Applied Machine Learning with Scikit-Learn

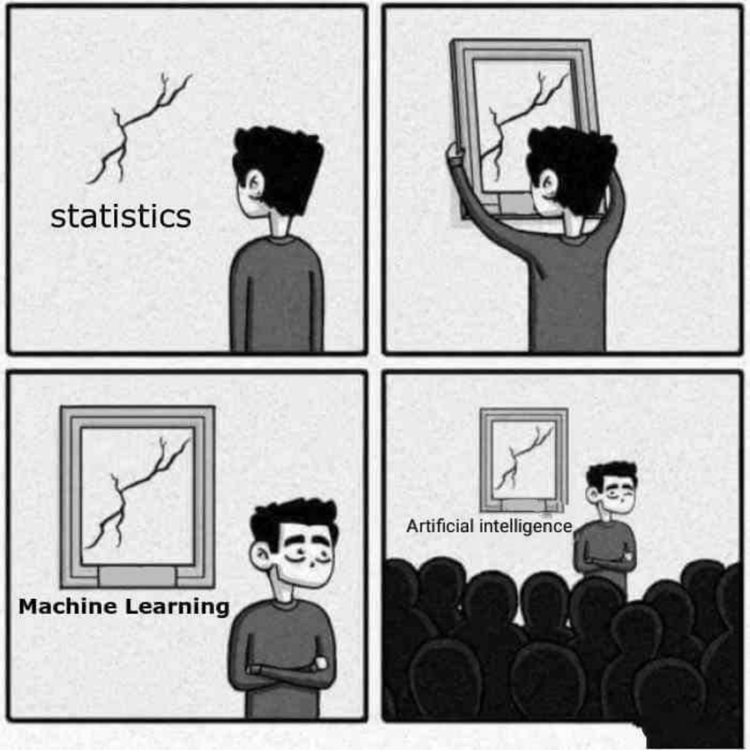
By Ted Petrou

Learning vs Machine Learning

- **Learning** is the process of acquiring new, or modifying existing, knowledge, behaviors, skills, values, or preferences.
  - The ability to improve on a certain task
  - Learning is possessed by humans, animals, and some machines
- **Machine Learning** is the ability of a machine to learn without explicitly being programmed to do so.
  - A branch of artificial intelligence
  - A model is built from data (inputs) to represent the real world
  - The model is trained by an algorithm to maximize some objective
  - The model can then be used in the future for predictions

Statistics vs Machine Learning

- Huge amount of overlap
- Statistics is more concerned with inference
  - Mathematical models to formalize understanding
  - Which variables have influence
  - Have confidence that relationship is true
- Machine learning cares more prediction
  - Validate on unseen data
  - Arose from computer science
  - Brute force, black box, bigger data
- Housing dataset example
  - Statistics - does age or income have a larger effect on whether someone has purchased a house
  - ML - Find model that will have highest accuracy for whether someone has bought a house

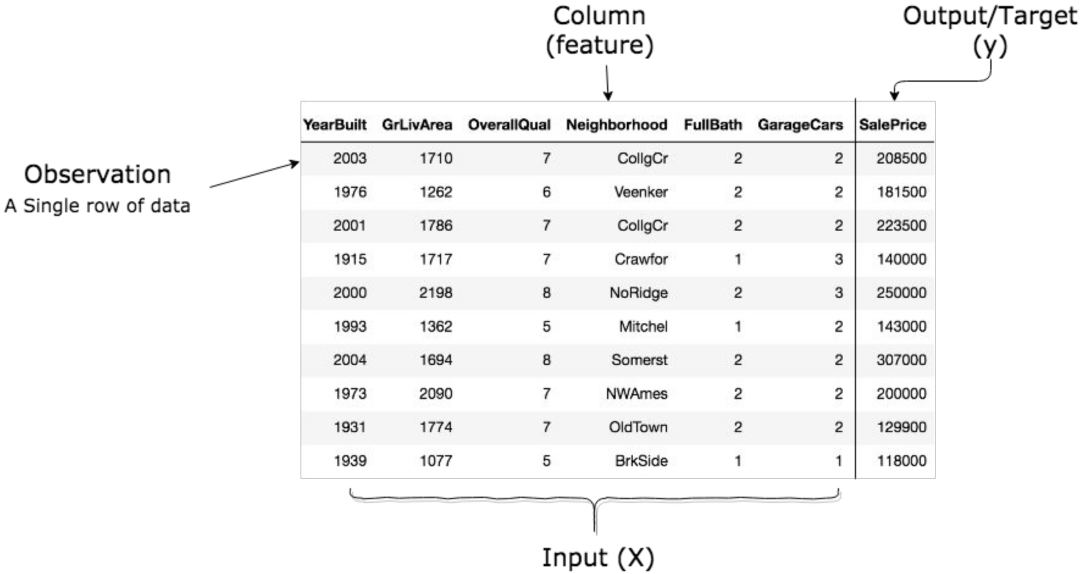


Types of Machine Learning

- Supervised Learning
    - All of the input data is labeled with an output. This output is considered the ground truth
    - Goal is to design a model to receive the input data and predict the output
    - Two types of supervised learning
        - Classification - Finite set of output labels - usually a word. (Is the image a cat or a dog?)
        - Regression - Output is a continuous value - always a number. (The final price of the house is $250k)
- Unsupervised Learning
    - The input data is not labeled with an output
    - Goal is to find inherent structure within the data
    - Used to cluster (group) similar observations together. (Classifying all species on Earth)

Supervised Learning - Regression

- Predicting the value of a house
- Supervised learning problem
  - regression output
  - 6 features
- Typically modeled with linear regression with regularization
- Output of model is house price



Column (feature)

Output/Target (y)

Observation

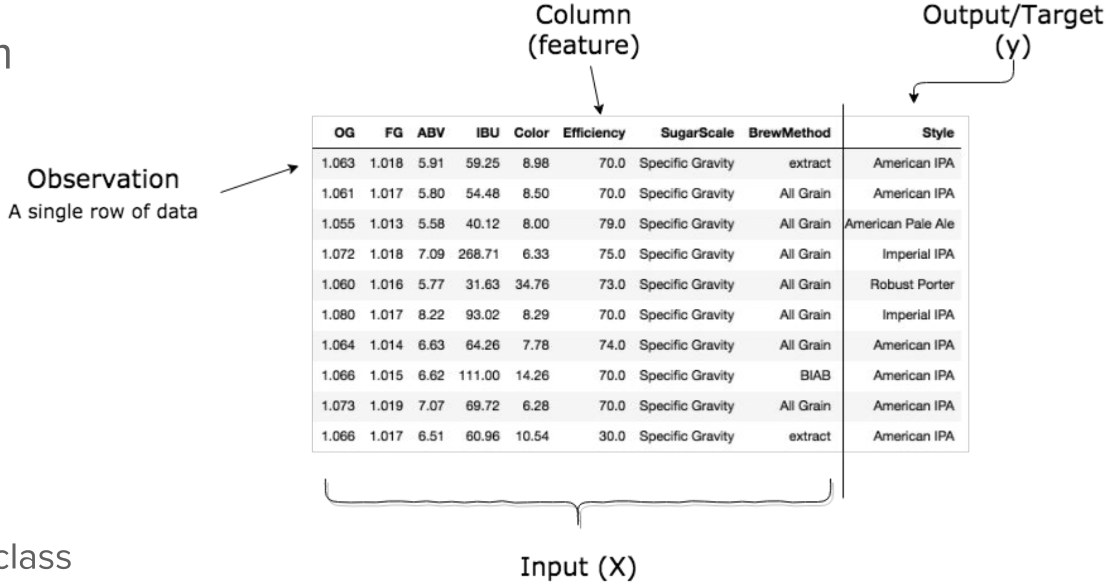
A Single row of data

YearBuilt	GrLivArea	OverallQual	Neighborhood	FullBath	GarageCars	SalePrice
2003	1710	7	CollgCr	2	2	208500
1976	1262	6	Veenker	2	2	181500
2001	1786	7	CollgCr	2	2	223500
1915	1717	7	Crawfor	1	3	140000
2000	2198	8	NoRidge	2	3	250000
1993	1362	5	Mitchel	1	2	143000
2004	1694	8	Somerst	2	2	307000
1973	2090	7	NWAmes	2	2	200000
1931	1774	7	OldTown	2	2	129900
1939	1077	5	BrkSide	1	1	118000

Input (X)

Supervised Learning - Classification

- Predicting the style of beer
- Supervised learning problem
  - Classification output
  - 8 features
- Common models include
  - Logistic Regression
  - Random Forests
  - Support Vector Machines
- Output of model
  - A single style
  - A probability of being in each class



Column (feature)

Output/Target (y)

Observation

A single row of data

OG	FG	ABV	IBU	Color	Efficiency	SugarScale	BrewMethod	Style
1.063	1.018	5.91	59.25	8.98	70.0	Specific Gravity	extract	American IPA
1.061	1.017	5.80	54.48	8.50	70.0	Specific Gravity	All Grain	American IPA
1.055	1.013	5.58	40.12	8.00	79.0	Specific Gravity	All Grain	American Pale Ale
1.072	1.018	7.09	268.71	6.33	75.0	Specific Gravity	All Grain	Imperial IPA
1.060	1.016	5.77	31.63	34.76	73.0	Specific Gravity	All Grain	Robust Porter
1.080	1.017	8.22	93.02	8.29	70.0	Specific Gravity	All Grain	Imperial IPA
1.064	1.014	6.63	64.26	7.78	74.0	Specific Gravity	All Grain	American IPA
1.066	1.015	6.62	111.00	14.26	70.0	Specific Gravity	BIAB	American IPA
1.073	1.019	7.07	69.72	6.28	70.0	Specific Gravity	All Grain	American IPA
1.066	1.017	6.51	60.96	10.54	30.0	Specific Gravity	extract	American IPA

Input (X)

All machine learning problems

- Must have Data
- Identify the input data
- Identify the target column (label) for each observation if it exists
- Identify the type of target (regression or classification)

How to start doing machine learning

- Must first select a model
  - Linear regression, support vector machine, neural network, etc...
- Train the model on historical data
- Once model is trained, it can make predictions