Machine Learning

Nathan Sprague
JMU Department of Computer Science

Spring 2024

Definitions...

- Here are some terms... How are they related?
  - Machine Learning
  - Statistics
  - Artificial Intelligence
  - Data Mining
  - Deep Learning

Definitions

- No universally accepted definitions for any of these things...

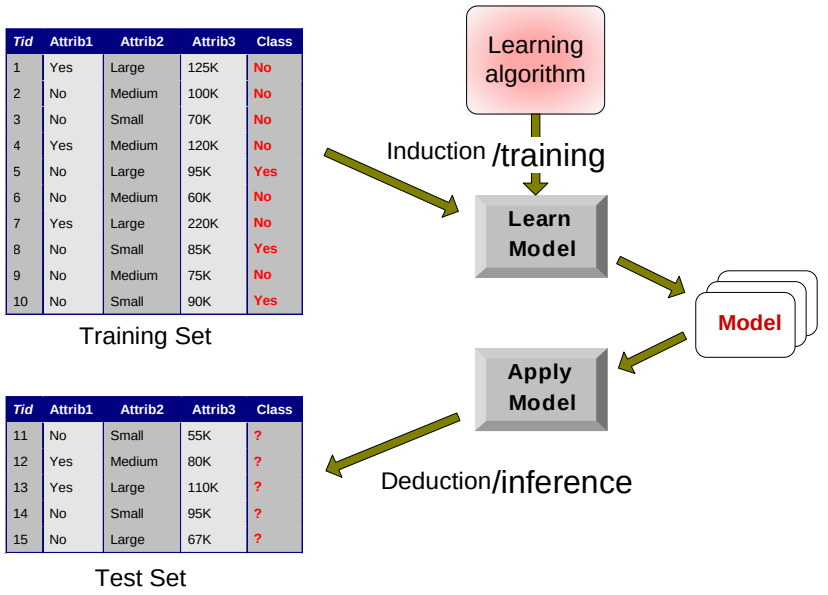
Machine Learning

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

  Mitchell, T. (1997). Machine Learning. McGraw Hill.

- "Field of study that gives computers the ability to learn without being explicitly programmed"

  Widely attributed to A. L. Samuel, "Some studies in machine learning using the game of checkers," in IBM Journal of Research and Development, vol. 44, no. 1.2, pp. 206-226, 1959. … but probably just paraphrasing the ideas in the paper.



Statistics

- “A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data”

  https://www.merriam-webster.com/dictionary/statistics

- Emphasis on mathematical rigor... Which tends to encourage relatively simple models with relatively few parameters

Artificial Intelligence

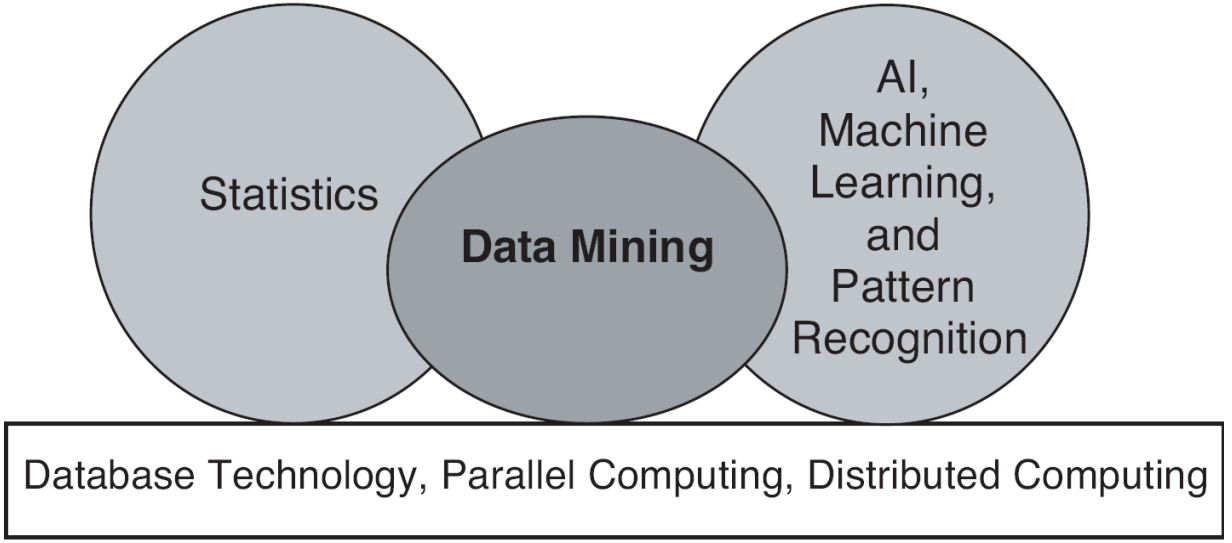
- AI:
  - “Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment.”

Nils J. Nilsson, The Quest for Artificial Intelligence: A History of Ideas and Achievements (Cambridge, UK: Cambridge University Press, 2010).

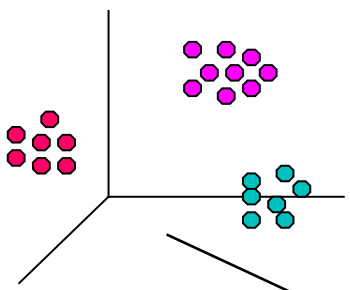
Data Mining

- "Data mining is the process of automatically discovering useful information in large data repositories"

Pang-Ning Tan et. al., Introduction to Data Mining, 2nd Edition (Pearson, 2019).



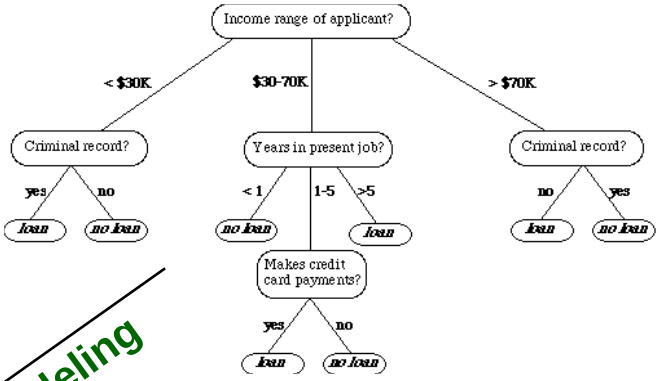
Data Mining Tasks



Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

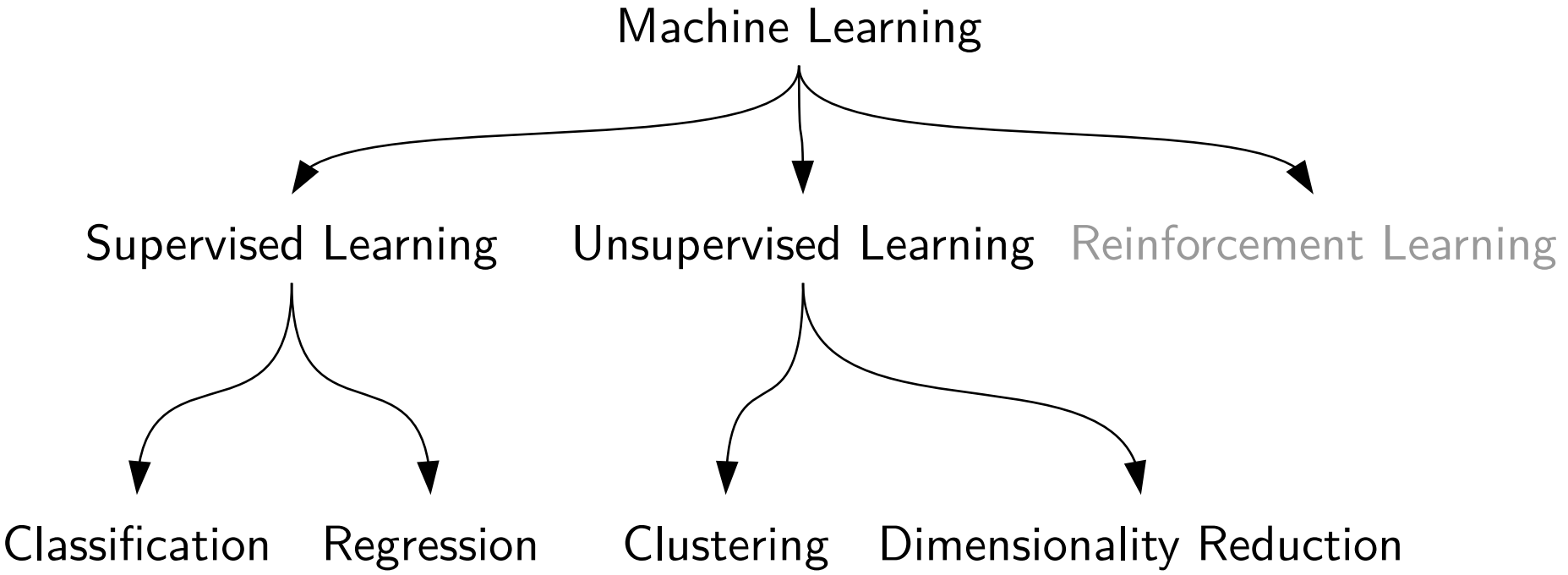


Predictive Modeling

Association Rules

Anomaly Detection

Traditional Machine Learning Task Breakdown



Interesting Times

- The last 10-15 years have seen dramatic progress in machine learning
- Much of this can be attributed to progress in **deep learning**

Deep Learning

- "**Representation learning** is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. **Deep-learning** methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level."

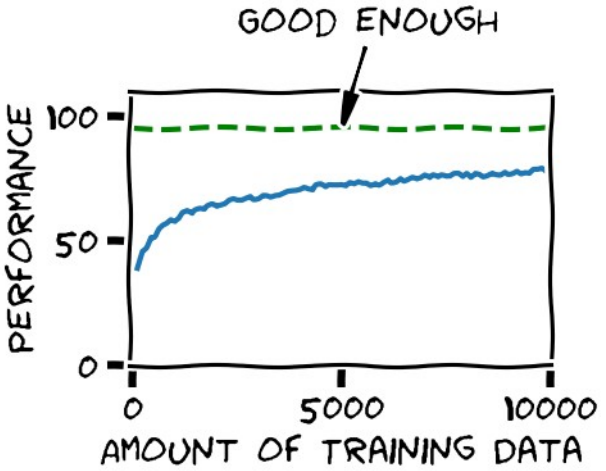
LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015) doi:10.1038/nature14539

“Shallow” Learning

- Decision Trees
- Random Forests
- Support Vector Machines
- Logistic Regression
- Three-layer Neural Networks
- Naive Bayes
- K-Nearest Neighbors
- Linear Discriminant Analysis
- ...

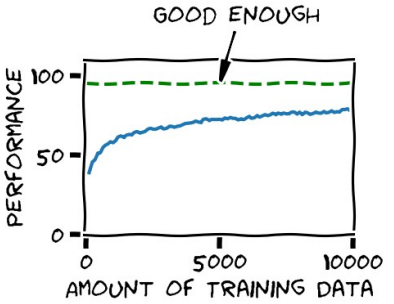
Shallow Learning Potential Problem #1

- Good news... More training data leads to higher accuracy:

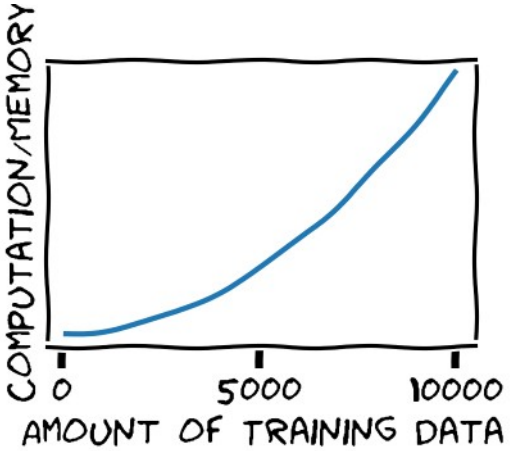


Shallow Learning Potential Problem #1

- Good news... More training data leads to higher accuracy:

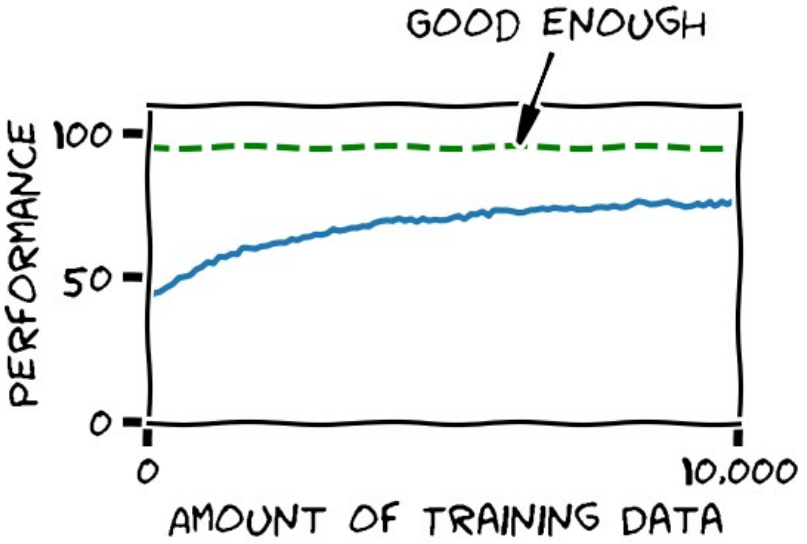


- Bad news... Algorithm doesn't scale:



Shallow Learning Potential Problem #2

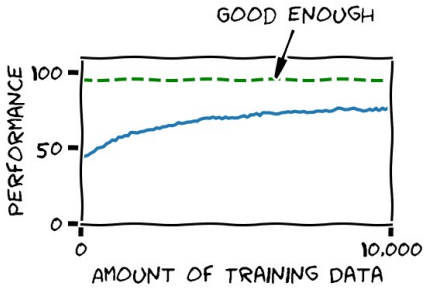
- Shallow algorithm that can handle massive training data:



- Promising! Let's try more data...

Shallow Learning Potential Problem #2

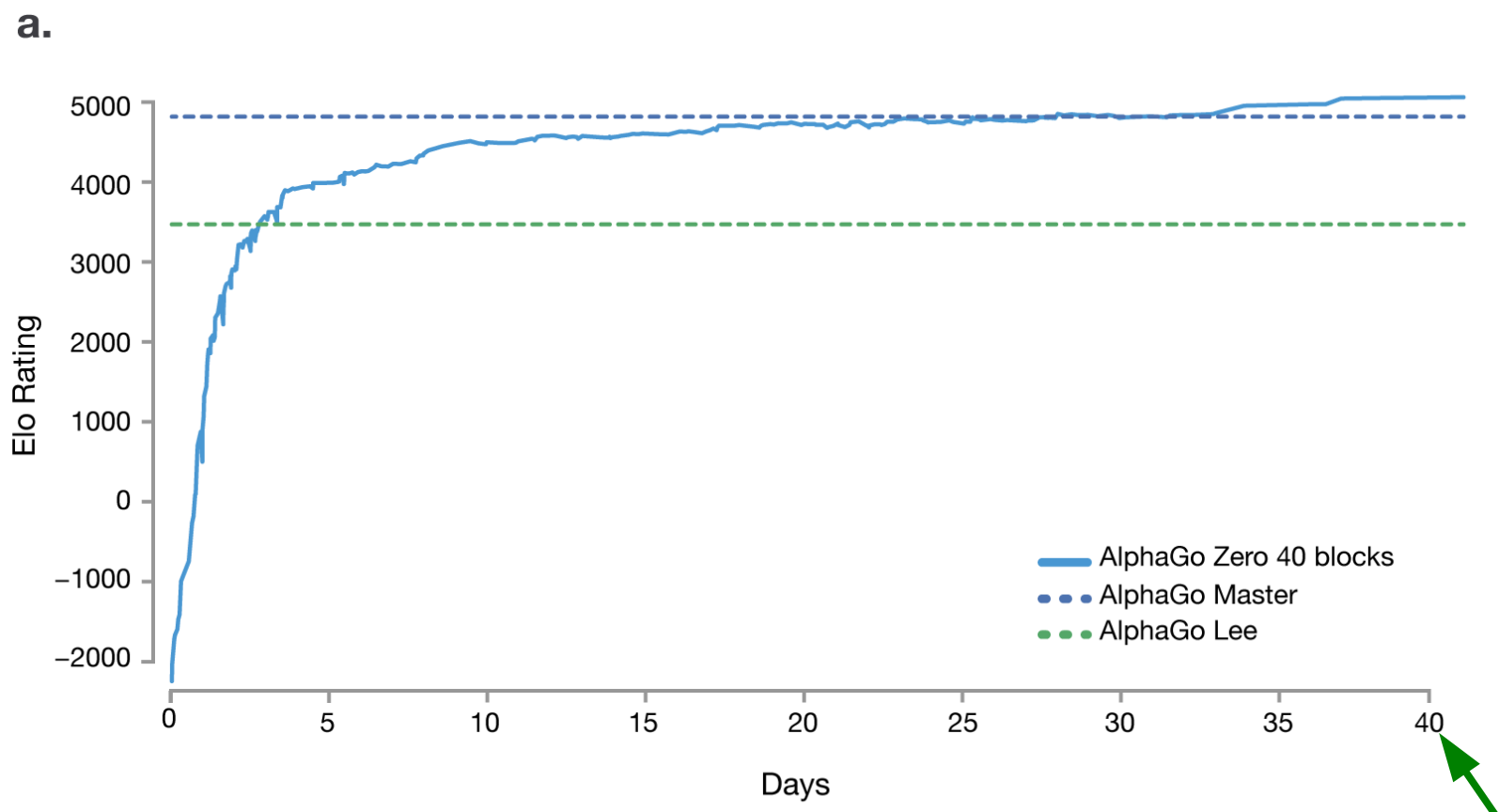
- Shallow algorithm that <u>can</u> handle massive training data:



- Promising! Let's try more data...

- Nope. Performance asymptote.

The Nice Thing About Deep Learning...

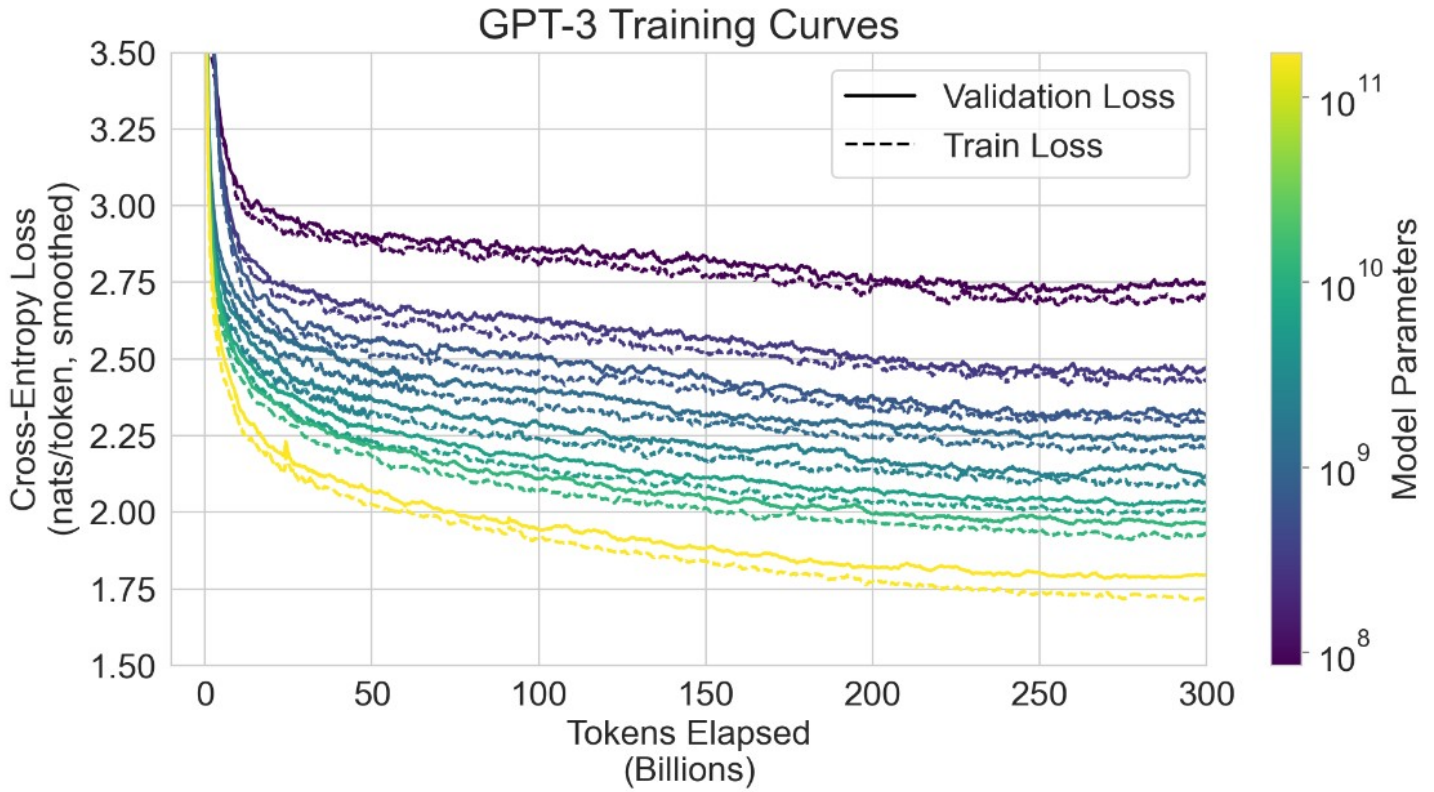


80+ Layers

6,000,000,000+ board positions

Silver, David, et al. "Mastering the game of Go without human knowledge." Nature 550.7676 (2017): 354.

GPT-3



Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020)

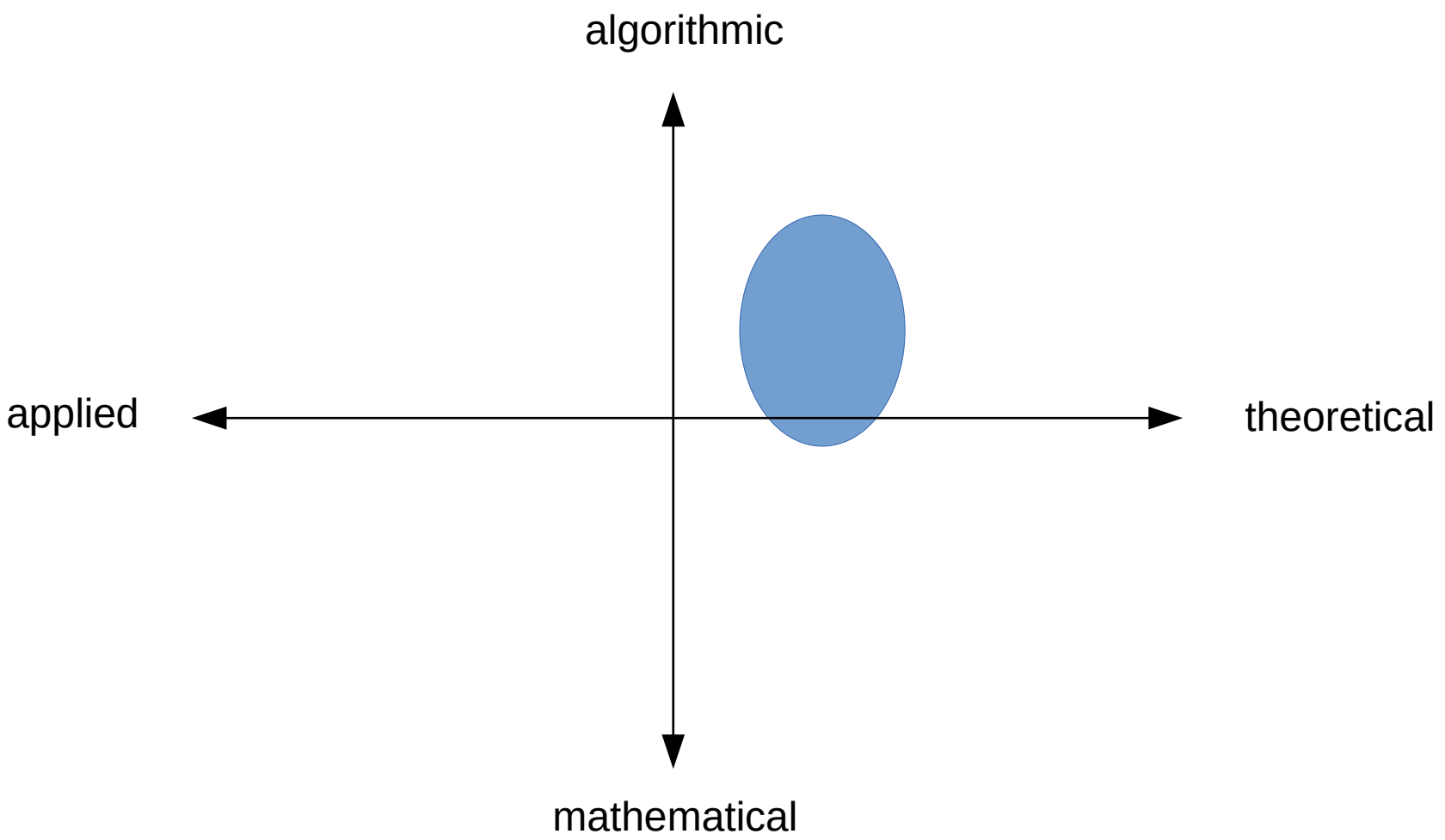
Buyer Beware! CS 445

- Math
- Python
- Independent initiative

Python

- Python is the most widely used language in machine learning (at least for research and education)
- Our toolset:
  - Anaconda
    - numpy
    - scikit-learn
    - PyTorch
    - Pandas

Course Emphasis



Course Logistics...

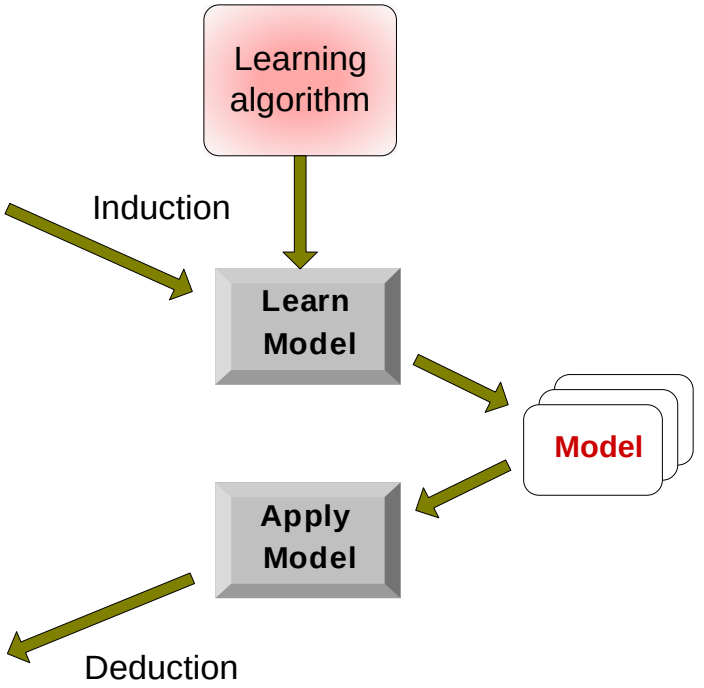
Classification/Data

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

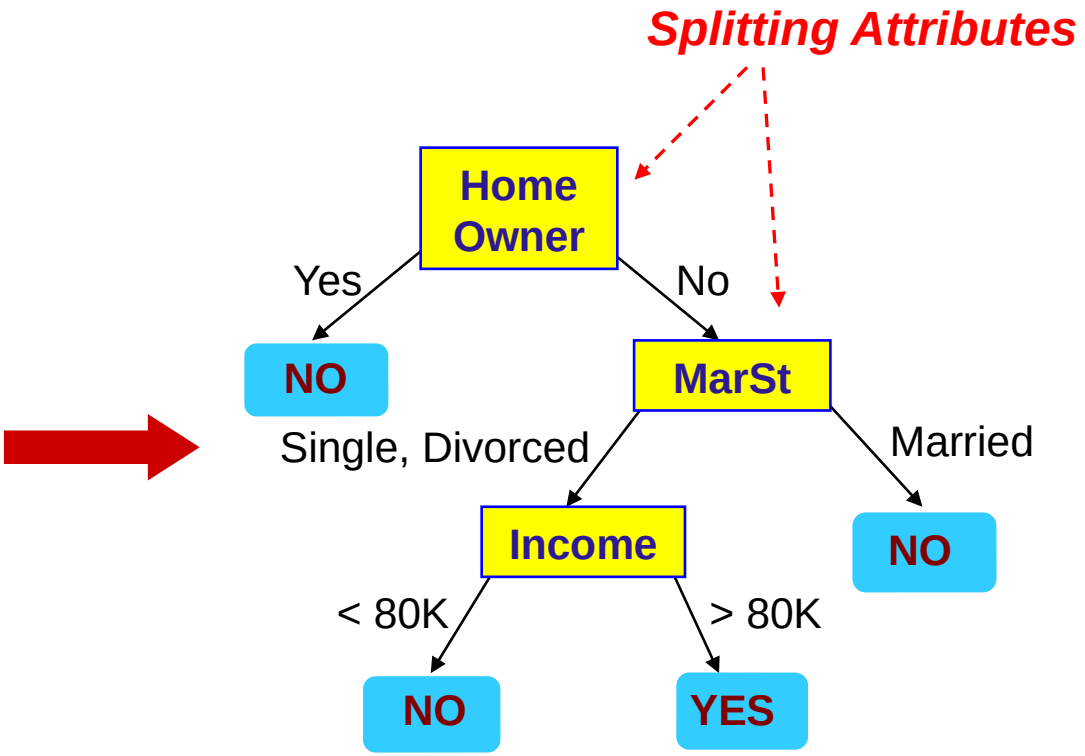


Decision Tree Classifier

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



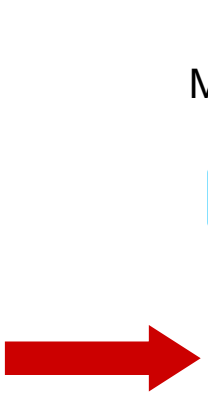
Model: Decision Tree

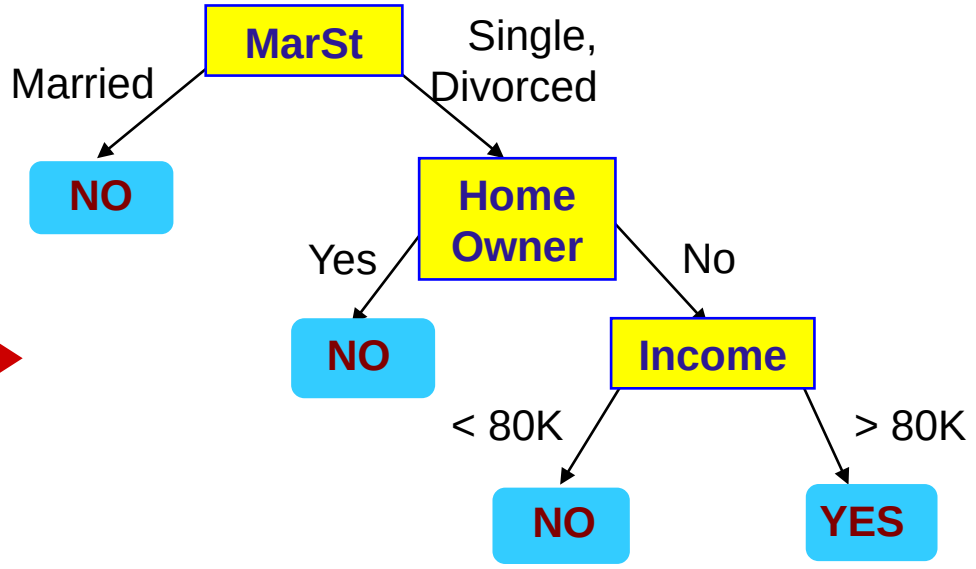
Decision Tree Classifier

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data





There could be more than one tree that fits the same data!

For Thursday

- Complete the posted reading
- Log into Canvas and complete the reading quiz and course survey
- Set up your Python environment
- Get Started on PA0