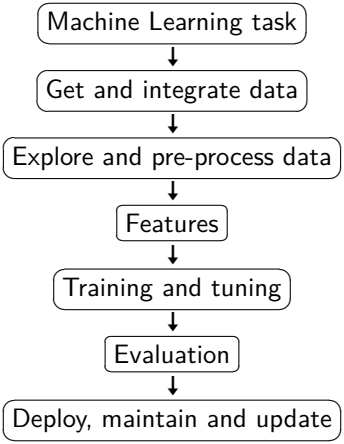
ML Basics

Machine Learning for Social Science

ML process


Machine Learning task
Get and integrate data
Explore and pre-process data
Features
Training and tuning
Evaluation
Deploy, maintain and update

Introduction

Unsupervised Learning

- Finding patterns in data using a set of input variables X

Supervised Learning

- Predicting an output variable Y based on a set of input variables X
  1. Learn the relationship between input and output using **training data** (with X and Y)

$$Y = f(X) + \varepsilon$$

  2. Predict the output based on the prediction model (of step 1) for **new test data** ($\sim$only X available)
- continuous Y: regression, categorical Y: classification
- Focus on **prediction**

Introduction

Unsupervised Learning

- Finding patterns in data using a set of input variables X

Supervised Learning

- Predicting an output variable Y based on a set of input variables X
  1. Learn the relationship between input and output using **training data** (with X and Y)

$$Y = f(X) + \varepsilon$$

  2. Predict the output based on the prediction model (of step 1) for **new test data** ($\sim$only X available)
- continuous Y: regression, categorical Y: classification
- Focus on **prediction**

Introduction

Unsupervised Learning

- Finding patterns in data using a set of input variables X

Supervised Learning

- Predicting an output variable Y based on a set of input variables X
  1. Learn the relationship between input and output using **training data** (with X and Y)

$$Y = f(X) + \varepsilon$$

  2. Predict the output based on the prediction model (of step 1) for **new test data** ($\sim$only X available)
- continuous Y: regression, categorical Y: classification
- Focus on **prediction**

Introduction

Unsupervised Learning

- Finding patterns in data using a set of input variables X

Supervised Learning

- Predicting an output variable Y based on a set of input variables X
  1. Learn the relationship between input and output using **training data** (with X and Y)

$$Y = f(X) + \varepsilon$$

  2. Predict the output based on the prediction model (of step 1) for **new test data** ($\sim$only X available)
- continuous Y: regression, categorical Y: classification
- Focus on **prediction**

ML basics

Supervised Learning: Find function $f(x)$ that makes optimal predictions in a **new data set**

Prerequisites:

- **Representation**: What is the *hypothesis space*, the family of functions to search over?
  - Describes possible relationships between X and Y
  - Examples: $f(x) = x'\beta$ is linear, or f is a tree.
- **Evaluation**: What is the criterion to choose between different functions?
  - Measures predictive performance
  - Examples: Mean Squared Error, Logistic Loss
- **Computation**: How is f actually calculated?
  - Speed and memory space may be limiting factors

ML basics

Supervised Learning: Find function $f(x)$ that makes optimal predictions in a **new data set**

Prerequisites:

- **Representation**: What is the *hypothesis space*, the family of functions to search over?
  - Describes possible relationships between X and Y
  - Examples: $f(x) = x'\beta$ is linear, or f is a tree.
- **Evaluation**: What is the criterion to choose between different functions?
  - Measures predictive performance
  - Examples: Mean Squared Error, Logistic Loss
- **Computation**: How is f actually calculated?
  - Speed and memory space may be limiting factors

ML basics

Supervised Learning: Find function $f(x)$ that makes optimal predictions in a **new data set**

Prerequisites:

- **Representation**: What is the *hypothesis space*, the family of functions to search over?
  - Describes possible relationships between X and Y
  - Examples: $f(x) = x'\beta$ is linear, or f is a tree.
- **Evaluation**: What is the criterion to choose between different functions?
  - Measures predictive performance
  - Examples: Mean Squared Error, Logistic Loss
- **Computation**: How is f actually calculated?
  - Speed and memory space may be limiting factors

ML basics

Supervised Learning: Find function $f(x)$ that makes optimal predictions in a **new data set**

Prerequisites:

- **Representation**: What is the *hypothesis space*, the family of functions to search over?
  - Describes possible relationships between X and Y
  - Examples: $f(x) = x'\beta$ is linear, or f is a tree.
- **Evaluation**: What is the criterion to choose between different functions?
  - Measures predictive performance
  - Examples: Mean Squared Error, Logistic Loss
- **Computation**: How is f actually calculated?
  - Speed and memory space may be limiting factors

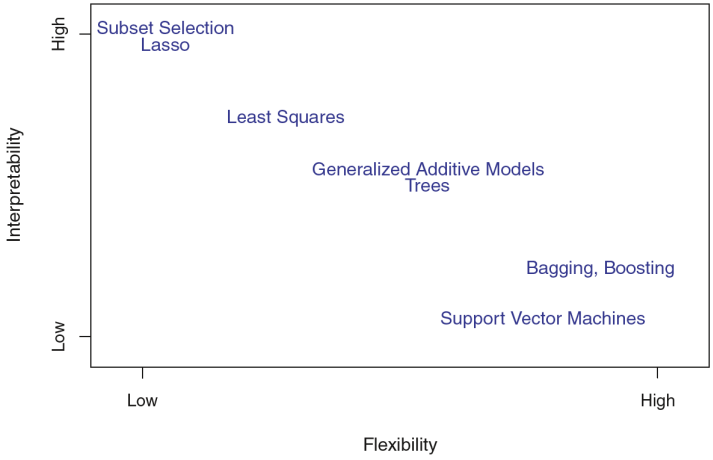
ML basics

Table: Estimating $f(x)$

Regression methods	(tree-based) ML methods
parametric	non-parametric
linearity, additivity	flexible functional form
prior model specification	"built-in" feature selection
theory-driven	data-driven
$\rightarrow$ Inference	$\rightarrow$ Prediction

ML basics

Figure: Flexibility-Interpretability Trade-Off



James et al. (2013)

ML basics

Estimation requires (at least implicitly):

- a target distribution $\mathbb{P}_{Y|X}$, or
- a loss function

Table: Possible choices

Setting	Loss	Target $f(x)$
Regression	$(y - f(x))^2$	$\text{mean}(y\|x)$
Regression	$\|y - f(x)\|$	$\text{median}(y\|x)$
Regression	$\rho_\tau(y - f(x))$	$F^{-1}_{y\|x}(\tau)$
Classification	Deviance	$\pi_{y\|x}$

Prediction perspective:

- Machine Learning mindset is more focused on evaluation criteria – and therefore – on loss functions

References

Buskirk, T. D., Kirchner, A., Eck, A., Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice 11*(1).

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM 55*(10), 78–87.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.