Advanced Machine Learning **Live Session 5**
SoSe 2023

Exercise 1: Factorizing distributions

(a) Can we factorize the joint distribution $\mathbb{P}(x)$ as $\mathbb{P}(x_S)\mathbb{P}(x_{-S})$? How can we factorize the joint distribution, such that the distribution is preserved? Formally prove your answer.

(b) Let $x_S \perp x_{-S}$. Does the factorization preserve the joint now? Formally prove your answer.

(c) Illustrate the two factorizations in a schematic drawing. *Hint:* You can draw a 2D scatterplot with two dependent variables. Given a fixed value for the conditioned variable, draw the range of values that conditional and marginal sampling consider.

Exercise 2: Feature importance and extrapolation

Based on the results in Exercise 1, explain why and when the different feature importance methods extrapolate.

(a) Over which distributions does PFI evaluate the model? Under which assumptions is the model evaluated outside the domain?

(b) Over which distributions does CFI evaluate the model? What about SAGE? Do the methods extrapolate?

(c) For both PFI and CFI evaluate whether/when the perturbed variables are dependent/independent of the target variable.

(d) What does that mean for the interpretation of PFI and CFI?

(e) Can a feature be relevant for CFI but not relevant for PFI?

Exercise 3: In class discussion

Discuss with your neighbor. Which of the aforementioned methods is superior? PFI or the extrapolation-free alternatives?

(a) Which method is most suitable for situations where we aim to understand the model's mechanism? If any?

(b) Which method is most suitable for situations where we want to understand the data generating mechanism?

  (i) In order to find features that are informative of the prediction target?

  (ii) In order to select the smallest possible set of features, which would enable the same prediction performance?

  (iii) In order to find variables that are causal for the prediction target?