- Hva vil et krav om forklarbarhet innebære dersom maskinlæringsmodeller skal brukes som beslutningsstøtte eller for å automatisere beslutninger i forvaltningen?*



$$(1) \quad \phi_j(v) = \phi_j = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}} \frac{|\mathcal{S}|!(M - |\mathcal{S}| - 1)!}{M!} (v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})), \quad j = 1, \dots, M,$$

that is, a weighted mean over all subsets \mathcal{S} of players not containing player j . Note that the empty set $\mathcal{S} = \emptyset$ is also part of this sum. The formula can be interpreted as follows: Imagine the coalition being formed for one player at a time, with each player demanding their contribution $v(\mathcal{S} \cup \{j\}) - v(\mathcal{S})$ as a fair compensation. Then, for each player, compute the average of this contribution over all possible combinations in which the coalition can be formed, yielding a weighted mean.

To illustrate the application of (1), let us consider a game with three players such that $\mathcal{M} = \{1, 2, 3\}$. Then, there are 8 possible subsets: $\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}$, and $\{1, 2, 3\}$. Using (1), the Shapley values for the three players are given by

$$\begin{aligned} \phi_1 &= \frac{1}{3} (v(\{1, 2, 3\}) - v(\{2, 3\})) + \frac{1}{6} (v(\{1, 2\}) - v(\{2\})) + \frac{1}{6} (v(\{1, 3\}) - v(\{3\})) + \frac{1}{3} (v(\{1\}) - v(\emptyset)), \\ \phi_2 &= \frac{1}{3} (v(\{1, 2, 3\}) - v(\{1, 3\})) + \frac{1}{6} (v(\{1, 2\}) - v(\{1\})) + \frac{1}{6} (v(\{2, 3\}) - v(\{3\})) + \frac{1}{3} (v(\{2\}) - v(\emptyset)), \\ \phi_3 &= \frac{1}{3} (v(\{1, 2, 3\}) - v(\{1, 2\})) + \frac{1}{6} (v(\{1, 3\}) - v(\{1\})) + \frac{1}{6} (v(\{2, 3\}) - v(\{2\})) + \frac{1}{3} (v(\{3\}) - v(\emptyset)). \end{aligned}$$

Let us also define the non-distributed gain $\phi_0 = v(\emptyset)$, that is, the fixed payoff which is not associated to the actions of any of the players, although this is often zero for coalition games.

By summarizing the right hand sides above, we easily see that they add up to the total worth of the game: $\phi_0 + \phi_1 + \phi_2 + \phi_3 = v(\{1, 2, 3\})$.

The Shapley value has the following desirable properties

Efficiency: The total gain is distributed:

$$\sum_{j=0}^M \phi_j = v(\mathcal{M})$$

Symmetry: If i and j are two players who contribute equally to all possible coalitions, i.e.

$$v(\mathcal{S} \cup \{i\}) = v(\mathcal{S} \cup \{j\})$$

for every subset \mathcal{S} which contains neither i nor j , then their Shapley values are identical:

$$\phi_i = \phi_j.$$

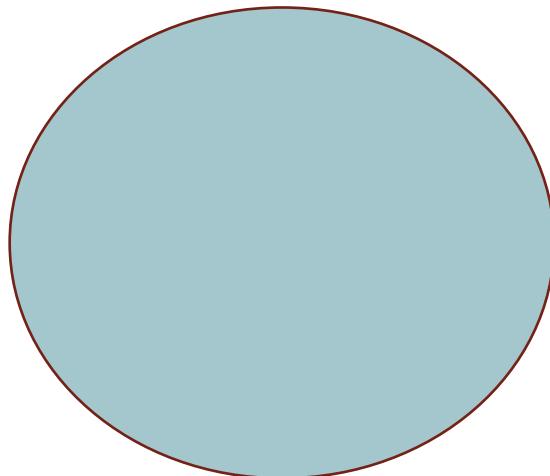
Dummy player: If $v(\mathcal{S} \cup \{j\}) = v(\mathcal{S})$ for a player j and all coalitions $\mathcal{S} \subseteq \mathcal{M} \setminus \{j\}$, then

$$\phi_j = 0.$$

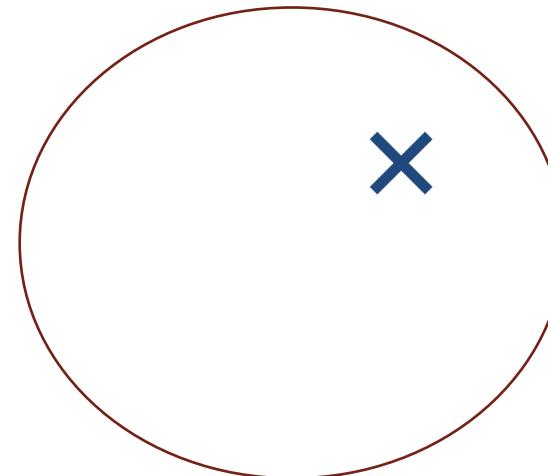
$$\min_{b_1, b_2, \dots, b_p \in \{-10, -9, \dots, 9, 10\}} \frac{1}{n} \sum_{i=1}^n \log \left(1 + \exp \left(- \sum_{j=1}^p b_j x_{i,j} \right) \right) + \lambda \sum_j \mathbb{1}_{[b_j \neq 0]}$$

Global og lokal forklaring

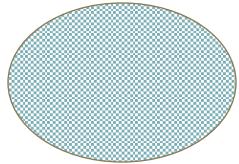
Modellnivå



Enkeltprediksjon



Global og lokal forklaring i en forvaltningskontekst



- Grl § 100 / offl § 3:
allmenhoffentlighet
- fvl § 16: forhåndsvarsel
- fvl § 11: veiledning
- GDPR art. 5, 13, 14, 15, 25, 35

- fvl § 11: veiledning
- fvl § 17: uttalerett
- fvl § 18: partsinnsyn
- fvl §§ 24-25: begrunnelse
- GDPR art. 22

Globale forklaringer

- En overordnet beskrivelse av modellens bruksområde
- Hvordan modellen opptrer «i snitt» for utvalgte grupper
- En «brukerveiledning»
- Lovfestede krav til systemdokumentasjon:
 - GDPR 25,35
 - EMK art. 8
 - Ny arkivlov og forvaltningslov
 - AI-lovgivning fra EU



- Overview of the model,
- How it learned,
- What training data,
- Limitations to the model and use restrictions.

Beaudouin et al. (2020)

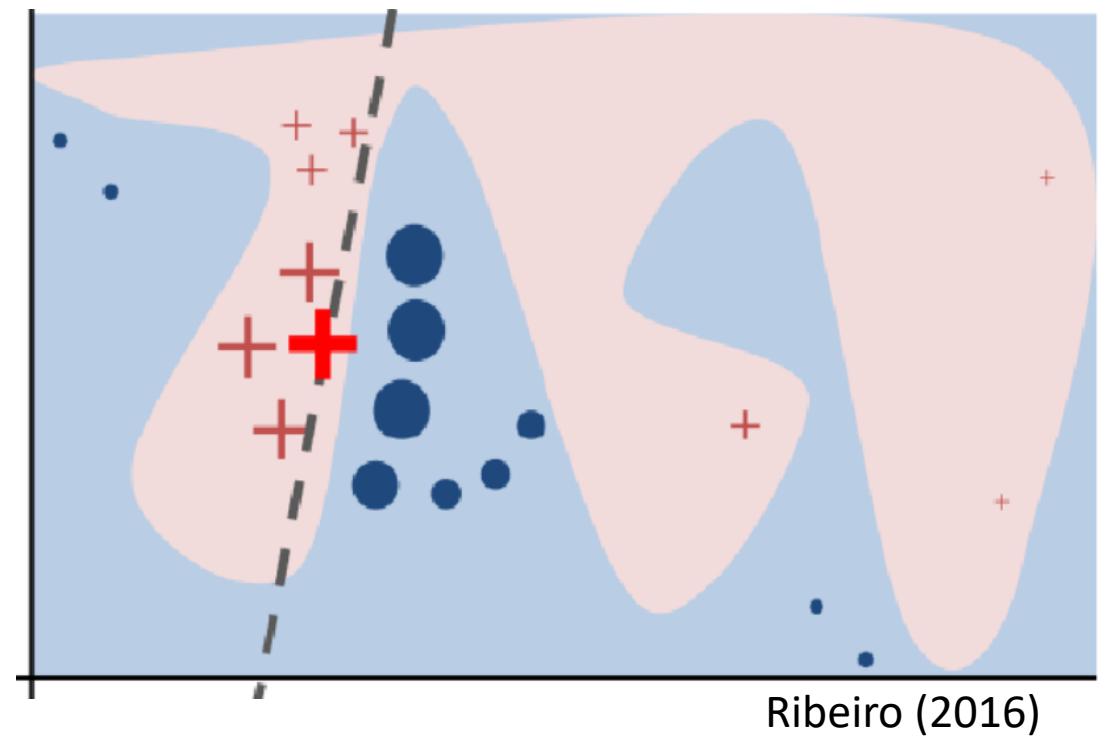
Lokale forklaringsmetoder

(I) Kontrafaktisk forklaring – «Hva hvis?»

- Hvilken endring av egenskaper ville ført til et annet utfall?
- Enkle å få til teknisk
- Man kan generere flere forklaringer for å fylle ut bildet
- Brukervennlige: Parten settes i stand til å forstå avgjørelsen og kan foreta de endringer som er nødvendig for å få ønsket resultat

(II) Approksimasjon med Shapley-verdier

- **Approksimasjon:** En maskinlæringsmodell som er en forenklet representasjon av en kompleks modell
- **Shapley-verdi:** Den relative betydningen av en egenskap (variabel) for et enkeltutfall.



Kan lokale forklaringsmetoder fungere?

Kontrafaktisk forklaring

- Tilfredsstiller ikke kravene til forvaltningsrettslig begrunnelse
- Kan være nyttig som supplement til begrunnelse eller der modellen fungerer som beslutningsstøtte

Approksimasjon med Shapley-verdier

- Strukturen minner om en forvaltningsrettslig begrunnelse
- Antakeligvis ikke stabil og rettssikker nok til å brukes i automatiserte begrunnelser
- Kan være hensiktsmessig i beslutningsstøttetilfellene – men saksbehandleren må ha global forståelse av modellen



Noen refleksjoner

- Lokale og globale forklaringer må suppleres med en form for «systembeskrivelse» som viser hvordan beslutningsprosessen er innrettet og hvordan systemet utøver myndighet
- Forvaltningslovens begrunnelseskrev er lite tilpasset algoritmisk forvaltning
- Forvaltningen må fokusere på formidling av forklaringer

Formidling av forklaringer

