What are Shapley values?

Shapley values, from cooperative game theory, tell us how much "credit" should each player in the game get for producing the outcome by assigning credit based on the marginal contribution \( \Delta \) each player makes when joining the group.

Alice, Bob, and Celine are farmers who produce 9 bushels of wheat when working together. Their Shapley values \( \phi \) are their average \( \Delta \) over all \( 3! \) permutations of the group order.

Goals for feature explanation

When interpreting machine learning models, it is important to consider the full process of producing model output [3]:

When interpreting machine learning models, it is important to consider the full process of producing model output [3]:

- **Explaining the model**: understanding why the machine learning model makes a prediction.
- **Explaining the world**: understanding a real-world mechanism through the data and model output.

Cotenability and Causality

- **Features are often correlated because they have a shared latent real-world cause \( Z \): there is a trade-off in breaking or respecting these dependencies.**
- **Cotenability** explanations respect correlations among features: changes in BMI must also change height or weight. **Explains the world.**
- **Model-based causal** explanations tell us how features intervene on the model: how does setting blood pressure to 180 affect predicted mortality? **Explains the model.**

Grouping features increases interpretability and moves closer to satisfying both cotenability and causality. Grouping features increases interpretability and moves closer to satisfying both cotenability and causality.

NHANES mortality case study

(a) The importance of blood pressure is spread across systolic, diastolic, and pulse pressure.

(b) Removing systolic and diastolic increases the importance of pulse pressure.

(c) Grouping all blood pressure features increases their relative importance.

References