Invited session: Opening the black box

Session program

Martin Jullum (Norwegian Computing Center)

Efficient Shapley value explanation through feature groups

Kary Främling (Umeå University, Sweden)

Why explainable AI should move from influence to contextual importance and utility

Homayun Afrabandpey (Nokia Technologies, Finland)

Model interpretability in Bayesian framework

Please feel free to post questions in the chat ***during*** the talks!

Chair: Martin Jullum (jullum@nr.no)

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

BigInsight

Efficient Shapley value explanation through covariate groups

Martin Jullum

Joint work with Kjersti Aas and Annabelle Redelmeier

Nordstat 2021, Tromsø June 21th-24th 2021

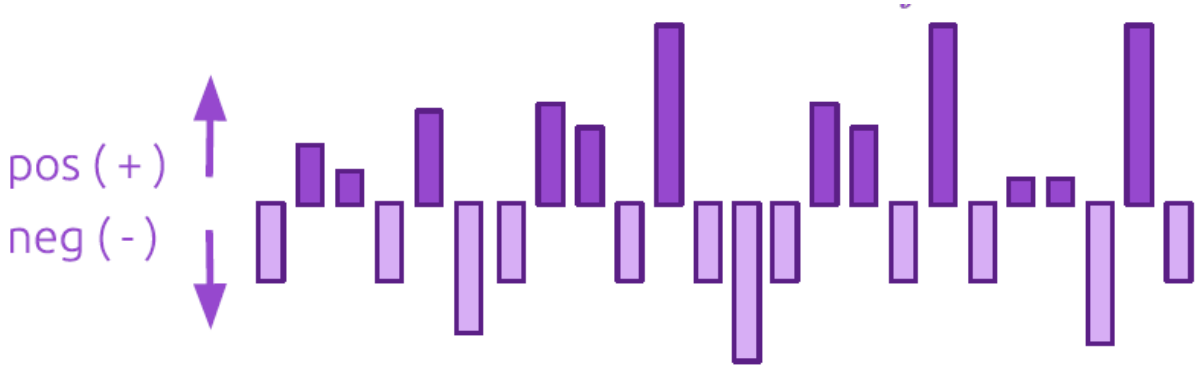
Prediction explanation

- Assume a standard regression situation
  - Response variable y
  - Set of covariates $\boldsymbol{x} = (x_1, \ldots, x_p)$
  - A fitted regression model $f(\boldsymbol{x})$
- We apply f to a new set of covariates $\boldsymbol{x} = \boldsymbol{x}^*$, providing the prediction $f(\boldsymbol{x}^*)$
- Prediction explanation
  - Which covariates or types of covariates contributed the most to the specific prediction $f(\boldsymbol{x}^*)$, and in what way did they contribute? (local explanation)
  - Note: Not the same as saying which covariates where most informative when fitting f (global explanation)

Prediction explanation – example

- Car insurance pricing
  - Advanced statistical/machine learning model f built based on historical data to model risk of a crash
  - **Covariates** $\boldsymbol{x}$:
    - Age of driver, education level, gender, # years driving
    - # claims last 5 years, # licence record points, any previous licence revokes?
    - type of car, age of car, value of car…

- Question: Why did a guy with covariate $\boldsymbol{x}^*$ get a probability for crashing the next year $f(\boldsymbol{x}^*)$= 0.3?



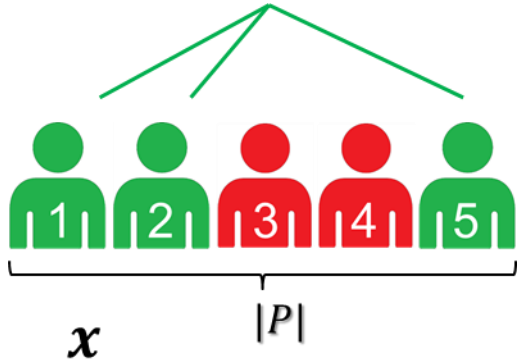
Shapley values

- General concept
  - Stems from cooperative game theory (Shapley, 1953)
  - Used to distribute the total payoff to the players ($\sum_j \phi_j = \text{payoff}$)
  - Explicit formula for the "fair" payment to every player j:

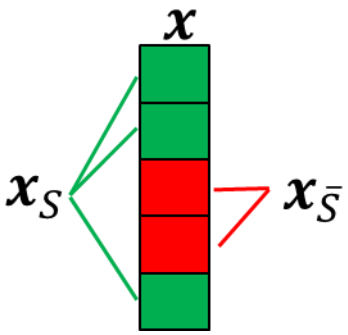
$$\phi_j = \sum_{\text{all } S \text{ without } j} w(|S|)\left(v(S \cup \{j\}) - v(S)\right)$$

w is a certain weight function,
$v(S)$ is the payoff with only players in subset S

  - Several mathematical optimality properties



- Within explainable AI/prediction explanation
  - Massive attention since Lundberg&Lee (2017) >3000 citations
    - Players = covariates ($x_1, \ldots, x_p$)
    - Payoff = prediction ($f(\boldsymbol{x}^*)$)
    - Contribution function: $v(S) = E[f(\boldsymbol{x})|\boldsymbol{x}_S = \boldsymbol{x}_S^*]$
  - Rough interpretation of ϕ_j
    - The prediction change caused by observing x_j



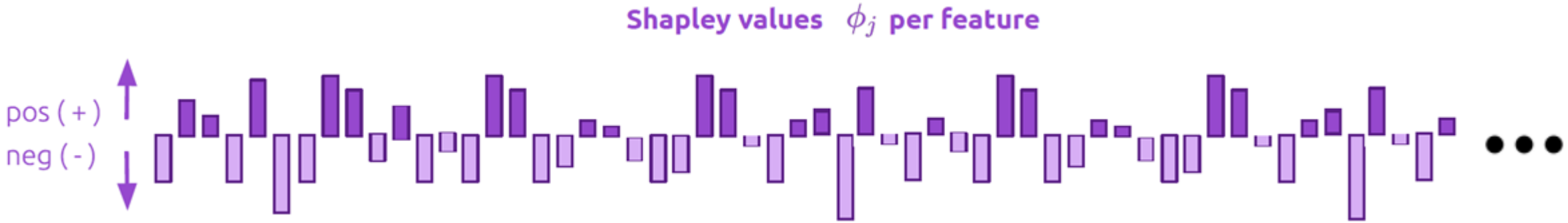
Bottlenecks

$$\phi_j = \sum_{\text{all } S \text{ without } j} w(|S|)\left(v(S \cup \{j\}) - v(S)\right)$$

$$p = 5 \Rightarrow 2^p = 32$$
$$p = 10 \Rightarrow 2^p = 1024$$
$$p = 20 \Rightarrow 2^p > 10^6$$
$$p = 40 \Rightarrow 2^p > 10^{12}$$
$$p = 100 \Rightarrow 2^p > 10^{30}$$
$$p = 1000 \Rightarrow 2^p > 10^{301}$$

1. The sum in the Shapley value formula is of size 2^p, growing exponentially in the number of covariates
   - Computational infeasible for large number of covariates

2. How can we visualize, interpret and extract knowledge from 100s or 1000s of Shapley values?



- Typically: the sum of many small ϕ_j > sum of the few large ones
- Many highly dependent covariates complicates the interpretation

Our suggestion: groupShapley

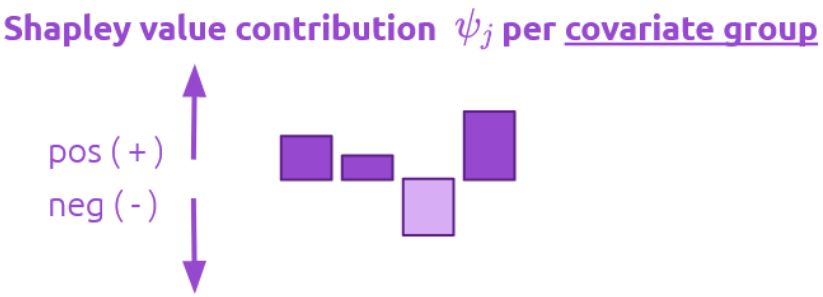
- Fundamentally very simple suggestion
  - Divide the p covariates into a small number G of groups
  - Replace the covariate subsets S in the Shapley formula by group subsets T:

$$\psi_g = \sum_{\text{all } T \text{ without } g} w(|T|)\left(v(T \cup \{g\}) - v(T)\right)$$

  - The scores are still Shapley values, so all mathematical properties are kept (on group level)
- What about the bottlenecks?
  - $2^G \ll 2^p \Rightarrow$ computationally tractable

G small $\Rightarrow$ easy to visualize

- Grouping method controls interpretation of ψ_g
  - Practical interpretation: Group by covariate type
  - Theoretical interpretation: Highly dependent covariates in the same group



Theoretical result

- To ease the interpretation (while ignoring the computational issue), an alternative grouping score has been suggested in the literature:

$$\psi'_g = \sum_{j \in g} \phi_j$$

- Question: Do we ever have $\psi'_g = \psi_g$?
  - Answer: Yes!
    - Any partially additively separable prediction function with between-group independence $\left(f(\boldsymbol{x}) = \sum_{g=1}^{G} f_g\left(\boldsymbol{x}_g\right)\right)$ $\left(\boldsymbol{x}_g \perp\!\!\!\perp \boldsymbol{x}_{g'}\right)$
- Curiosity
  - If we decompose ψ_g on the covariates in the group using Shapley values, we get the so-called Owen values – an alternative game theoretic distribution method

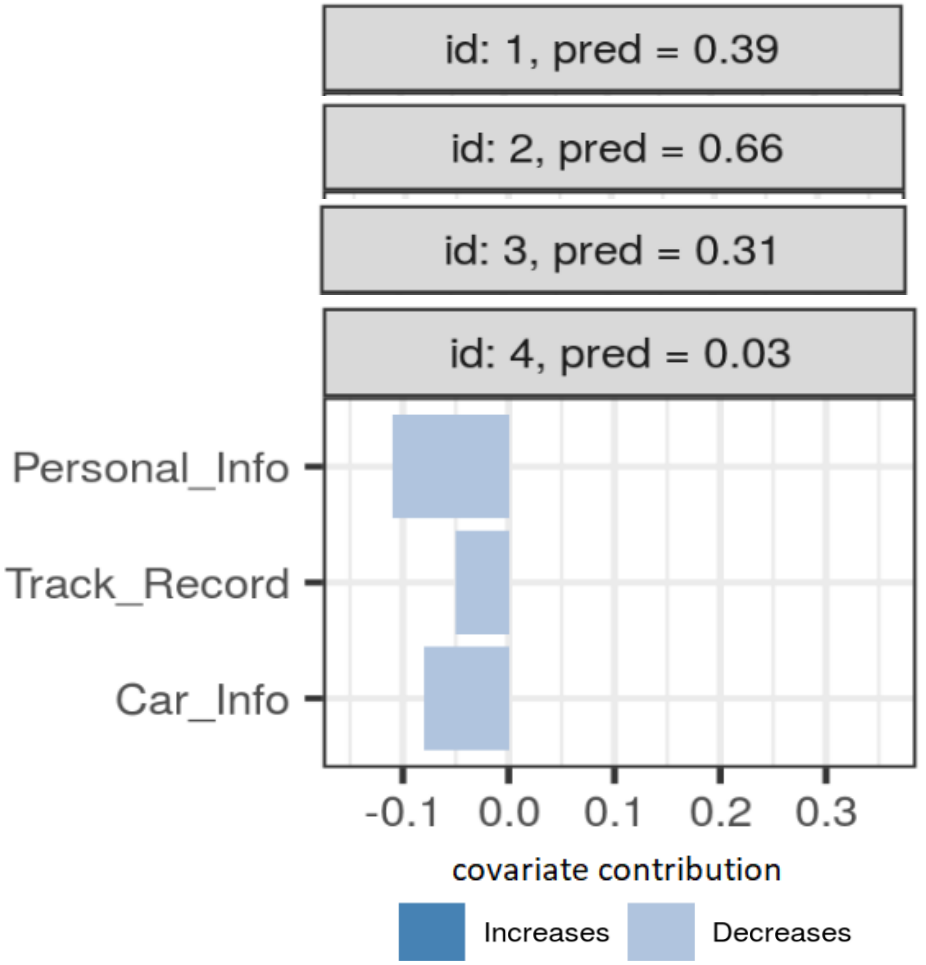
Practical example of groupShapley

- US Car insurance dataset
  - Data from 10 302 customers with records of crash/no crash + claim size, and 23 covariates
  - We concentrate on the binary response crash/no crash and fit a random forest model with 500 trees to model $f(\boldsymbol{x}) = P(\text{customer with covariate } \boldsymbol{x} \text{ crashes})$
  - Group covariates based on type
    - Personal information (13 covariates): age of driver, education level, # children, job type, # driving children, marital status, gender, distance to work +++
    - Track record (4 covariate): # claims last 5 years, # licence record points, previous licence revokes, time as customer
    - Car information (4 covariates) type of car, age of car, type of car, whether car is red

Practical example of groupShapley

- We apply the model to 4 individuals

1. Single mother of 4 (2 driving).
   1 claim last 5 years, 3 licence record points.
   Driving a SUV, 27 miles to work.
2. 37 years old father of 2 (1 driving).
   1 claim last 5 years, got licence revoked and 10 licence record points.
3. 60 year old married male doctor with no children, holding a Phd.
   3 claims last 5 years
   Red sports car.
4. 50 year old female, no kids.
   No claims or license revoked.
   Drives a minivan



Conclusion

- Our contribution
  - Explain individual predictions $f(\boldsymbol{x})$ through Shapley values for covariate groups instead of the single covariates
- Implementation
  - groupShapley is available in the GitHub version (https://github.com/NorskRegnesentral/shapr) of our R-package **shapr**. Will be included in the next CRAN release
- Paper
  - Will be made available on arXiv later this week (https://arxiv.org/search/?searchtype=author&query=Jullum,M)

(jullum@nr.no)