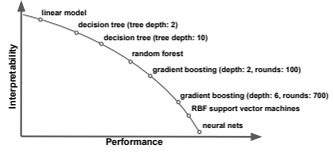
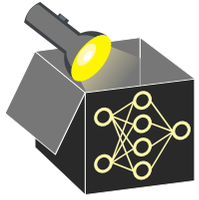
Interpretable Machine Learning

Introduction, Motivation, and History

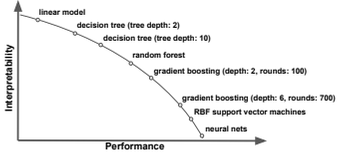


Learning goals

- Why interpretability?
- Developments until now?
- Use cases for interpretability

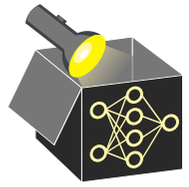


Interpretable Machine Learning
Introduction, Motivation, and History



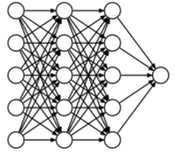
Learning goals

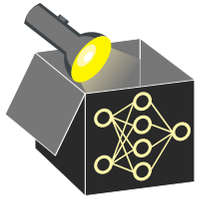
- Why interpretability?
- Developments until now?
- Use cases for interpretability



WHY INTERPRETABILITY?

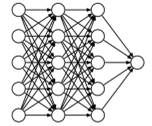
- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  $\rightsquigarrow$ too complex to be understood by humans
- Some applications are "learn to understand"

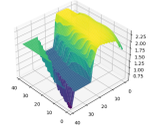


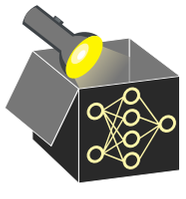


WHY INTERPRETABILITY?

- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  $\rightsquigarrow$ too complex to be understood by humans
- Some applications are "learn to understand"



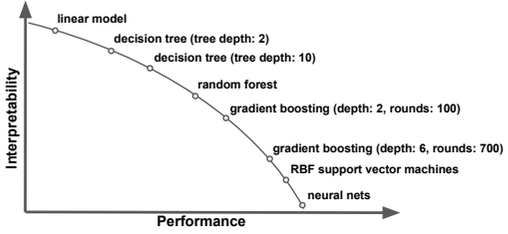




WHY INTERPRETABILITY?

- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"
- When deploying ML models, lack of explanations
  1. hurts trust
  2. creates barriers

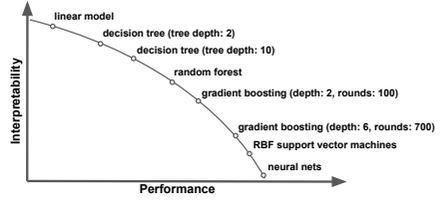
⇝ Many disciplines with required trust rely on traditional models, e.g., linear models, with less predictive performance



WHY INTERPRETABILITY?

- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"
- When deploying ML models, lack of explanations
  1. hurts trust
  2. creates barriers

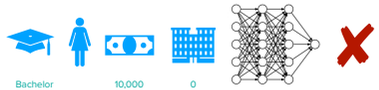
⇝ Many disciplines with required trust rely on traditional models, e.g., linear models, with less predictive performance

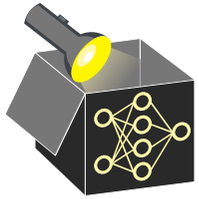


INTERPRETABILITY IN HIGH-STAKES DECISIONS

Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications
  ▸ Society of Actuaries
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

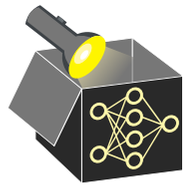




INTERPRETABILITY IN HIGH-STAKES DECISIONS

Examples of critical areas where decisions based on ML models can affect human life

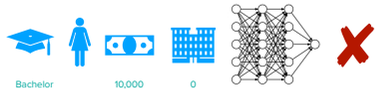
- Credit scoring and insurance applications ▸ Click for source
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

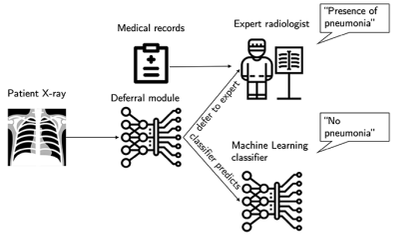


INTERPRETABILITY IN HIGH-STAKES DECISIONS

Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications
  ▸ Society of Actuaries
  - Reasons for not granting a loan
  - Fraud detection in insurance claims
- Medical applications
  - Identification of diseases
  - Recommendations of treatments
- ...



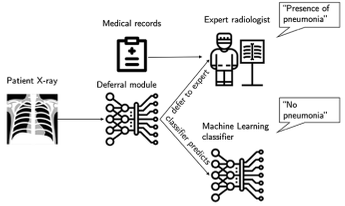


▸ Miliard (2020)

INTERPRETABILITY IN HIGH-STAKES DECISIONS

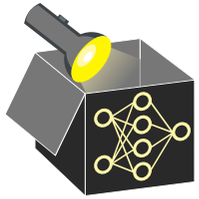
Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications ▸ Click for source
  - Reasons for not granting a loan
  - Fraud detection in insurance claims
- Medical applications
  - Identification of diseases
  - Recommendations of treatments
- ...



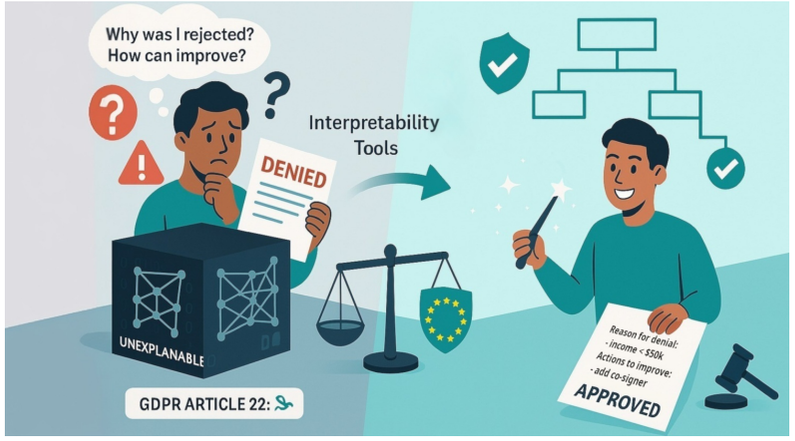
Miliard (2020) ▸ Click for source

NEED FOR INTERPRETABILITY

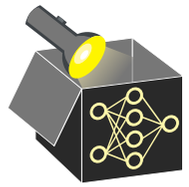


Need for interpretability becoming increasingly important from a legal perspective

- General Data Protection Regulation (GDPR) requires for some applications that models have to be explainable ▸ Goodman & Flaxman (2017)
  ⇝ *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation"*
- *Ethics guidelines for trustworthy AI* ▸ European Commission (2019)

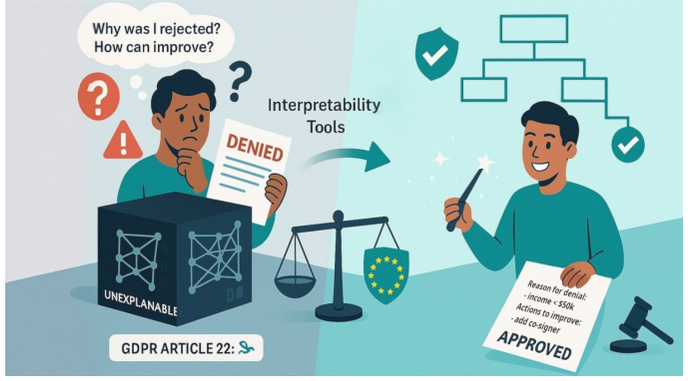


NEED FOR INTERPRETABILITY



Need for interpretability becoming increasingly important from a legal perspective

- General Data Protection Regulation (GDPR) requires for some applications that models have to be explainable ▸ Flaxman 2017
  ⇝ *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation"*
- *Ethics guidelines for trustworthy AI* ▸ "European Commission" 2019

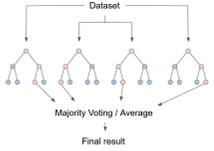


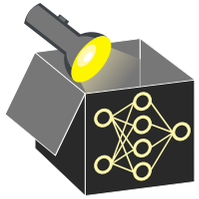
BRIEF HISTORY OF INTERPRETABILITY

- 18th and 19th century:
  Linear regression models (Gauss, Legendre, Quetelet)
- 1940s:
  Emergence of sensitivity analysis (SA)
- Middle of 20th century:
  Rule-based ML, incl. decision rules and decision trees
- 2001:
  Built-in feature importance measure of random forests
- >2010:
  Explainable AI (XAI) for deep learning
- >2015:
  IML as an independent field of research

▸ Carl Friedrich Gauss

▸ Wikipedia





BRIEF HISTORY OF INTERPRETABILITY

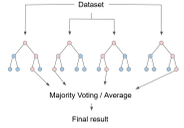
- 18th and 19th century:
  Lin. regression models (Gauss, Legendre, Quetelet)
- 1940s:
  Emergence of sensitivity analysis (SA)
- Middle of 20th century:
  Rule-based ML, incl. decision rules and trees
- 2001:
  Built-in feature imp. measure of random forests
- >2010:
  Explainable AI (XAI) for deep learning
- >2015:
  IML as an independent field of research

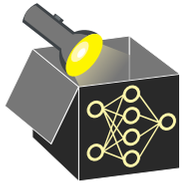
Carl Friedrich Gauss

▸ Click for source

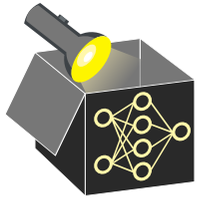
Wikipedia

▸ Click for source

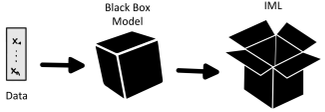




Interpretable Machine Learning



Interpretation Goals



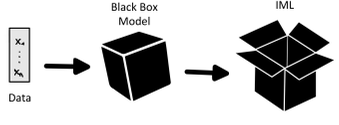
Learning goals

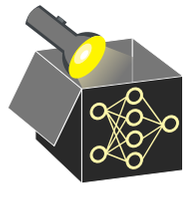
Understand Interpretation Goals:

- Global insights (discovery)
- Improve model (debug and audit)
- Understand and control individual predictions
- Justification and fairness

Interpretable Machine Learning

Interpretation Goals



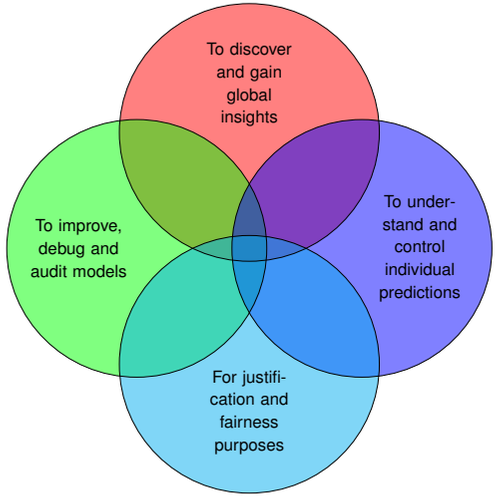


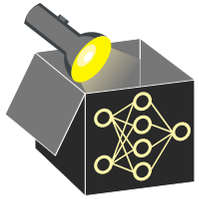
Learning goals

Understand Interpretation Goals:

- Global insights (discovery)
- Improve model (debug and audit)
- Understand and control individual predictions
- Justification and fairness

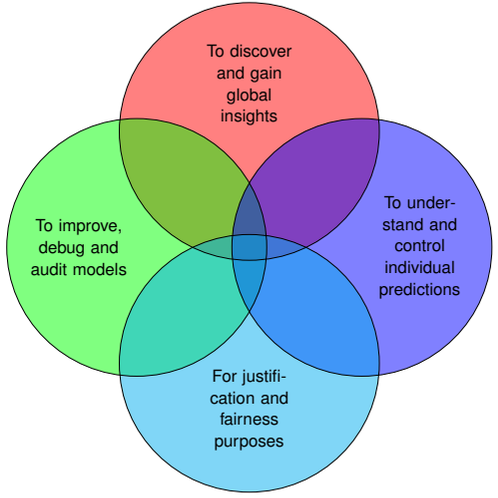
POTENTIAL INTERPRETATION GOALS

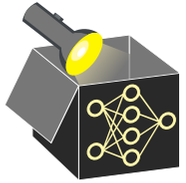




A related presentation can be found in ▸ Adadi and Berrada 2018.

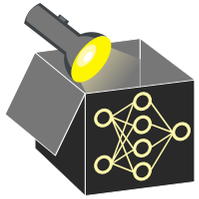
POTENTIAL INTERPRETATION GOALS





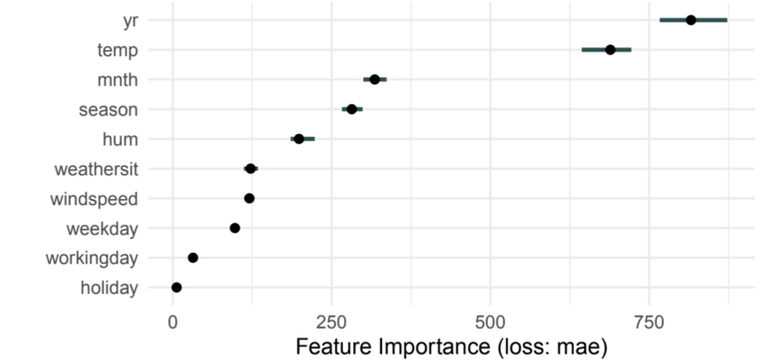
A related presentation can be found in ▸ Adadi and Berrada 2018.

DISCOVER AND GAIN GLOBAL INSIGHTS



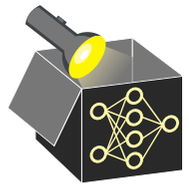
⇝ Gain insights about data, model, and underlying data-generating process

Example: Bike Sharing Dataset (predict number of bike rentals per day)
Exemplary question: Which feature influences model performance and how much?



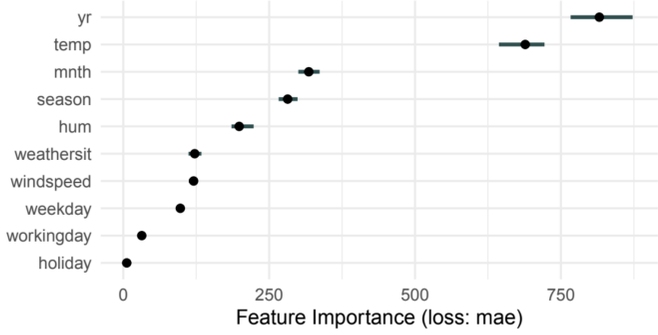
- Year (`yr`) and Temperature (temp) most important features
- Holiday (`holiday`) less important (Can we drop it?)

DISCOVER AND GAIN GLOBAL INSIGHTS



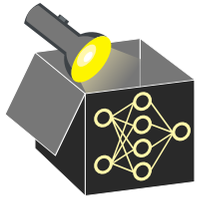
⇝ Gain insights about data, model, and underlying data-generating process

Example: Bike Sharing Dataset (predict number of bike rentals per day)
Exemplary question:
Which feature influences model performance and how much?



- Year (`yr`) and Temperature (temp) most important features
- Holiday (`holiday`) less important (Can we drop it?)

IMPROVE, DEBUG AND AUDIT MODELS



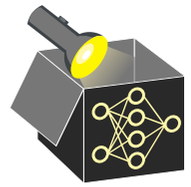
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure

IMPROVE, DEBUG AND AUDIT MODELS



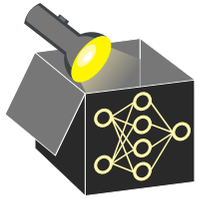
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank (gwern.net) ▸ Click for source

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure

IMPROVE, DEBUG AND AUDIT MODELS



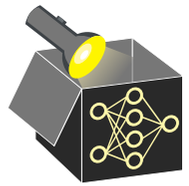
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels

IMPROVE, DEBUG AND AUDIT MODELS



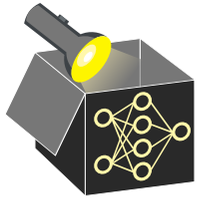
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank (gwern.net) ▸ Click for source

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels

IMPROVE, DEBUG AND AUDIT MODELS



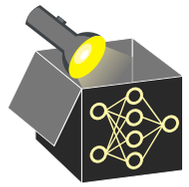
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels
- E.g. model detects weather based on sky:
  ⇝ All photos with tanks show cloudy sky
  ⇝ Photos without tanks show sunny sky

IMPROVE, DEBUG AND AUDIT MODELS



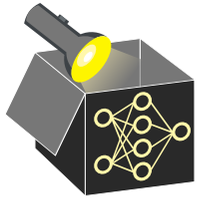
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank (gwern.net) ▸ Click for source

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels
- E.g. model detects weather based on sky:
  ⇝ All photos with tanks show cloudy sky
  ⇝ Photos without tanks show sunny sky

IMPROVE, DEBUG AND AUDIT MODELS



⇝ Insights help to identify flaws (in data or model), which can be corrected

Comment on tank example:

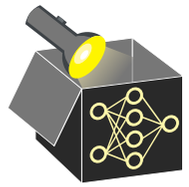
"We made exactly the same mistake in one of my projects on insect recognition. We photographed 54 classes of insects. Specimens had been collected, identified, and placed in vials. Vials were placed in boxes sorted by class. I hired student workers to photograph the specimens.

Naturally they ***did this one box at a time; hence, one class at a time.*** *Photos were taken in alcohol.* ***Bubbles would form in the alcohol. Different bubbles on different days.*** *The learned classifier was surprisingly good. But a* ***saliency map revealed that it was reading the bubble patterns*** *and ignoring the specimens.*

I was so embarrassed that I had made the oldest mistake in the book (even if it was apocryphal). Unbelievable. Lesson: always randomize even if you don't know what you are controlling for!"

▸ Thomas G. Dietterich

IMPROVE, DEBUG AND AUDIT MODELS



⇝ Insights help to identify flaws (in data or model), which can be corrected

Comment on tank example:

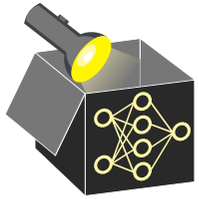
"We made exactly the same mistake in one of my projects on insect recognition. We photographed 54 classes of insects. Specimens had been collected, identified, and placed in vials. Vials were placed in boxes sorted by class. I hired student workers to photograph the specimens.

Naturally they ***did this one box at a time; hence, one class at a time.*** *Photos were taken in alcohol.* ***Bubbles would form in the alcohol. Different bubbles on different days.*** *The learned classifier was surprisingly good. But a* ***saliency map revealed that it was reading the bubble patterns*** *and ignoring the specimens.*

I was so embarrassed that I had made the oldest mistake in the book (even if it was apocryphal). Unbelievable. Lesson: always randomize even if you don't know what you are controlling for!"

(Thomas G. Dietterich) ▸ Click for source

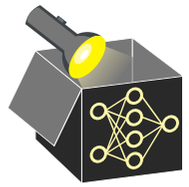
DEBUG AND AUDIT



- Nearly all computer programs have bugs
  - ⇝ Minimizing such bugs extremely relevant
- Process with multiple steps to locate, understand and solve a problem
  - ⇝ Classical debugging
- **In ML** we have a program (learner) writing another program (model)
- How to debug or audit programs which contain ML models?
- Based on a single cross-val score?

⇝ Being able to interpret your model will always be helpful – if possible!

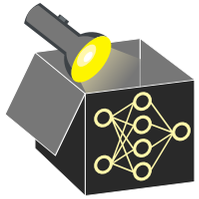
DEBUG AND AUDIT



- Nearly all computer programs have bugs
  - ⇝ Minimizing such bugs extremely relevant
- Process with multiple steps to locate, understand and solve a problem
  - ⇝ Classical debugging
- **In ML** we have a program (learner) writing another program (model)
- How to debug or audit programs which contain ML models?
- Based on a single cross-val score?

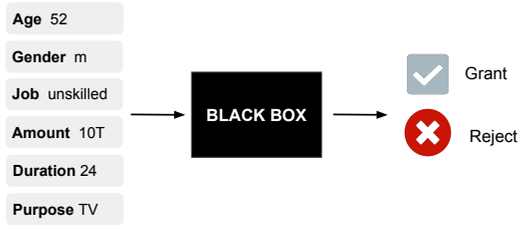
⇝ Being able to interpret your model will always be helpful – if possible!

UNDERSTAND & CONTROL INDIVIDUAL PREDICTIONS



⇝ Explaining individual decisions can prevent unwanted actions based on the model

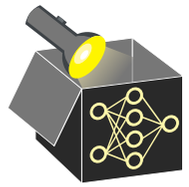
Example: Credit Risk Application
x: customer and credit information; *y*: grant or reject credit



Questions:

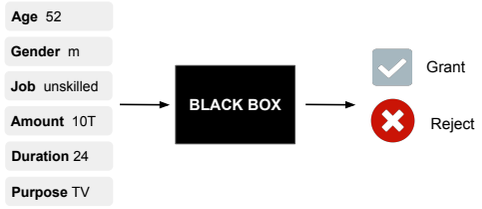
- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

UNDERSTAND & CONTROL INDIVIDUAL PRED.-S



⇝ Explaining individual decisions can prevent unwanted model-based actions

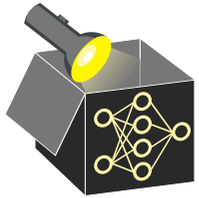
Example: Credit Risk Application
x: customer and credit information; *y*: grant or reject credit



Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should be changed so that the credit is accepted?**

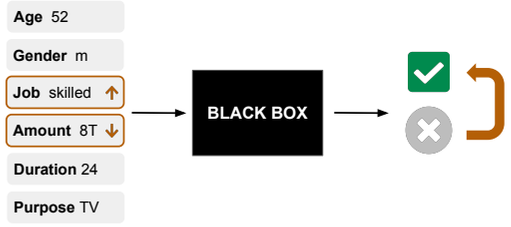
UNDERSTAND & CONTROL INDIVIDUAL PREDICTIONS



⇝ Explaining individual decisions can prevent unwanted actions based on the model

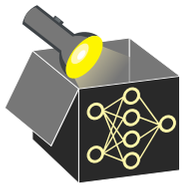
Example: Credit Risk Application
x: customer and credit information; *y*: grant or reject credit

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**



"If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

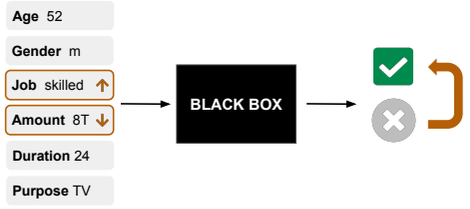
UNDERSTAND & CONTROL INDIVIDUAL PRED.-S



⇝ Explaining individual decisions can prevent unwanted model-based actions

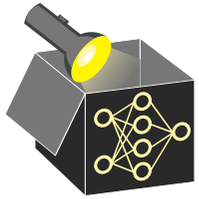
Example: Credit Risk Application
x: customer and credit information; *y*: grant or reject credit

- Why was the credit rejected?
- Is it a fair decision?
- **How should be changed so that the credit is accepted?**



"If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

JUSTIFICATION AND FAIRNESS

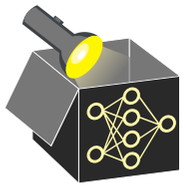


$\rightsquigarrow$ Investigate if and why biased, unexpected or discriminatory predictions were made

Example: COMPAS

- COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
- Commercial tool used in courts to assess a defendant's risk of re-offending
- Predicts **recidivism risk**:
  - Likelihood of an individual with a past offense is arrested again
  - Features: race, gender, age, number of prior prison sentences, ...
  - Output: COMPAS score from 1 (low risk) to 10 (high risk) risk of recidivism
- Based on a questionnaire completed by the defendant

JUSTIFICATION AND FAIRNESS



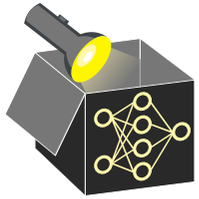
$\rightsquigarrow$ Investigate if and why biased, unexpected or discriminatory predictions were made

Example: COMPAS

- COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
- Commercial tool used in courts to assess a defendant's risk of re-offending
- Predicts **recidivism risk**:
  - Likelihood of an individual with a past offense is arrested again
  - Features: race, gender, age, number of prior prison sentences, ...
  - Output: COMPAS score from 1 (low) to 10 (high) risk of recidivism
- Based on a questionnaire completed by the defendant

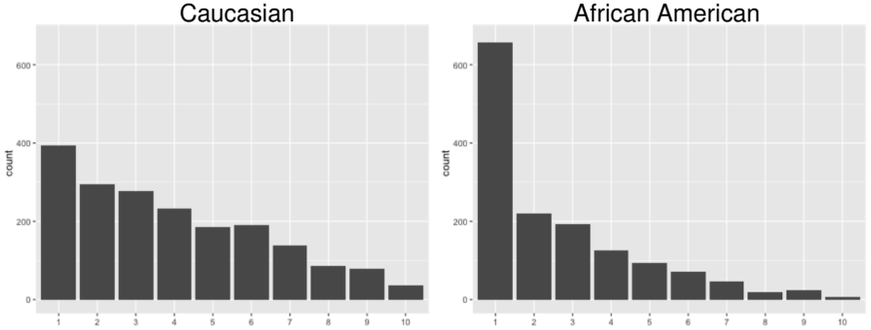
JUSTIFICATION AND FAIRNESS: COMPAS

Larson et al. 2016



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

Descriptive data analysis of the target (COMPAS score) by a feature encoding race:

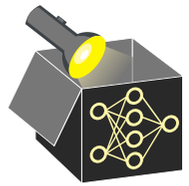


COMPAS score: 1 (low risk) to 10 (high risk)

⇝ Model skewed towards low risk for Caucasians
⇝ Strong indication that the model is discriminating against African American
⇝ Use IML to investigate if and how much the model uses the defendant's race

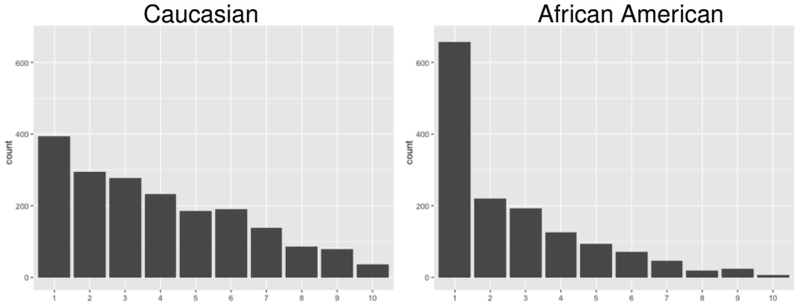
JUSTIFICATION AND FAIRNESS: COMPAS

LARSON



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

Descriptive data analysis of the target (COMPAS score) by a feature encoding race:

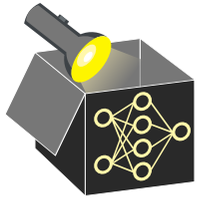


COMPAS score: 1 (low risk) to 10 (high risk)

⇝ Model skewed towards low risk for Caucasians
⇝ Strong indication that the model is discriminating against African American
⇝ Use IML to assess if and how much the model uses the defendant's race

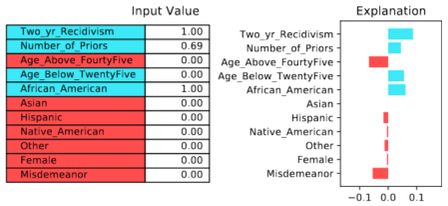
JUSTIFICATION AND FAIRNESS: COMPAS

▸ Alvarez-Melis and Jaakkola 2018



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

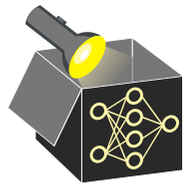
IML: Analyze how strongly a feature influences an individual prediction (e.g., LIME):



- Pick a defendant
- LIME quantifies a feature's impact on the defendant's COMPAS score
- `African_American` has a large positive weight on COMPAS score
- Occurs for many individuals, see ▸ XAI Stories ⇝ Suggests racial bias

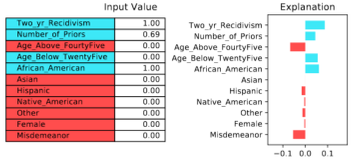
JUSTIFICATION AND FAIRNESS: COMPAS

▸ ALVAREZ_MELIS_JAAKKOLA



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

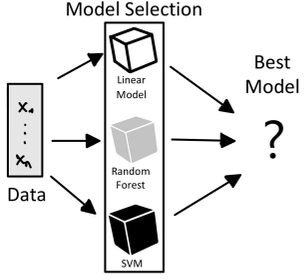
IML: Analyze how strongly a feature influences an individual pred. (e.g., LIME):



- Pick a defendant
- LIME quantifies a feature's impact on the defendant's COMPAS score
- `African_American` has a large positive weight on COMPAS score
- Occurs for many individuals, see "XAI Stories" ▸ Click for source ⇝ Suggests racial bias

Interpretable Machine Learning

Dimensions of Interpretability

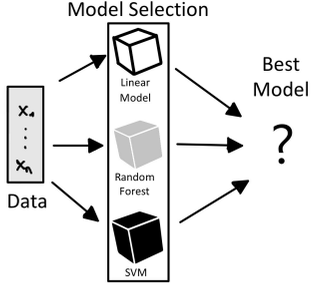


Learning goals

- Difference between intrinsic, model-specific, and model-agnostic interpretability
- Different types of explanations
- Local, global, and regional explanations
- Model/learner explanation (without/with refits)
- Levels of interpretability

Interpretable Machine Learning

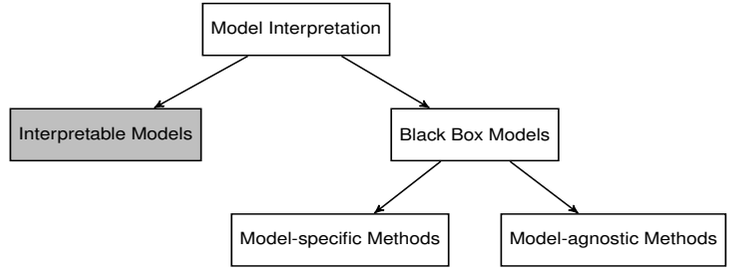
Dimensions of Interpretability

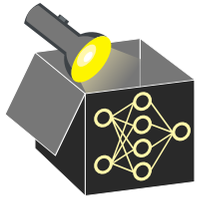


Learning goals

- Difference between intrinsic, model-specific, and model-agnostic interpretability
- Different types of explanations
- Local, global, and regional explanations
- Model/learner explanation (with(out) refits)
- Levels of interpretability

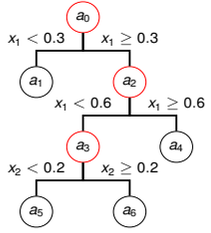
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



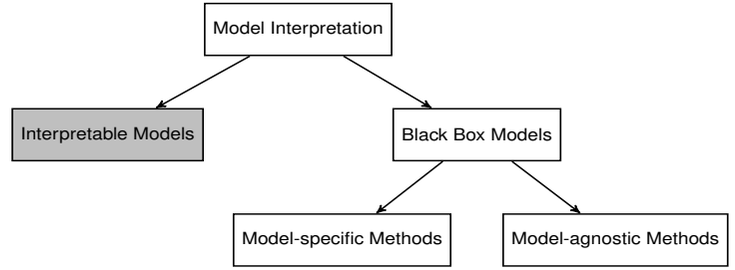


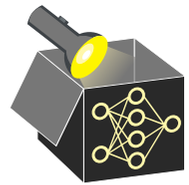
Intrinsically Interpretable Models:

- Simple model structure (e.g., weighted sum or tree)
- Examples: GLMs, decision trees
- Pro: Additional IML methods not necessarily required
- Con: Limited model complexity can reduce performance; can still be hard to interpret with many features /interactions



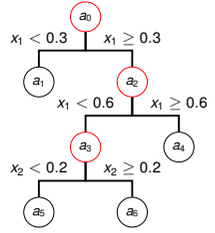
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



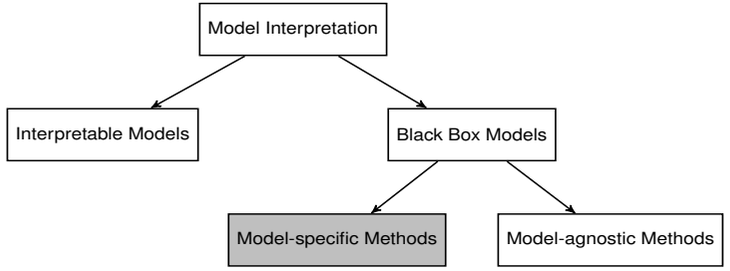


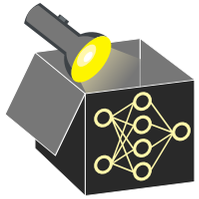
Intrinsically Interpretable Models:

- Simple model structure (e.g., weighted sum or tree)
- Examples: GLMs, decision trees
- Pro: Additional IML methods not necessarily required
- Con:
  Limited model complexity can reduce performance, can still be hard to interpret (many features/interactions)



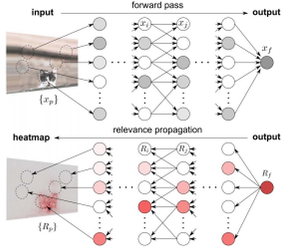
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



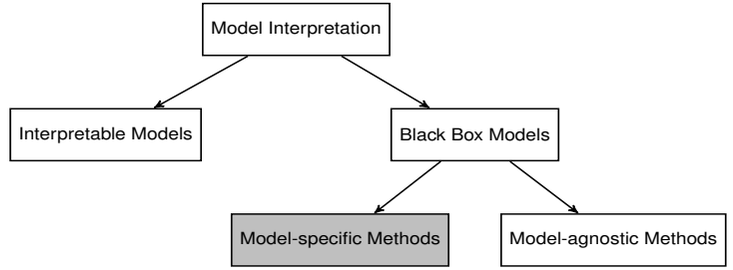


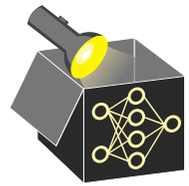
Model-specific Methods:

- Designed for specific model types (e.g., NNs)
- Examples: Gini importance of tree-based models, Layer-wise relevance propagation (LRP)
- Pro: Exploit model structure
- Con: Restricted to specific model class



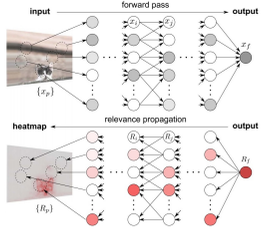
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



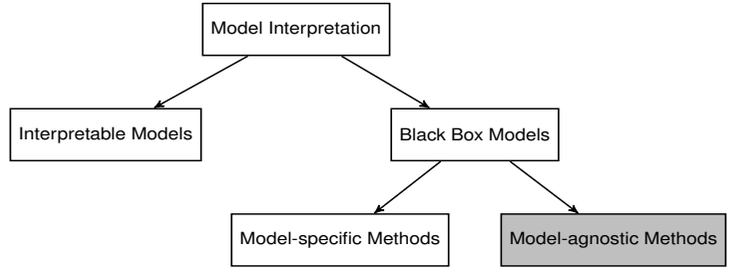


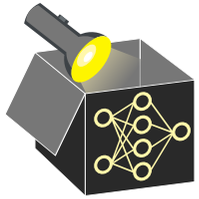
Model-specific Methods:

- Designed for specific model types (e.g., NNs)
- Examples:
  Gini importance of tree-based models,
  Layer-wise relevance propagation (LRP)
- Pro: Exploit model structure
- Con: Restricted to specific model class



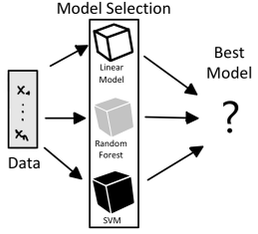
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



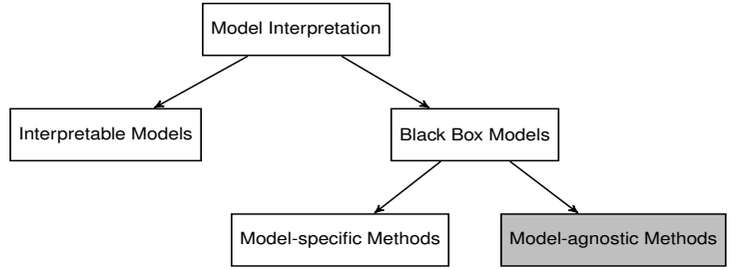


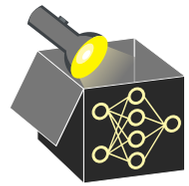
Model-agnostic Methods:

- In ML: Tune over many model classes
  ⇝ Unknown which model is best / deployed
  ⇝ Need for IML methods that work for any model
- Applied after training (post-hoc)
- Applicable to intrinsically interpretable models
  ⇝ provides insights into explanations



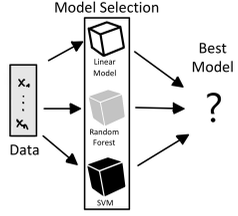
INTRINSIC, MODEL-SPECIFIC, MODEL-AGNOSTIC



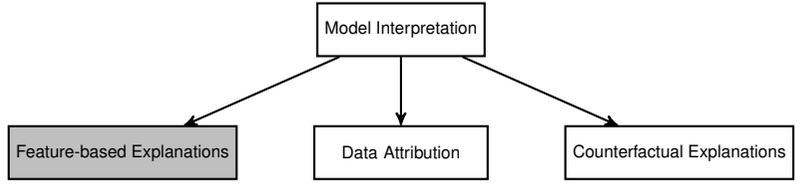


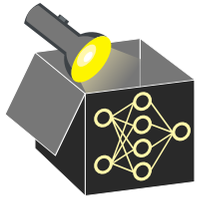
Model-agnostic Methods:

- In ML: Tune over many model classes
  ⇝ Unknown which model is best / deployed
  ⇝ Need for IML methods that work for any model
- Applied after training (post-hoc)
- Applicable to intrinsically interpretable models
  ⇝ provides insights into explanations



TYPES OF EXPLANATIONS

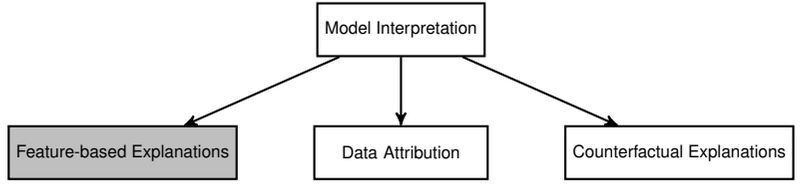


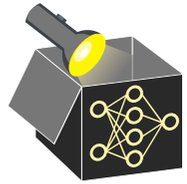


Feature-based Explanations:

- Analyze the role of individual features in model behavior.
- Types of feature-based explanations:
  - Feature Importance
  - Feature Effects
  - Feature Interactions
- Common principle: Vary or perturb feature values and observe changes in predictions, variance, or performance.

TYPES OF EXPLANATIONS

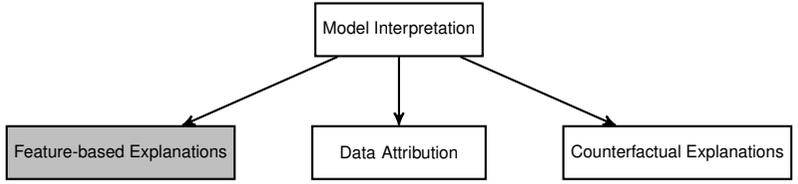


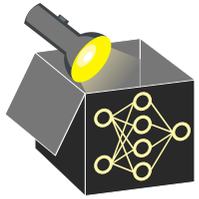


Feature-based Explanations:

- Analyze the role of individual features in model behavior.
- Types of feature-based explanations:
  - Feature Importance
  - Feature Effects
  - Feature Interactions
- Common principle: Vary or perturb feature values and observe changes in predictions, variance, or performance.

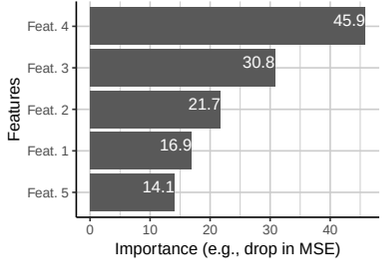
TYPES OF EXPLANATIONS



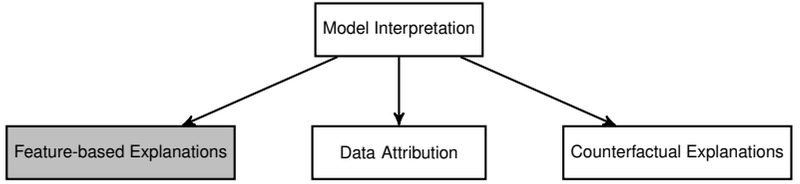


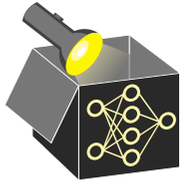
Feature Importance quantifies relevance of features, e.g., their contribution to model prediction, predictive performance, or prediction variance.

- Model-agnostic methods: PFI, . . .
- Pendant in linear models: t-statistic, p-value (significant effect)



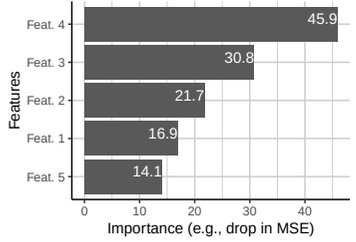
TYPES OF EXPLANATIONS



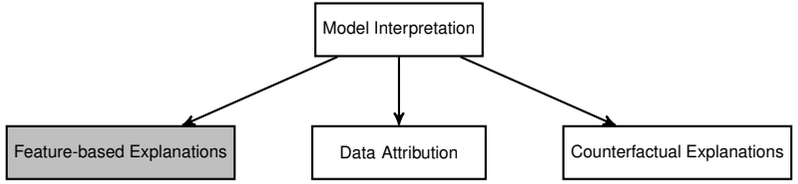


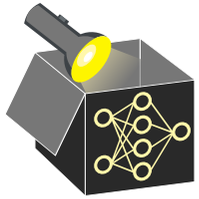
Feature Importance quantifies relevance of features, e.g., their contribution to model prediction, predictive performance, or prediction variance.

- Model-agnostic methods: PFI, . . .
- Pendant in linear models: t-statistic, p-value (significant effect)



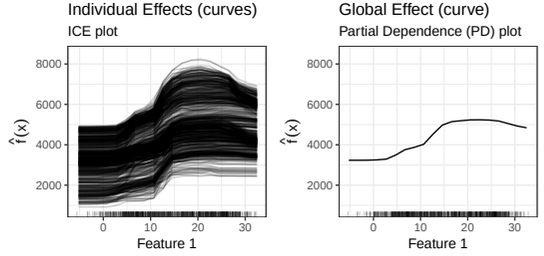
TYPES OF EXPLANATIONS



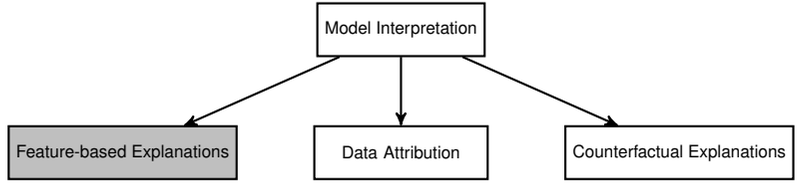


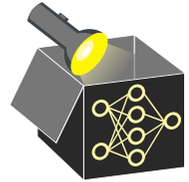
Feature Effects indicate changes (direction and magnitude) in model prediction due to changes in feature values.

- Model-agnostic methods: ICE curves, PD plots . . .
- Pendant in linear models: Weights / coefficients θ_j
- Further examples: ALE, SHAP, and LIME



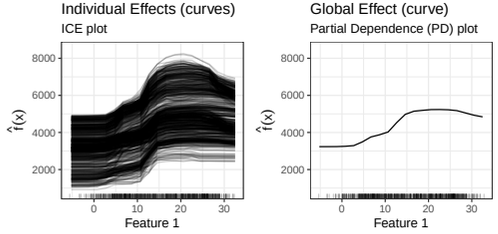
TYPES OF EXPLANATIONS



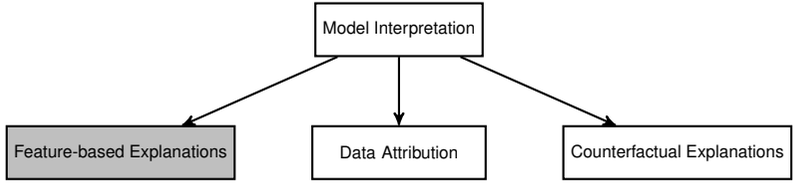


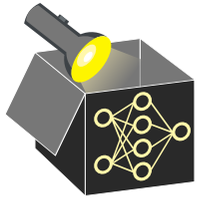
Feature Effects indicate changes (direction and magnitude) in model prediction due to changes in feature values.

- Model-agnostic methods: ICE curves, PD plots . . .
- Pendant in linear models: Weights / coefficients θ_j
- Further examples: ALE, SHAP, and LIME



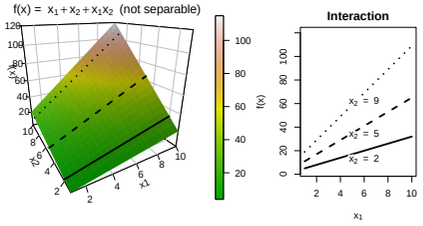
TYPES OF EXPLANATIONS



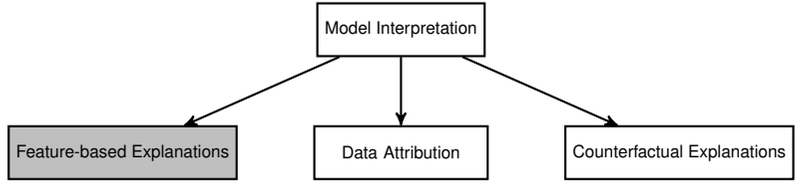


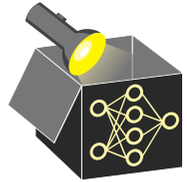
Feature Interaction: How combinations of features jointly affect predictions.

- Model-agnostic methods: Friedman's H-statistic
- Pendant in linear models: Coefficients of interaction terms θ_{jk}



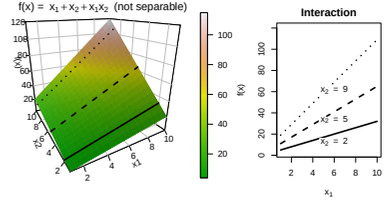
TYPES OF EXPLANATIONS



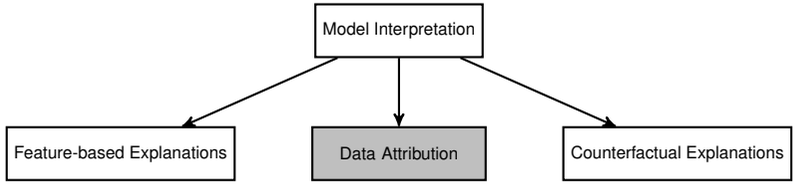


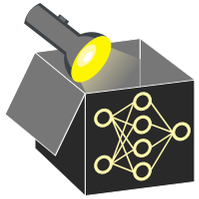
Feature Interaction:
How combinations of features jointly affect predictions.

- Model-agnostic methods: Friedman's H-statistic
- Pendant in linear models: Coefficients of interaction terms θ_{jk}



TYPES OF EXPLANATIONS

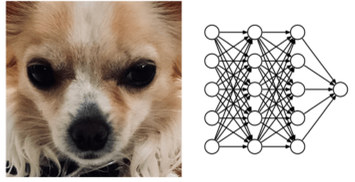




Data Attribution: Identify which training instances most influenced a prediction.

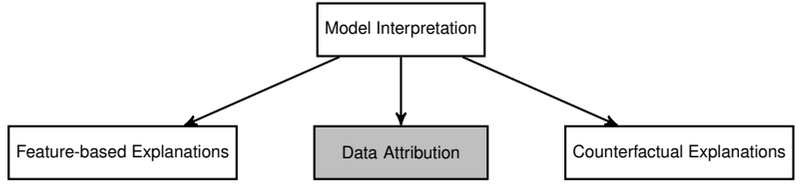
Example: A model should distinguish muffins and dogs.

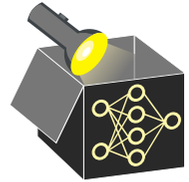
Question: Why does it misclassify this dog image (test point) as a muffin?



Muffin

TYPES OF EXPLANATIONS

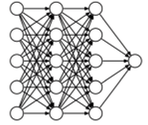




Data Attribution: Identify training instances that most influenced a prediction.

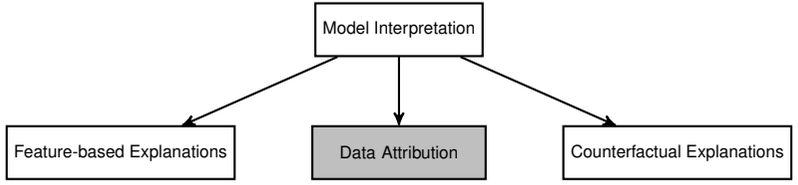
Example: A model should distinguish muffins and dogs.

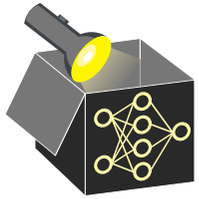
Question: Why does it misclassify this dog image (test point) as a muffin?



Muffin

TYPES OF EXPLANATIONS





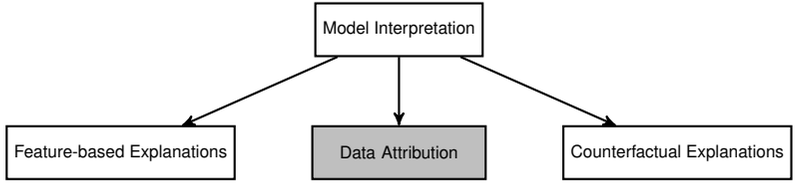
Data Attribution: Identify which training instances most influenced a prediction.

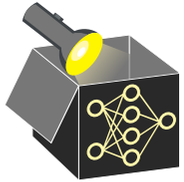
Example: A model should distinguish muffins and dogs.

Approach: Measure how perturbations to training instances affect prediction or loss.

- ⇝ Influential training instances drive prediction of test points.
- ⇝ If these resemble muffins, the model may predict muffin instead of dog.

TYPES OF EXPLANATIONS





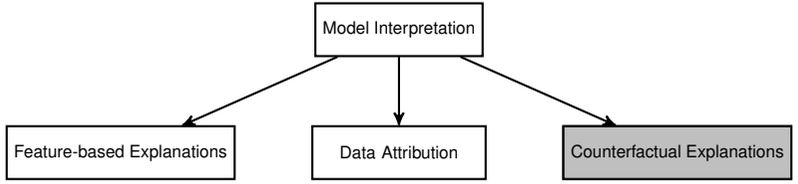
Data Attribution: Identify training instances that most influenced a prediction.

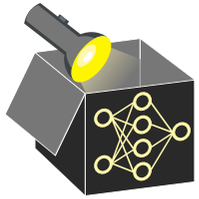
Example: A model should distinguish muffins and dogs.

Approach: Measure how perturbations to training data affect prediction/loss.

- ⇝ Influential training instances drive prediction of test points.
- ⇝ If these resemble muffins, the model may predict muffin instead of dog.

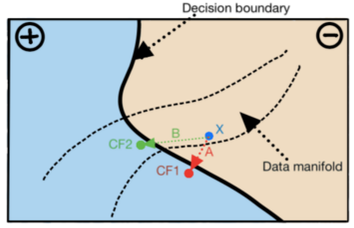
TYPES OF EXPLANATIONS



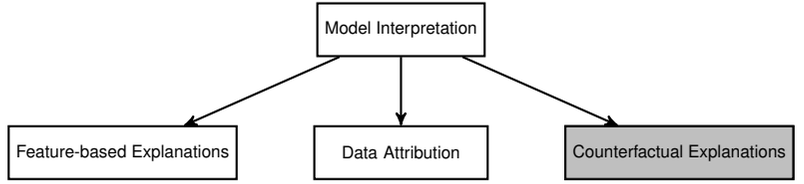


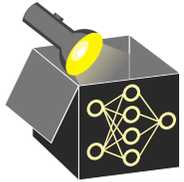
Counterfactual Explanations:

- Identify smallest necessary change in feature values so that a desired outcome is predicted
- Contrastive explanations
- Diverse counterfactuals
- Feasible & actionable explanations



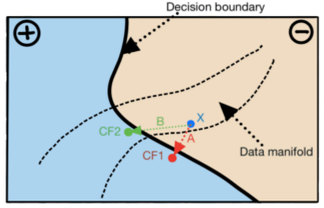
TYPES OF EXPLANATIONS



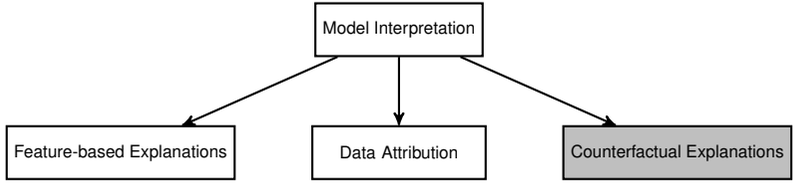


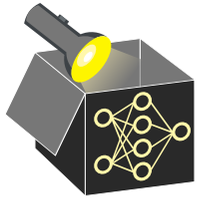
Counterfactual Explanations:

- Identify smallest necessary change in feature values so that a desired outcome is predicted
- Contrastive explanations
- Diverse counterfactuals
- Feasible & actionable explanations

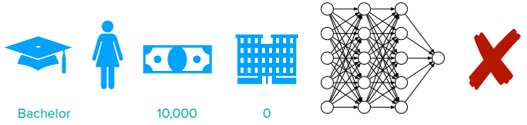


TYPES OF EXPLANATIONS

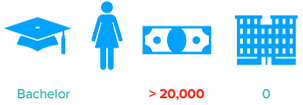




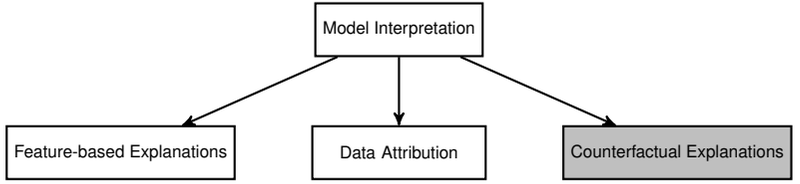
Example (loan application):

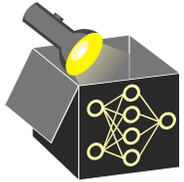


What can a person do to obtain a favorable prediction from a given model ?

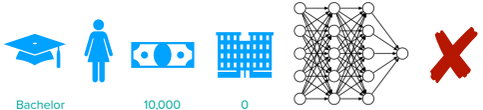


TYPES OF EXPLANATIONS

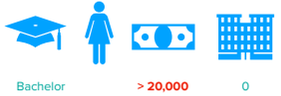




Example (loan application):



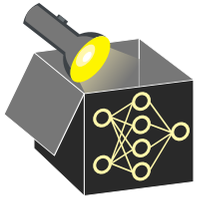
What can a person do to obtain a favorable prediction from a given model ?

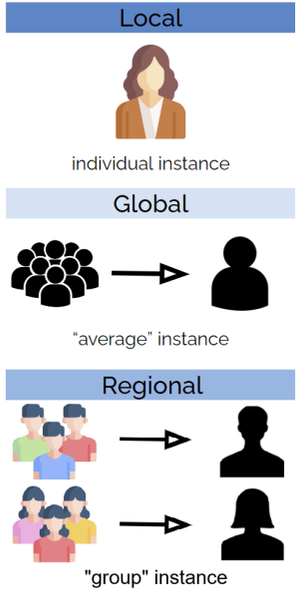


LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

Local – Explain model behavior for **single instances**:

- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

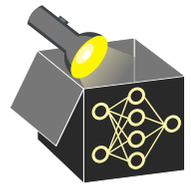


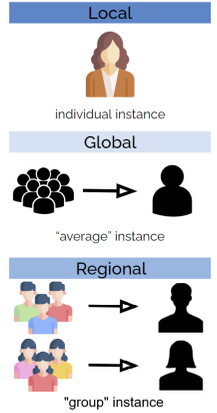


LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

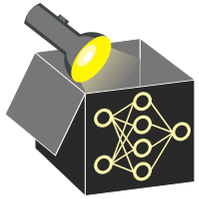
Local: Explain model behavior for **single instances**:

- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE





LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

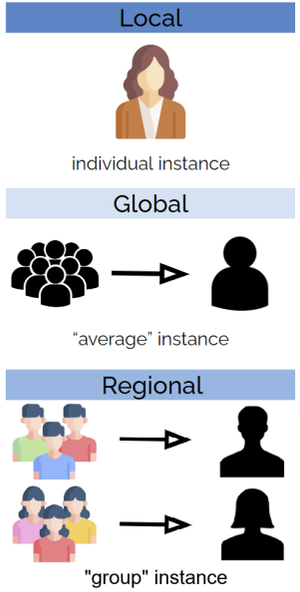


Local – Explain model behavior for **single instances**:

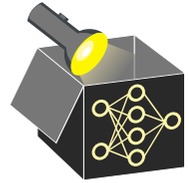
- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

Global – Explain model behavior for **entire input space**:

- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI



LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

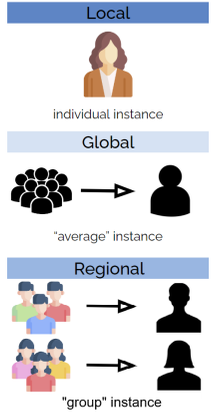


Local: Explain model behavior for **single instances**:

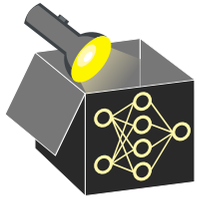
- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

Global: Explain model behavior for **entire input space**:

- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI



LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS



Local – Explain model behavior for **single instances**:

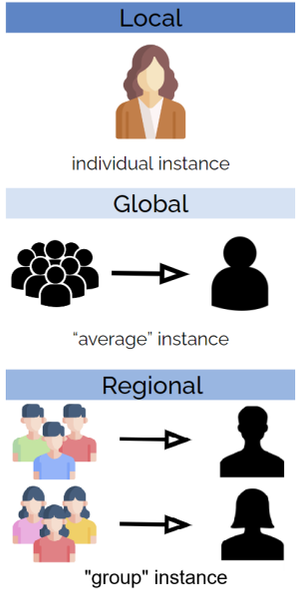
- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

Global – Explain model behavior for **entire input space**:

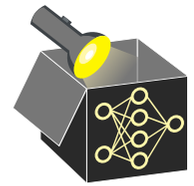
- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI

Regional explanations – for **subspaces** / **regions**:

- Compromise between nuanced & high-level insights
- Useful when local explanations group well without losing much detail



LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS



Local: Explain model behavior for **single instances**:

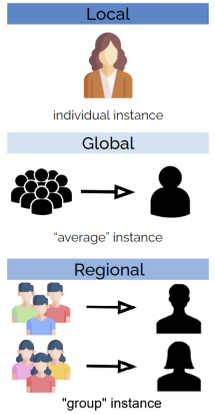
- Provide nuanced instance-specific insights
- Crucial for complex models where features typically affect instances differently (due to interactions)
- Examples: Counterfactuals, LIME, SHAP, ICE

Global: Explain model behavior for **entire input space**:

- Provide high-level insights into model behavior, often by aggregating local explanations
- Easier to communicate but loss of detail & over-simplification (hides differences)
- Examples: PD plots, ALE plots, PFI

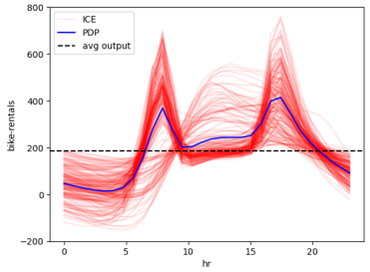
Regional explanations – for **subspaces** / **regions**:

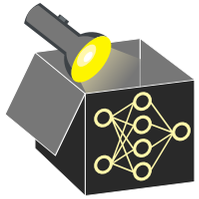
- Compromise between nuanced & high-level insights
- Useful when local explanations group well without losing much detail



LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

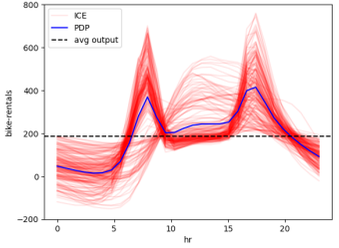
- **Local** (red): ICE curves for one instance
  ⇝ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  ⇝ Averaged curve hides heterogeneity

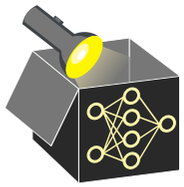




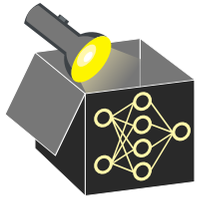
LOCAL, GLOBAL, REGIONAL EXPLANATIONS

- **Local** (red): ICE curves for one instance
  ⇝ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  ⇝ Averaged curve hides heterogeneity



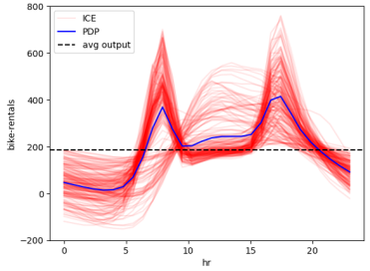


LOCAL, GLOBAL, AND REGIONAL EXPLANATIONS

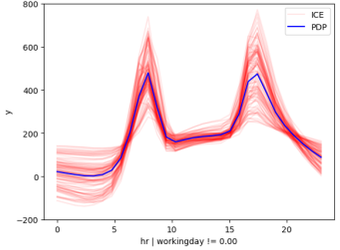


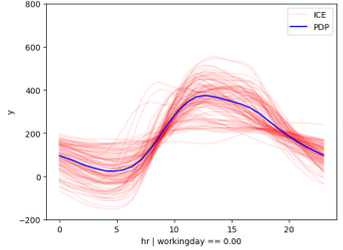
- **Local** (red): ICE curves for one instance
  ⇝ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  ⇝ Averaged curve hides heterogeneity
- **Regional**: Split data on `workingday`
  - Region 1: morning and evening peak
  - Region 2: late-morning leisure peak

  ⇝ Preserves detail without overload (challenge: find regions automatically)



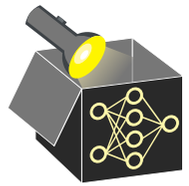
Region 1 (working day)





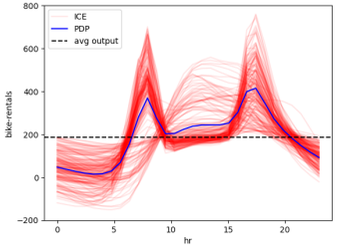
Region 2 (non-working day)

LOCAL, GLOBAL, REGIONAL EXPLANATIONS

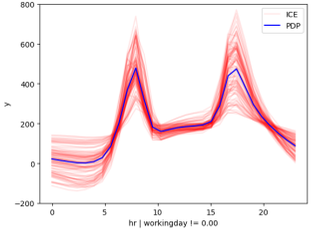


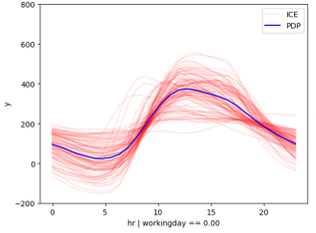
- **Local** (red): ICE curves for one instance
  ⇝ Detailed but cluttered/obscure pattern
- **Global** (blue): PDP averaged over *all* days
  ⇝ Averaged curve hides heterogeneity
- **Regional**: Split data on `workingday`
  - Region 1: morning and evening peak
  - Region 2: late-morning leisure peak

  ⇝ Preserves detail without overload
  ⇝ Challenge: find regions automatically



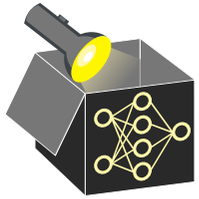
Region 1 (working day)



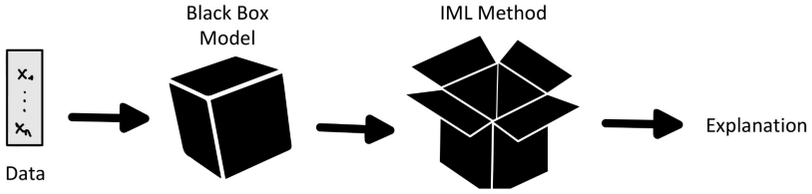


Region 2 (non-working day)

FIXED MODEL VS. REFITS

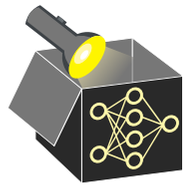


- Input of global interpretation methods: model + data, output: explanations
  ⇝ Explanations can be viewed as statistical estimators

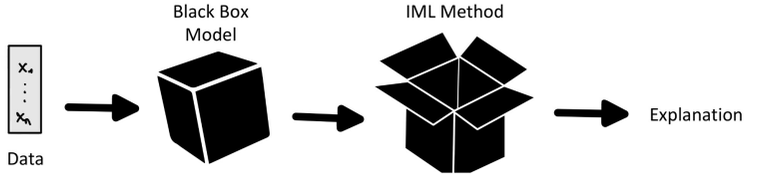


- Situation in ML: Deployed model is trained on all available data
  ⇝ No unseen test data left to, e.g., reliably estimate performance
  ⇝ IML method could use same data model was trained on
  ⇝ But: Some IML methods rely on measuring loss requiring unseen test data
- Alternative: Explain the inducer that created the model (instead of a fixed model)
  ⇝ Idea: Use resample strategies (e.g., 4-fold CV) as in performance estimation
  ⇝ Requires refitting

FIXED MODEL VS. REFITS



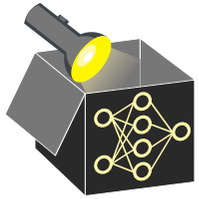
- Global interpretation methods: Input: model + data, output: explanations
  ⇝ Explanations can be viewed as statistical estimators

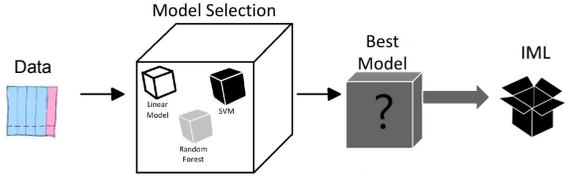


- Situation in ML: Deployed model is trained on all available data
  ⇝ No unseen test data left to, e.g., reliably estimate performance
  ⇝ IML method could use same data model was trained on
  ⇝ But: Some IML methods require measuring loss on unseen test data
- Alternative: Explain the inducer that created the model (not a fixed model)
  ⇝ Idea: Use resample strategies (e.g. CV) as in performance estimation
  ⇝ Requires refitting

LEVELS OF INTERPRETABILITY

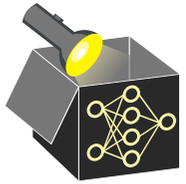
	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$

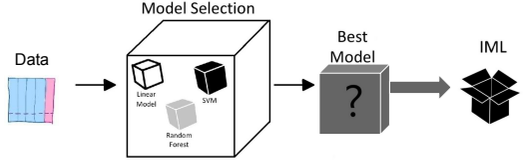




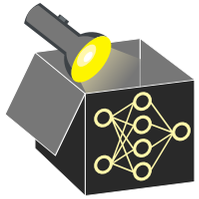
LEVELS OF INTERPRETABILITY

	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$

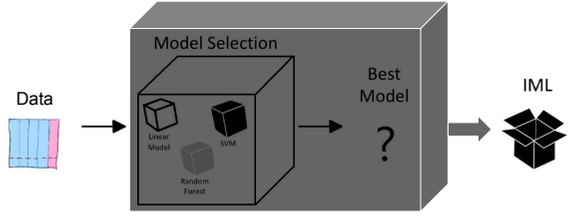




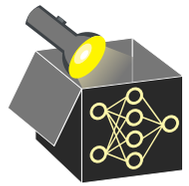
LEVELS OF INTERPRETABILITY



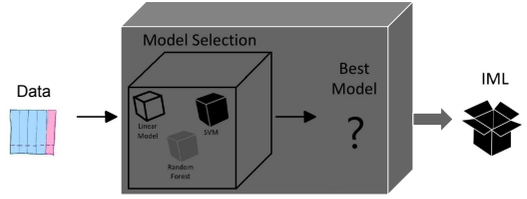
	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2nd level view	How does an optimizer choose a model based on a data set?	Model selection process (e.g., decisions made by AutoML systems or HPO process)



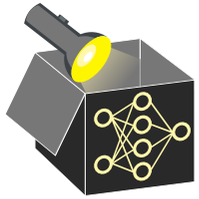
LEVELS OF INTERPRETABILITY



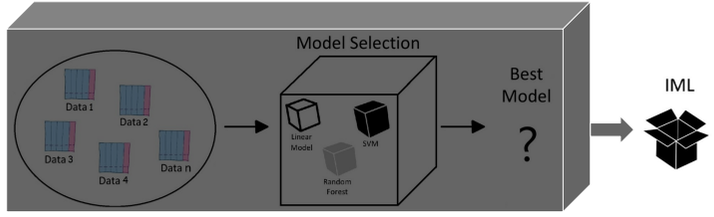
	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2nd level view	How does an optimizer choose a model based on a data set?	Model selection process (e.g., decisions made by AutoML systems or HPO)



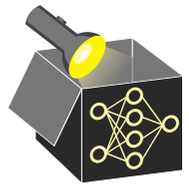
LEVELS OF INTERPRETABILITY



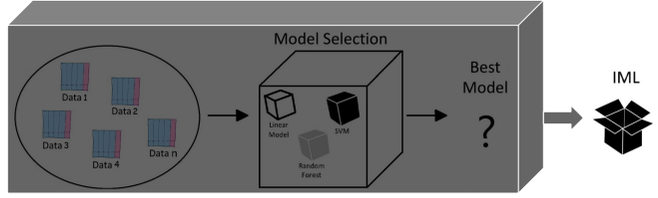
	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2nd level view	How does an optimizer choose a model based on a data set?	Model selection process (e.g., decisions made by AutoML systems or HPO process)
3rd level view	How do data properties relate to performance of a learner and its hyperparameters?	properties of ML algorithms in general (benchmark)



LEVELS OF INTERPRETABILITY

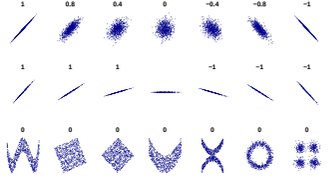


	Research Question	Objects of analysis
1st level view	How to explain a given model fitted on a data set?	(deployed) model $\theta \mapsto \hat{f}(\theta)$
2nd level view	How does an optimizer choose a model based on a data set?	Model selection process (e.g., decisions made by AutoML systems or HPO)
3rd level view	How do data properties relate to performance of a learner and its hyperparameters?	Properties of ML algorithms in general (benchmark)



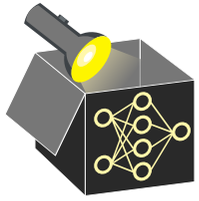
Interpretable Machine Learning

Correlation and Dependencies



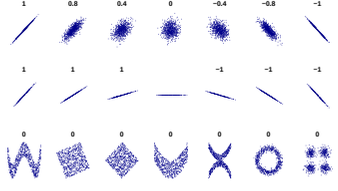
Learning goals

- Pearson correlation
- Coefficient of determination R^2
- Mutual information
- Correlation vs. dependence



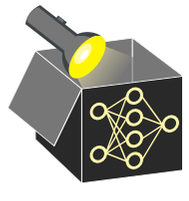
Interpretable Machine Learning

Correlation and Dependencies

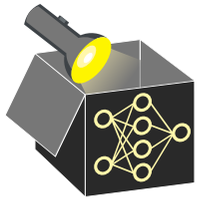


Learning goals

- Pearson correlation
- Coefficient of determination R^2
- Mutual information
- Correlation vs. dependence

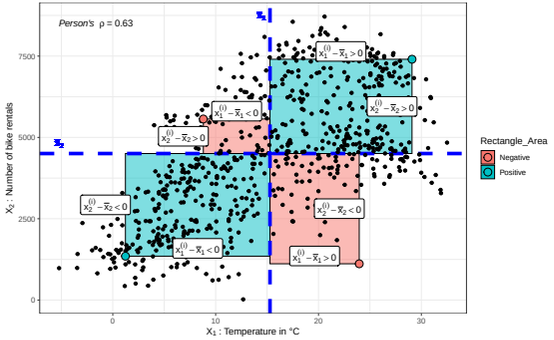


PEARSON'S CORRELATION COEFFICIENT ρ



Correlation often refers to Pearson's correlation (measures only **linear relationship**)

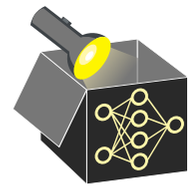
$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$



Geometric interpretation of ρ:

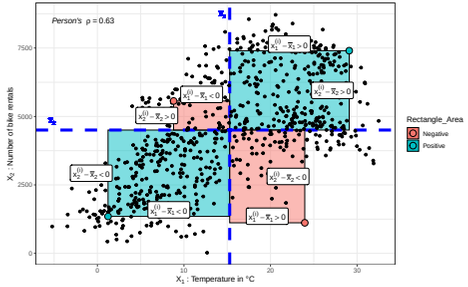
- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

PEARSON'S CORRELATION COEFFICIENT ρ



Correlation often refers to Pearson's correlation (measures only **linear relationship**)

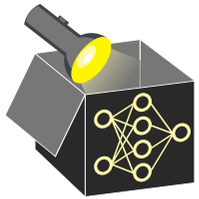
$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$



Geometric interpretation of ρ:

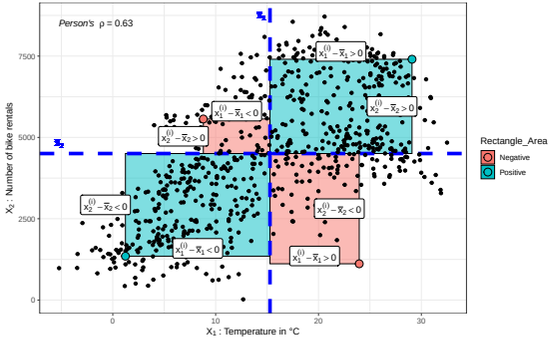
- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

PEARSON'S CORRELATION COEFFICIENT ρ



Correlation often refers to Pearson's correlation (measures only **linear relationship**)

$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$

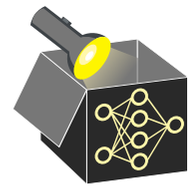


Geometric interpretation of ρ:

- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

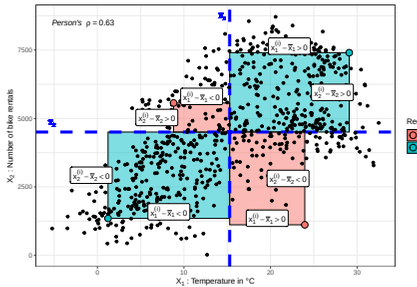
- $\rho > 0$ if positive areas dominate negative areas $\rightsquigarrow X_1, X_2$ positive correlated
- $\rho < 0$ if negative areas dominate positive areas $\rightsquigarrow X_1, X_2$ negative correlated
- $\rho = 0$ if area of rectangles cancels out $\rightsquigarrow X_1, X_2$ linearly uncorrelated

PEARSON'S CORRELATION COEFFICIENT ρ



Correlation often refers to Pearson's correlation (measures only **linear relationship**)

$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$



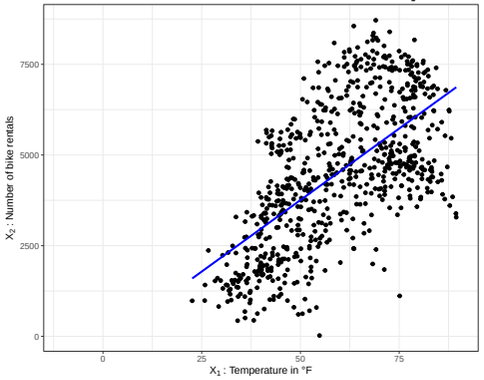
Geometric interpretation of ρ:

- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

- $\rho > 0$ if positive areas dominate negative areas $\rightsquigarrow X_1, X_2$ positive correlated
- $\rho < 0$ if negative areas dominate positive areas $\rightsquigarrow X_1, X_2$ negative correlated
- $\rho = 0$ if area of rectangles cancels out $\rightsquigarrow X_1, X_2$ linearly uncorrelated

COEFFICIENT OF DETERMINATION R^2

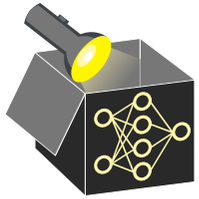
Another method to evaluate **linear dependency** between features is R^2



- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

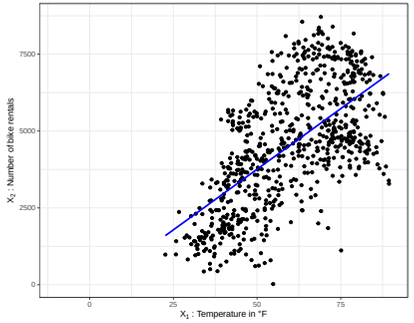
$\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence

$\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence



COEFFICIENT OF DETERMINATION R^2

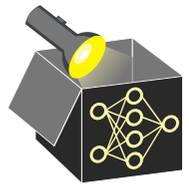
Another method to evaluate **linear dependency** between features is R^2



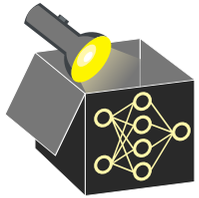
- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

$\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence

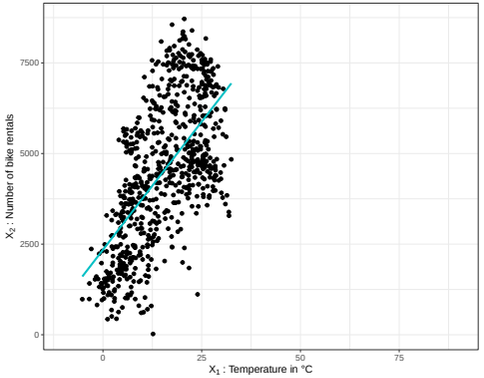
$\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence



COEFFICIENT OF DETERMINATION R^2

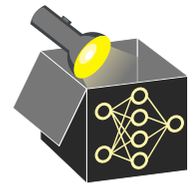


Another method to evaluate **linear dependency** between features is R^2

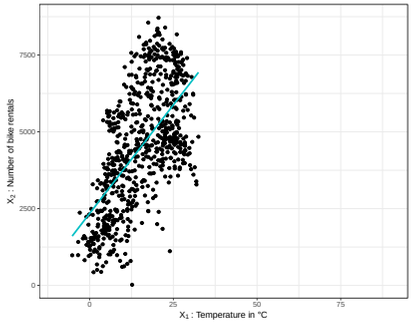


- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence
- $\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
- $\rightsquigarrow$ Re-scaling of x_1 or x_2 changes θ_1
- $\rightsquigarrow$ °F $\rightarrow$ °C $\Rightarrow \theta_1 = 78 \rightarrow \theta_1^* = 141$

COEFFICIENT OF DETERMINATION R^2

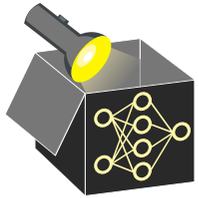


Another method to evaluate **linear dependency** between features is R^2

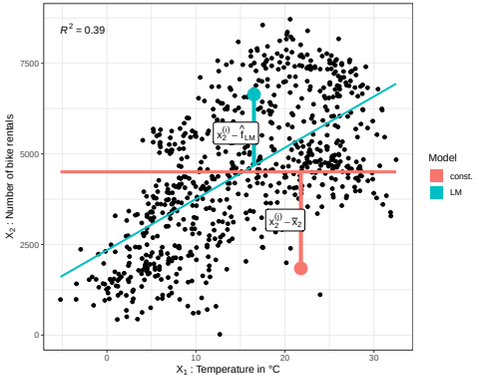


- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence
- $\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
- $\rightsquigarrow$ Re-scaling of x_1 or x_2 changes θ_1
- $\rightsquigarrow$ °F $\rightarrow$ °C $\Rightarrow \theta_1 = 78 \rightarrow \theta_1^* = 141$

COEFFICIENT OF DETERMINATION R^2



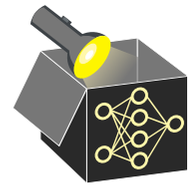
Another method to evaluate **linear dependency** between features is R^2



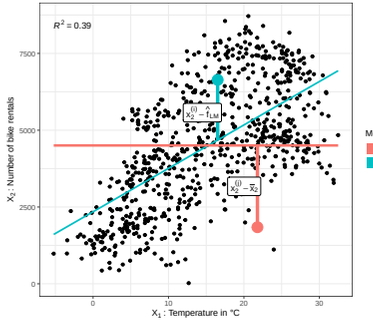
- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence
- $\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
- $\rightsquigarrow$ Re-scaling of x_1 or x_2 changes θ_1
- Set SSE_{LM} in relation to SSE of a constant model $\hat{f}_c = \bar{x}_2$
  $SSE_{LM} = \sum_{i=1}^{n} (x_2^{(i)} - \hat{f}_{LM}(x_1^{(i)}))^2$
  $SSE_c = \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2$

$\Rightarrow$ Measure of fitting quality of LM: $R^2 = 1 - \frac{SSE_{LM}}{SSE_c} \in [0, 1]$
$\Rightarrow \rho(X_1, X_2) = R$

COEFFICIENT OF DETERMINATION R^2



Another method to evaluate **linear dependency** between features is R^2



- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence
- $\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
- $\rightsquigarrow$ Re-scaling of x_1 or x_2 changes θ_1
- Set SSE_{LM} in relation to SSE of a constant model $\hat{f}_c = \bar{x}_2$
  $SSE_{LM} = \sum_{i=1}^{n} (x_2^{(i)} - \hat{f}_{LM}(x_1^{(i)}))^2$
  $SSE_c = \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2$

$\Rightarrow$ Measure of fitting quality of LM: $R^2 = 1 - \frac{SSE_{LM}}{SSE_c} \in [0, 1]$
$\Rightarrow \rho(X_1, X_2) = R$

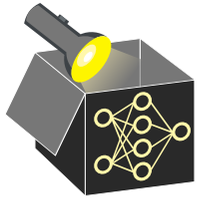
JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1



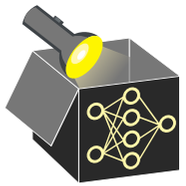
JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables X_1, X_2:

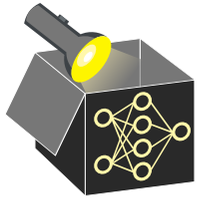
Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1



JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

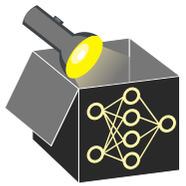
Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

$\rightsquigarrow$ In continuous case with integrals

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

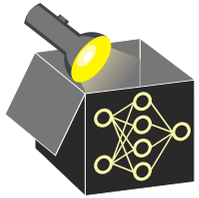
Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

$\rightsquigarrow$ In continuous case with integrals

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

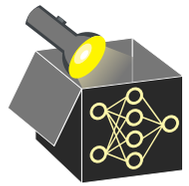
$\rightsquigarrow$ In continuous case with integrals

Conditional distribution

$$p_{X_1|X_2}(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

	$x_2 = 0$	$x_2 = 1$
$\mathbb{P}(X_1 = 0\|X_2 = x_2)$	0.67	0.43
$\mathbb{P}(X_1 = 1\|X_2 = x_2)$	0.33	0.57
$\sum$	1	1

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

$\rightsquigarrow$ In continuous case with integrals

Conditional distribution

$$p_{X_1|X_2}(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2) = \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

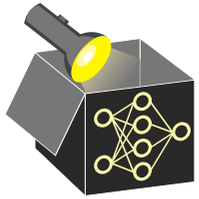
	$x_2 = 0$	$x_2 = 1$
$\mathbb{P}(X_1 = 0\|X_2 = x_2)$	0.67	0.43
$\mathbb{P}(X_1 = 1\|X_2 = x_2)$	0.33	0.57
$\sum$	1	1

DEPENDENCE

Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

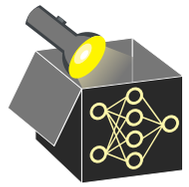


DEPENDENCE

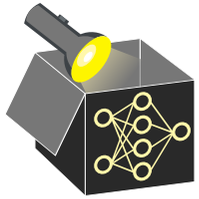
Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$



DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

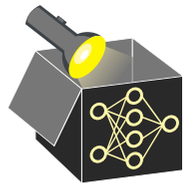
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowledge of X_k says nothing about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

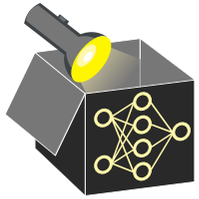
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

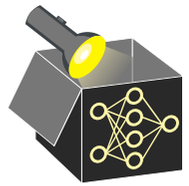
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowledge of X_k says nothing about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

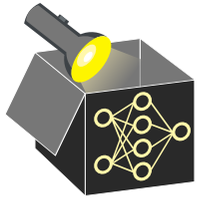
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

- Measuring complex dependencies is difficult but different measures exist
  Examples
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

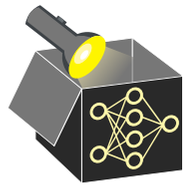
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowledge of X_k says nothing about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:** X_j, X_k independent $\Rightarrow \rho(X_j, X_k) = 0$ **but** $\rho(X_j, X_k) = 0 \not\Rightarrow X_j$, X_k indep. Equivalency holds if distribution is jointly normal

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

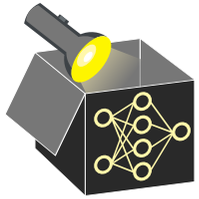
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

$$\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j) \text{ and } \mathbb{P}(X_k|X_j) = \mathbb{P}(X_k) \text{ (follows from cond. probability)}$$

- Measuring complex dependencies is difficult but different measures exist
  Examples
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:** X_j, X_k indep. $\Rightarrow \rho(X_j, X_k) = 0$ **but** $\rho(X_j, X_k) = 0 \not\Rightarrow X_j$, X_k indep. Equivalency holds if distribution is jointly normal

MUTUAL INFORMATION

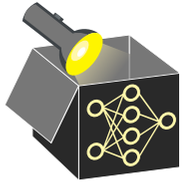


- MI describes expected amount of information shared by two random variables:

$$MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\right)\right]$$

- MI measures amount of "dependence" between features by looking how different the joint distribution is from pure independence $p(x_1, x_2) = p(x_1)p(x_2)$
  $\rightsquigarrow MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1,x_2)}{p(x_1,x_2)}\right)\right] = \mathbb{E}_{p(x_1,x_2)}\left[log(1)\right] = 0$
  $\rightsquigarrow MI(X_j, X_k) = 0$ if and only if the features are independent
- Unlike (Pearson) correlation, MI can also be computed for categorical features

MUTUAL INFORMATION



- MI describes expected amount of information shared by two RVs:

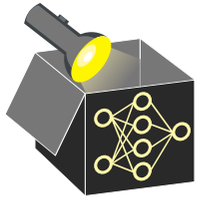
$$MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\right)\right]$$

- MI measures amount of "dependence" between features by looking how different the joint distribution is from pure indep. $p(x_1, x_2) = p(x_1)p(x_2)$
  $\rightsquigarrow MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1,x_2)}{p(x_1,x_2)}\right)\right] = \mathbb{E}_{p(x_1,x_2)}\left[log(1)\right] = 0$
  $\rightsquigarrow MI(X_j, X_k) = 0$ if and only if the features are independent
- Unlike (Pearson) correlation, MI is also defined for categorical features

MUTUAL INFORMATION: EXAMPLE

For two discrete RV X_1 and Y:

$$MI(X_1; Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1 \in \mathcal{X}_1}\sum_{y \in \mathcal{Y}} p(x_1,y) log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$



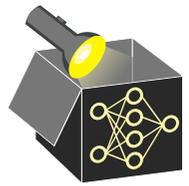
$\mathbf{X}_1$	...	**Y**
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

MUTUAL INFORMATION: EXAMPLE

For two discrete RV X_1 and Y:

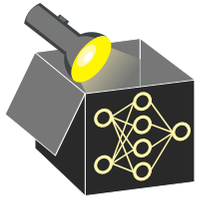
$$MI(X_1; Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1 \in \mathcal{X}_1}\sum_{y \in \mathcal{Y}} p(x_1,y) log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$



1	...	
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

MUTUAL INFORMATION: EXAMPLE



For two discrete RV X_1 and Y:

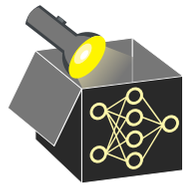
$$MI(X_1; Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1 \in \mathcal{X}_1}\sum_{y \in \mathcal{Y}} p(x_1,y) log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$

X_1	...	Y
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

$$\begin{aligned} MI(X_1; Y) &= 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &\quad + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &= 0.25\, log\left(\frac{0.25}{0.25}\right) \cdot 4 \\ &= 0.25\, log\,(1) \cdot 4 = 0 \end{aligned}$$

MUTUAL INFORMATION: EXAMPLE



For two discrete RV X_1 and Y:

$$MI(X_1; Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1 \in \mathcal{X}_1}\sum_{y \in \mathcal{Y}} p(x_1,y) log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$

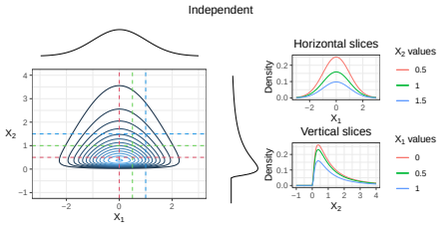
1	...	
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

$$\begin{aligned} MI(X_1; Y) &= 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &\quad + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25\, log\left(\frac{0.25}{0.5 \cdot 0.5}\right) \\ &= 0.25\, log\left(\frac{0.25}{0.25}\right) \cdot 4 \\ &= 0.25\, log\,(1) \cdot 4 = 0 \end{aligned}$$

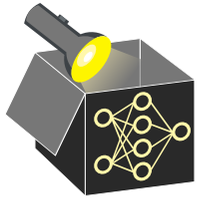
DEPENDENCE AND INDEPENDENCE

Example:



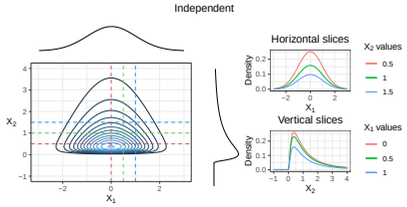
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$



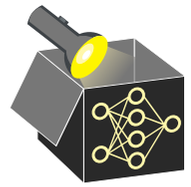
DEPENDENCE AND INDEPENDENCE

Example:



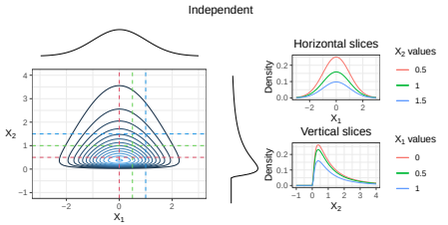
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$



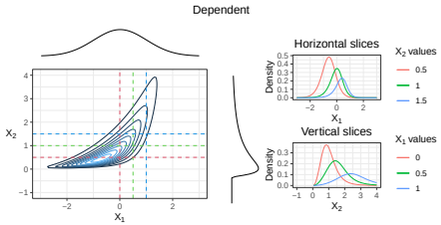
DEPENDENCE AND INDEPENDENCE

Example:

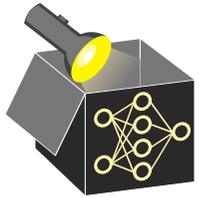


Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$

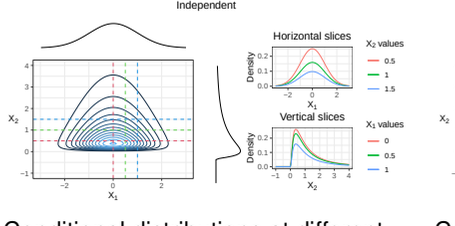


Conditional distributions do not match their marginal distributions



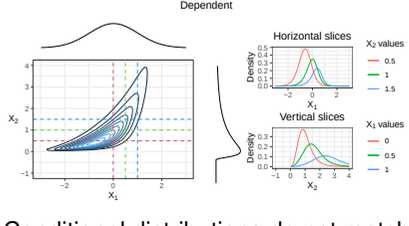
DEPENDENCE AND INDEPENDENCE

Example:

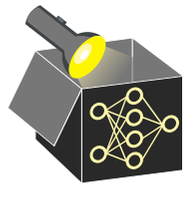


Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$

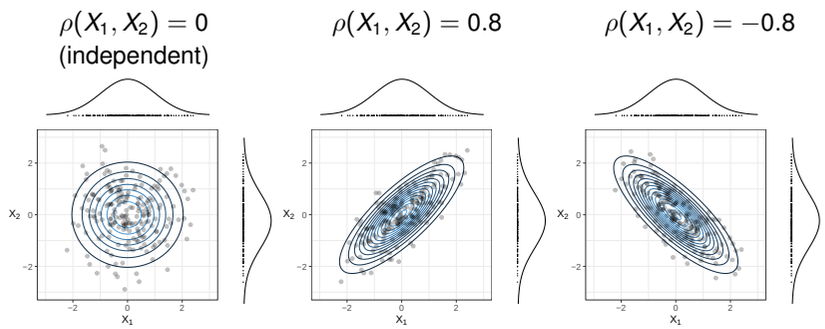


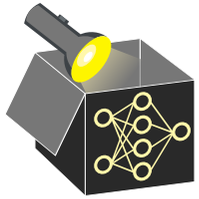
Conditional distributions do not match their marginal distributions



CORRELATION VS. DEPENDENCE

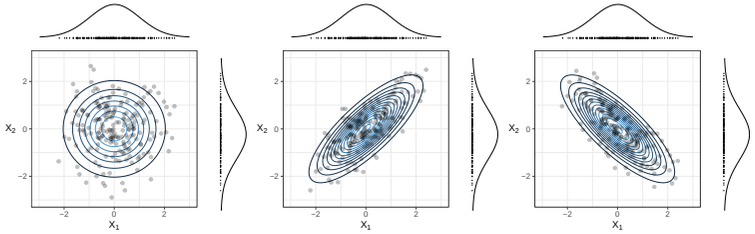
Illustration of bivariate normal distribution with different correlations $X_1, X_2 \sim N(0, 1)$

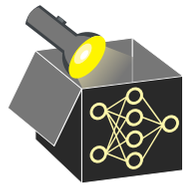




CORRELATION VS. DEPENDENCE

Illustration of bivariate normal distribution with different correlations $X_1, X_2 \sim N(0, 1)$





CORRELATION VS. DEPENDENCE

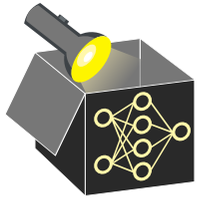
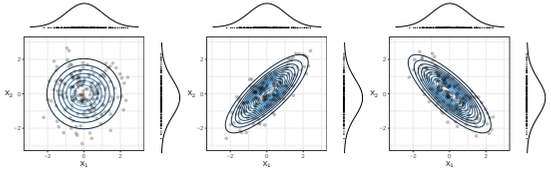


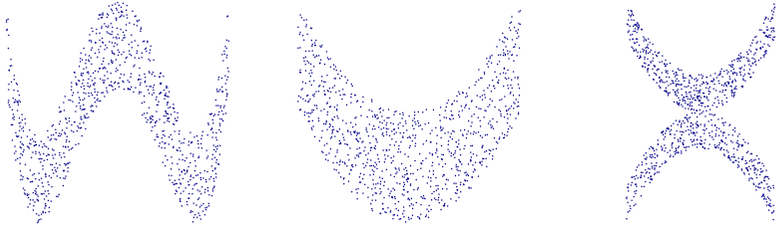
Illustration of bivariate normal distribution with different correlations $X_1, X_2 \sim N(0, 1)$

$\rho(X_1, X_2) = 0$ (independent) $\quad \rho(X_1, X_2) = 0.8 \quad \rho(X_1, X_2) = -0.8$



Examples with Pearson's correlation $\rho \approx 0$ but non-linear dependencies (MI $\neq 0$):

$\rho(X_1, X_2) = 0$, $MI(X_1, X_2) = 0.52$ $\quad \rho(X_1, X_2) = 0.01$, $MI(X_1, X_2) = 0.37$ $\quad \rho(X_1, X_2) = -0.06$, $MI(X_1, X_2) = 0.61$



CORRELATION VS. DEPENDENCE

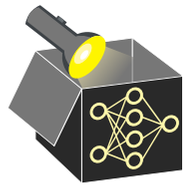
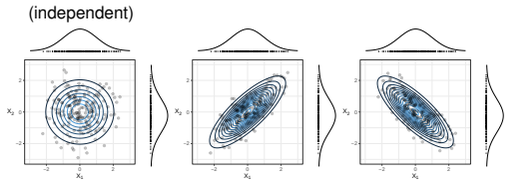


Illustration of bivariate normal distribution with different correlations $X_1, X_2 \sim N(0, 1)$

$\rho(X_1, X_2) = 0$ (independent) $\quad \rho(X_1, X_2) = 0.8 \quad \rho(X_1, X_2) = -0.8$

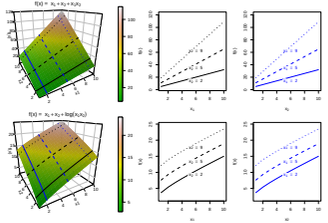


Examples with Pearson's corr. $\rho \approx 0$ but non-linear dependencies (MI $\neq 0$):

$\rho(X_1, X_2) = 0$, $MI(X_1, X_2) = 0.52$ $\quad \rho(X_1, X_2) = 0.01$, $MI(X_1, X_2) = 0.37$ $\quad \rho(X_1, X_2) = -0.06$, $MI(X_1, X_2) = 0.61$

Interpretable Machine Learning

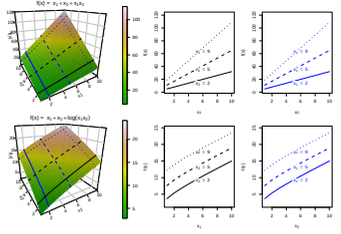
Feature Interactions



Learning goals

- Feature interactions
- Difference to feature dependencies

Interpretable Machine Learning
Feature Interactions

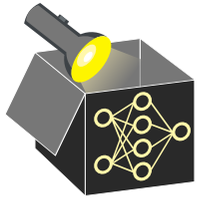


Learning goals

- Feature interactions
- Difference to feature dependencies

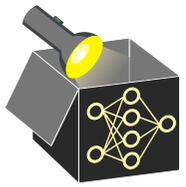
FEATURE INTERACTIONS

- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and $\hat{f}(X)$ or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model

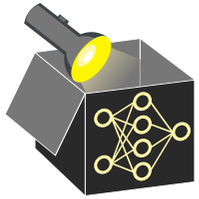


FEATURE INTERACTIONS

- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and (X) or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model

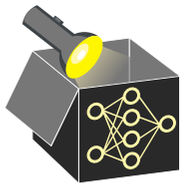


FEATURE INTERACTIONS



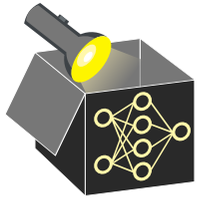
- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and $\hat{f}(X)$ or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  $\rightsquigarrow$ Difficult to identify interactions, especially when features are dependent

FEATURE INTERACTIONS

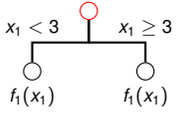


- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and (X) or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  $\rightsquigarrow$ Difficult to identify interactions, especially when features are dep.

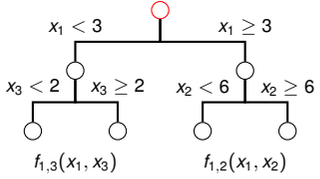
FEATURE INTERACTIONS



- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and $\hat{f}(X)$ or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  $\rightsquigarrow$ Difficult to identify interactions, especially when features are dependent
- Interactions: A feature's effect on the prediction depends on other features
  $\rightsquigarrow$ Example: $\hat{f}(\mathbf{x}) = x_1 x_2 \Rightarrow$ Effect of x_1 on $\hat{f}$ depends on x_2 and vice versa

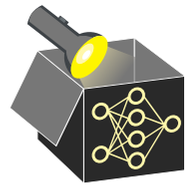


No interaction

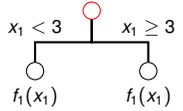


Interactions: x_1 and x_3,
x_1 and x_2
No interactions: x_2 and x_3

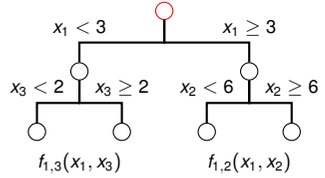
FEATURE INTERACTIONS



- Feature dependencies concern data distribution
- Feature interactions may occur in structure of **both** model or DGP (e.g., functional relationship between X and (X) or X and $Y = f(X)$)
  $\rightsquigarrow$ Feature dependencies may lead to feature interactions in a model
- No. of potential interactions increases exponentially with no. of features
  $\rightsquigarrow$ Difficult to identify interactions, especially when features are dep.
- Interactions: Feature's effect on the prediction depends on other features
  $\rightsquigarrow$ Example: $() = x_1 x_2 \Rightarrow$ Effect of x_1 on depends on x_2 and vice versa



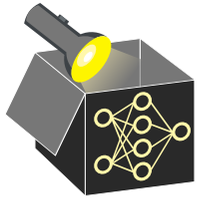
No interaction



Interactions: x_1 and x_3,
x_1 and x_2
No interactions: x_2 and x_3

FEATURE INTERACTIONS

Friedman and Popescu (2008)



Definition: A function $f(\mathbf{x})$ contains an interaction between x_j and x_k if a difference in $f(\mathbf{x})$-values due to changes in x_j will also depend on x_k, i.e.:

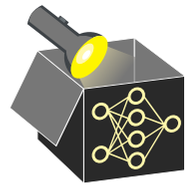
$$\mathbb{E}\left[\frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right]^2 > 0$$

$\Rightarrow$ If x_j and x_k do not interact, $f(\mathbf{x})$ is sum of 2 functions, each independent of x_j, x_k:

$$f(\mathbf{x}) = f_{-j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) + f_{-k}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)$$

FEATURE INTERACTIONS

FRIEDMAN_POPESCU



Definition: A function $f()$ contains an interaction between x_j and x_k if a difference in $f()$-values due to changes in x_j will also depend on x_k, i.e.:

$$\mathbb{E}\left[\frac{\partial^2 f()}{\partial x_j \partial x_k}\right]^2 > 0$$

$\Rightarrow$ If x_j and x_k don't interact, $f()$ is sum of 2 functions, each indep. of x_j, x_k:

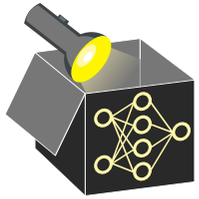
$$f() = f_{-j}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) + f_{-k}(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p)$$

FEATURE INTERACTIONS

Example: $f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2$ (not separable)

$$\mathbb{E}\left[\frac{\partial^2(x_1+x_2+x_1 \cdot x_2)}{\partial x_1 \partial x_2}\right]^2 = \mathbb{E}\left[\frac{\partial(1+x_2)}{\partial x_2}\right]^2 = 1 > 0$$

$\Rightarrow$ interaction between x_1 and x_2

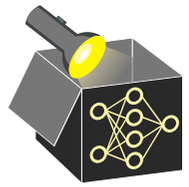


FEATURE INTERACTIONS

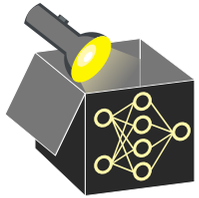
Example: $f() = x_1 + x_2 + x_1 \cdot x_2$ (not separable)

$$\mathbb{E}\left[\frac{\partial^2(x_1+x_2+x_1 \cdot x_2)}{\partial x_1 \partial x_2}\right]^2 = \mathbb{E}\left[\frac{\partial(1+x_2)}{\partial x_2}\right]^2 = 1 > 0$$

$\Rightarrow$ interaction between x_1 and x_2



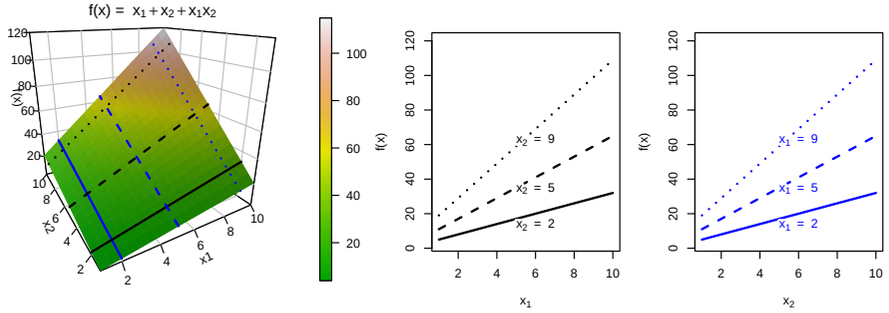
FEATURE INTERACTIONS



Example: $f(\mathbf{x}) = x_1 + x_2 + x_1 \cdot x_2$ (not separable)

$$\mathbb{E}\left[\frac{\partial^2(x_1+x_2+x_1 \cdot x_2)}{\partial x_1 \partial x_2}\right]^2 = \mathbb{E}\left[\frac{\partial(1+x_2)}{\partial x_2}\right]^2 = 1 > 0$$

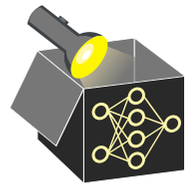
$\Rightarrow$ interaction between x_1 and x_2



- Effect of x_1 on $f(\mathbf{x})$ varies with x_2 (and vice versa)

$\Rightarrow$ Different slopes

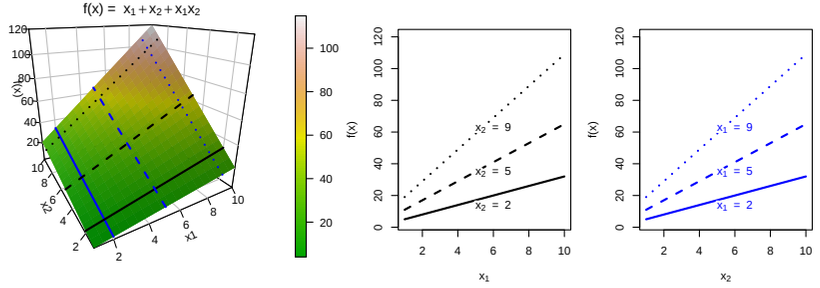
FEATURE INTERACTIONS



Example: $f() = x_1 + x_2 + x_1 \cdot x_2$ (not separable)

$$\mathbb{E}\left[\frac{\partial^2(x_1+x_2+x_1 \cdot x_2)}{\partial x_1 \partial x_2}\right]^2 = \mathbb{E}\left[\frac{\partial(1+x_2)}{\partial x_2}\right]^2 = 1 > 0$$

$\Rightarrow$ interaction between x_1 and x_2



- Effect of x_1 on $f()$ varies with x_2 (and vice versa)

$\Rightarrow$ Different slopes

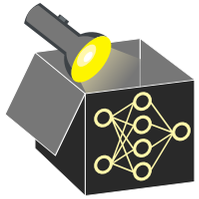
FEATURE INTERACTIONS

Example of separable function:

$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$\Rightarrow f(\mathbf{x}) = f_1(x_1) + f_2(x_2)$ with $f_1(x_1) = x_1 + \log(x_1)$ and $f_2(x_2) = x_2 + \log(x_2)$

$\Rightarrow$ no interactions due to separability, also $\mathbb{E}\left[\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right]^2 = 0$



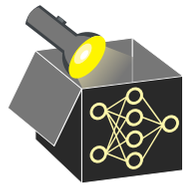
FEATURE INTERACTIONS

Example of separable function:

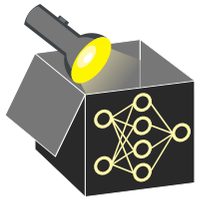
$$f() = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$\Rightarrow f() = f_1(x_1) + f_2(x_2)$ with $f_1(x_1) = x_1 + \log(x_1)$ and $f_2(x_2) = x_2 + \log(x_2)$

$\Rightarrow$ no interactions due to separability, also $\mathbb{E}\left[\frac{\partial^2 f()}{\partial x_1 \partial x_2}\right]^2 = 0$



FEATURE INTERACTIONS

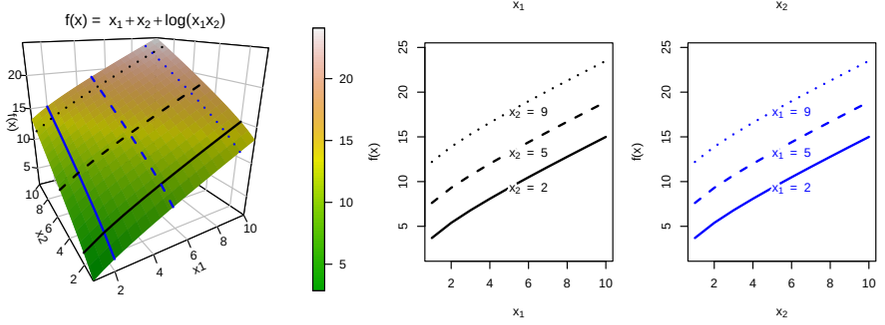


Example of separable function:

$$f(\mathbf{x}) = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$\Rightarrow f(\mathbf{x}) = f_1(x_1) + f_2(x_2)$ with $f_1(x_1) = x_1 + \log(x_1)$ and $f_2(x_2) = x_2 + \log(x_2)$

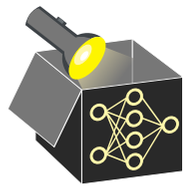
$\Rightarrow$ no interactions due to separability, also $\mathbb{E}\left[\frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2}\right]^2 = 0$



- Effect of x_1 on $f(\mathbf{x})$ stays the same for different x_2 values (and vice versa)

$\Rightarrow$ Parallel lines at different horizontal (blue) or vertical (black) slices

FEATURE INTERACTIONS

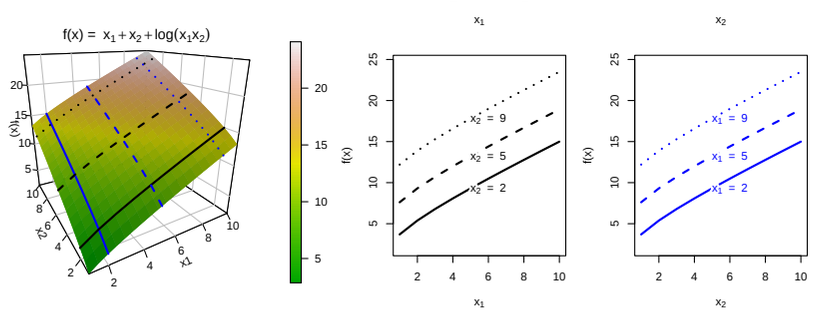


Example of separable function:

$$f() = x_1 + x_2 + \log(x_1 \cdot x_2) = x_1 + x_2 + \log(x_1) + \log(x_2)$$

$\Rightarrow f() = f_1(x_1) + f_2(x_2)$ with $f_1(x_1) = x_1 + \log(x_1)$ and $f_2(x_2) = x_2 + \log(x_2)$

$\Rightarrow$ no interactions due to separability, also $\mathbb{E}\left[\frac{\partial^2 f()}{\partial x_1 \partial x_2}\right]^2 = 0$



- Effect of x_1 on $f()$ stays the same for different x_2 values (and vice versa)

$\Rightarrow$ Parallel lines at different horizontal (blue) or vertical (black) slices