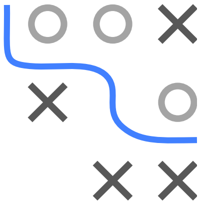
Introduction to Machine Learning



The Two Cultures of Statistical Modeling

MODELING: TWO CULTURES

Statistics, the Data Modeling Culture

- In a strongly simplified world an arbitrary outcome y is produced by “nature” from the features given in x
- The knowledge about nature’s true mechanisms ranges from entirely unknown (or stochastic) to established (scientific), possibly deterministic explanations

MODELING: TWO CULTURES / 2

- Focus on the modeling of data, which can be reduced to two targets:
  1. Learn a model to predict the outcome for new covariates
  2. Get a better understanding about the relationship between covariates and outcome

- Find a stochastic model of the data-generating process:

$$y = f(x, \text{parameters}, \text{random error})$$

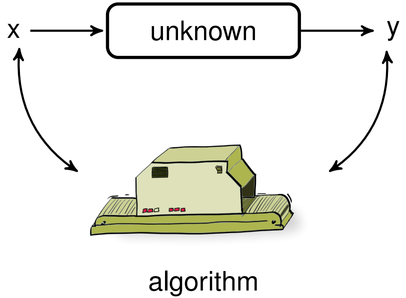
MODELING: TWO CULTURES / 3

In this “data modeling culture”, a stochastic model for the data-generating process is assumed

Typical assumptions and restrictions

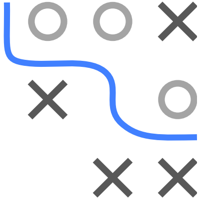
- Specific stochastic model that generated the data
- Distribution of residuals
- Linearity, additivity (e.g. linear predictor)
- Manual specification of interactions

Machine Learning, the Algorithmic Modeling Culture



Find a function $f(\mathbf{x})$ that minimizes the loss: $L\left(y, f(\mathbf{x})\right)$

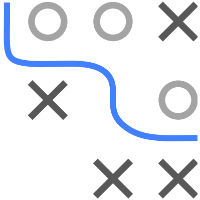
MODELING: TWO CULTURES / 5



- In the "algorithmic modeling culture", the true mechanism is treated as unknown
- The goal is not finding the true data-generating process but developing an algorithm that imitates/predicts (specific aspects of) a data-generating process as closely as possible
- Modeling is reduced to a mere problem of function optimization: Given the covariates x, outcome y and a loss function, find a function $f(\mathbf{x})$ which minimizes the loss for the prediction of the outcome

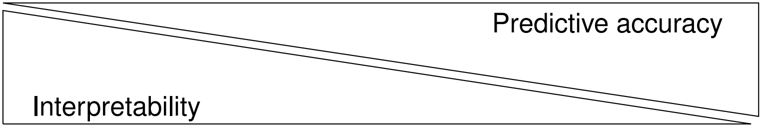
MODELING: TWO CULTURES / 6

Algorithm in Machine Learning



- Boosting
- Support Vector Machines
- Artificial neural networks
- Random Forests
- Hidden Markov
- Bayes-Nets
- . . .

PREDICTION VS. INTERPRETATION

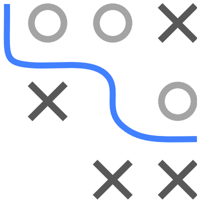


Tree	Random Forest
Logistic Regression	Neural networks
...	...

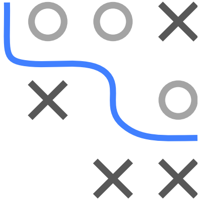
PREDICTION VS. INTERPRETATION / 2

- There is a trade-off between interpretability and predictive accuracy: models that yield accurate predictions are often complex and models that are easy to interpret are often bad predictors

- Example logistic regression and *k* Nearest Neighbors: in LR, we can inspect each coefficient and understand how changes in a single feature affect the class probabilities. kNN offers no such interpretability, but if the class boundaries are very nonlinear, it will have much better predictive accuracy.

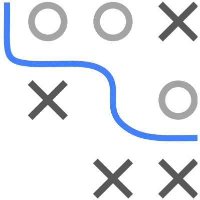


DIMENSIONALITY OF THE DATA



- The higher the dimensionality of the data (# covariates) the more difficult is the separation of signal and noise
- Common practice in data modeling: variable selection (by expert selection or data driven) and reduction of dimensionality (e.g. PCA)
- Common practice in algorithmic modeling: Engineering of new features (covariates) to increase predictive accuracy; algorithms robust for many covariates

MODELING: TWO CULTURES

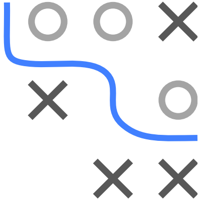


Problems and Blindspots of Data Modeling Culture:

- Conclusions about assumed model are interpreted as being about nature (reification).
- Model assumptions often violated.
- Often improper model evaluation presuming model validity
  $\Rightarrow$ can lead to irrelevant theory and questionable statistical conclusions
- Data models fail in areas like image and speech recognition

MODELING: TWO CULTURES / 2

Problems and Blindspots of Algorithmic Modeling Culture:



- Uncertainty quantification often difficult / impossible, almost always an afterthought.
- Models are often uninterpretable “black boxes”:
  $\Rightarrow$ Can you trust something you don’t understand?
- Often ignores suitable sampling plans or issues with data provenance that can jeopardize generalizability

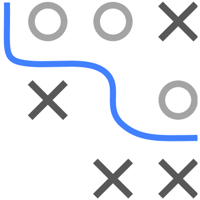
MODELING: TWO CULTURES / 3

Different terminology for machine learning and statistics:

Machine Learning	Statistics
Feature, Attribute	Covariate
Label	Response
Example, Instance	Observation
Weight	Parameter, Coefficient
Bias term	Intercept
Minimizing loss	Maximizing likelihood / Estimating posterior
Learning	Fitting, Estimation
Hypothesis	(Fitted) Model
Learner	Model (Class)
Supervised Learning	Regression / Classification
Unsupervised Learning	Density estimation / Clustering
Data Mining (good)	Data Mining (bad)

see also: `https://ubc-mds.github.io/resources_pages/terminology`

MODELING: TWO CULTURES / 4



Summary

Data modeling culture: “The model is true.”

Tries to estimate stochastic properties of the true data-generating process and focuses on parameters and their uncertainty.

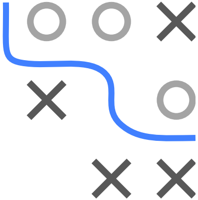
Algorithmic modeling culture: “The model is useful.”

Tries to minimize some measure of divergence between observations from the data-generating process and a function that imitates its behavior and focuses on predictive accuracy.

These are broad generalizations, there is much overlap and synergy between the two perspectives.

MODELING: TWO CULTURES / 5

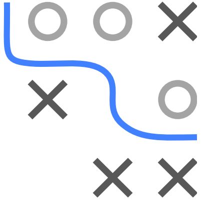
Rashomon Effect



In practice, many different models often describe a given set of data equally well, which makes it difficult to identify a “true” data-generating process.

In practice, using different loss functions / evaluation schemes will yield different optimal models, which makes it difficult to identify the “most useful” model.

PARAMETERS, STATISTICS AND SUPERVISED MACHINE LEARNING



- Supervised ML additionally assumes that f is of a certain “form” or comes from a certain *class of functions*.
  This is necessary to make the problem of automatically finding a “good” model feasible at all.
- The specific behavior of a mapping from this class can then be described by **parameters** which defines its shape.
- Statistics also studies how to learn such functions (or, rather: their parameters) from example data and how to perform inference on them and interpret the results.
- For historical reasons, statistics is mostly focused on fairly simple classes of mappings, like (generalized) linear models.
- Supervised ML also includes more complex kinds of mappings that can often deal with more complicated and high-dimensional inputs.