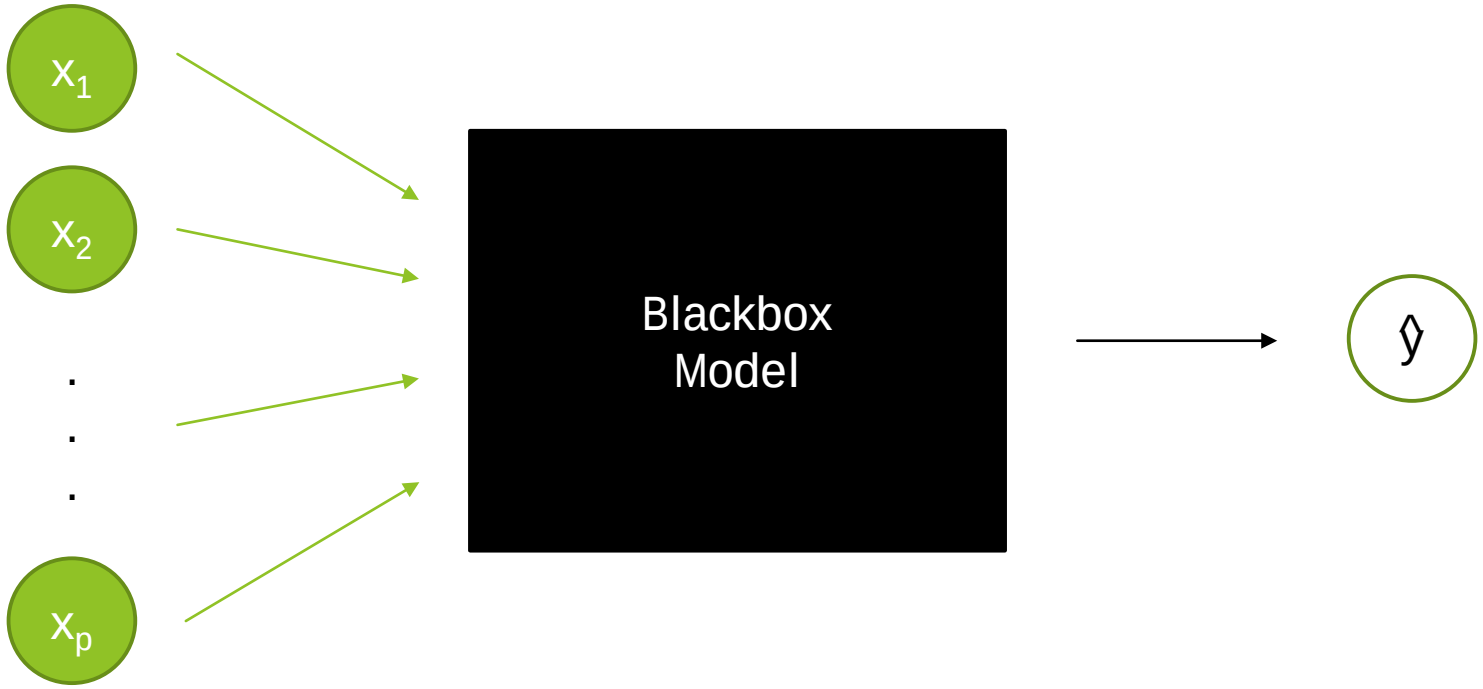
Decoding the Blackbox: Interpretable Machine Learning

Alexander Gerharz

„Statistical Methods for Big Data“

Concept for Interpretable Machine Learning



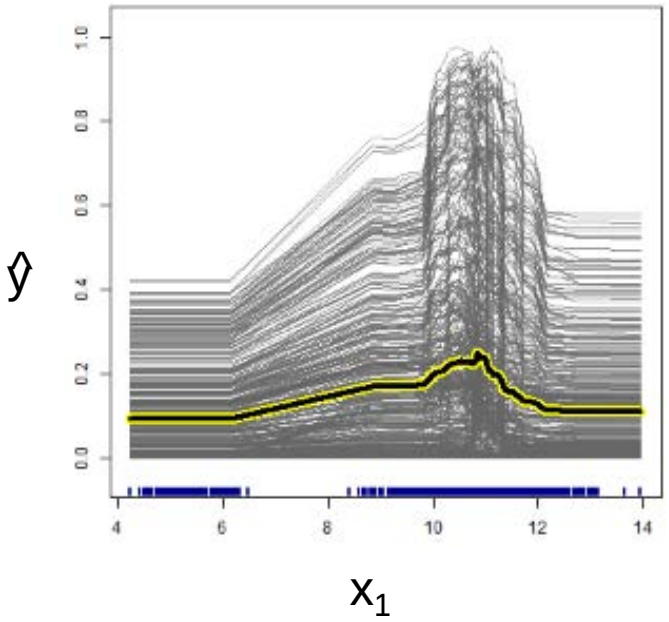
Determine a feature/some features of interest (foi)

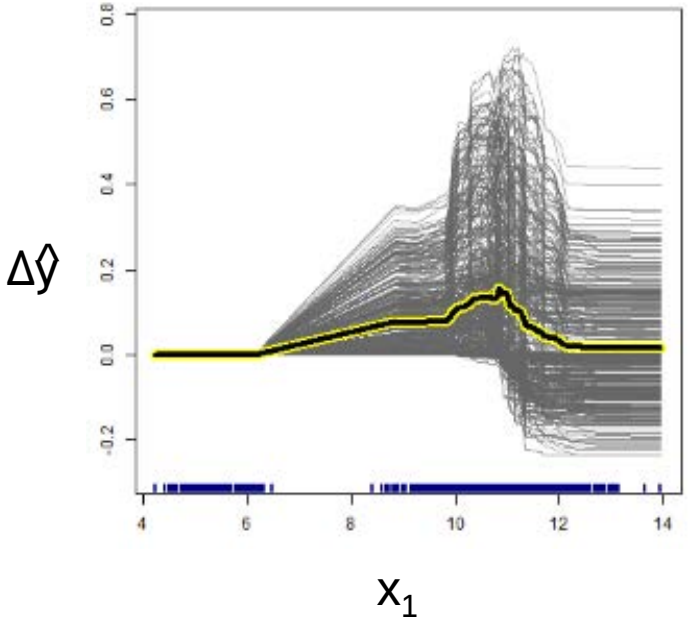
Manipulate foi

Observe the output-changes

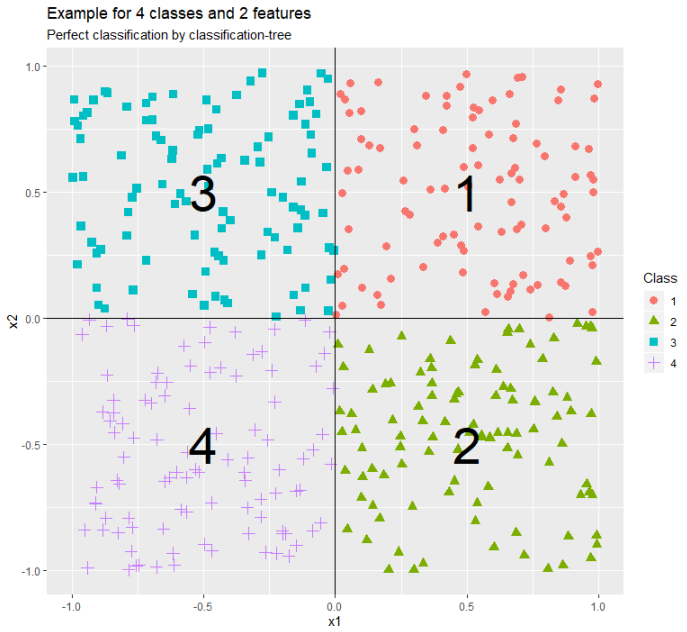
Individual Conditional Expectation Plot

- Determine one feature of interest
- Use real covariable settings from observed data
- Observe the prediction changes while testing the feature of interest on its whole range of values





Chordgraphs used for Neighborhoods in Models



1. Artifically raise x_1

2. Observe changes in the classification