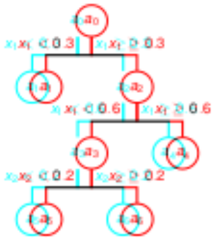
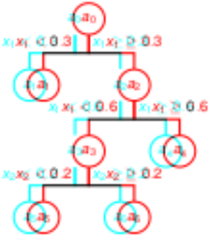
ADVANTAGES

- For inherently interpretable models some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error



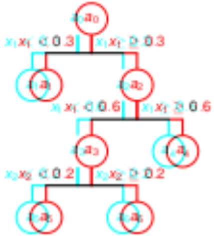
ADVANTAGES

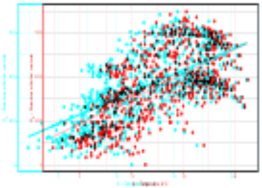
- For inherently interpretable models some additional model-agnostic interpretation methods not required
  ⇝ Eliminates a source of error
- Interpretable models often simple
  ⇝ training time is fairly small



ADVANTAGES

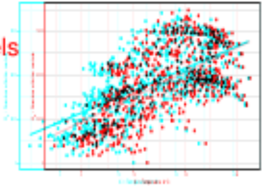
- For inherently interpretable models some additional model-agnostic interpretation methods not required
  $\rightsquigarrow$ Eliminates a source of error
- Interpretable models often simple
  $\rightsquigarrow$ training time is fairly small
- Some interpretable models estimate monotonic effects
  $\rightsquigarrow$ Simple to explain as larger feature values always lead to higher (or smaller) outcomes (e.g. GLMs)





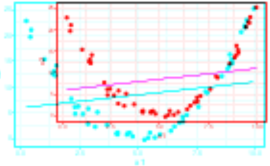
ADVANTAGES

- For inherently interpretable models some additional model-agnostic interpretation methods not required
  ⇝ Eliminates a source of error
- Interpretable models often simple
  ⇝ training time is fairly small
- Some interpretable models estimate monotonic effects
  ⇝ Simple to explain as larger feature values always lead to higher (or smaller) outcomes (e.g., GLMs)
- Many people are familiar with simple interpretable models
  ⇝ Increases trust, facilitates communication of results



DISADVANTAGES

- Often require assumptions about data / model structure
  $\rightsquigarrow$ If assumptions are wrong, models may perform bad



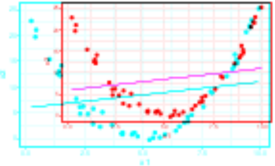
DISADVANTAGES

- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
  - Linear model with lots of features and interactions
  - Decision trees with huge tree depth

- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad

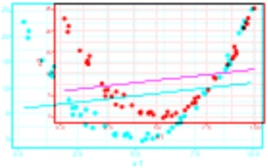
- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions

  - Decision trees with huge tree depth



DISADVANTAGES

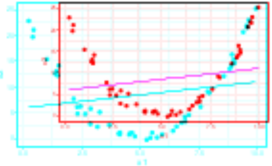
- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
  - Linear model with lots of features and interactions
  - Decision trees with huge tree depth
- Often do not automatically model complex relationships due to limited flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM



- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad
- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions
  - Decision trees with huge tree depth
- Often not able to automatically model complex relationships due to limited model flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM

DISADVANTAGES

- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad
- Interpretable models may also be hard to interpret, e.g.:
  - Linear model with lots of features and interactions
  - Decision trees with huge tree depth
- Often do not automatically model complex relationships due to limited flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM
- Inherently interpretable models do not provide all types of explanations
  ⇝ Methods providing other types of explanations still useful (e.g., counterfactual explanations)



- Often require assumptions about data / model structure
  ⇝ If assumptions are wrong, models may perform bad
- Interpretable models may also be difficult to interpret, e.g.:
  - Linear model with hundreds of features and interactions
  - Decision trees with huge tree depth
- Often not able to automatically model complex relationships due to limited model flexibility
  e.g., high-order main or interaction effects need to be specified manually in a LM
- Inherently interpretable models do not provide all types of explanations
  ⇝ Methods providing other types of explanations still useful (e.g., counterfactual explanations)

FURTHER COMMENTS

- Some argue that interpretable models should be preferred ▸ Rudin 2019
  - … instead of explaining uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing or manual feature engineering

FURTHER COMMENTS

- Some argue that interpretable models should be preferred (Rudin 2019)
  - …instead of explaining uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing or manual feature engineering

⇝ Drawback: Hard to achieve for data for which end-to-end learning is crucial

⇝ e.g., hard to extract good features for image / text data

⇝ information loss = bad performance

FURTHER COMMENTS

- Some argue that interpretable models should be preferred in the first place ▸ Rudin 2019
  - … instead of explaining uninterpretable models post-hoc
  - Can sometimes work out by spending enough time and energy on data pre-processing or manual feature engineering

⇝ Drawback: Hard to achieve for data for which end-to-end learning is crucial

⇝ e.g., hard to extract good features for image / text data
(⇝ information loss = bad performance)

- Often there is a trade-off between interpretability and model performance