### Plan for the week

- M: Simple Regression
  - Ordinary Least Squares (OLS)
- W: Regression in Practice
  - Regression Diagnostics
- F: Properties of Estimators
  - OLS Estimators are BLUE: Best Linear Unbiased Estimators

# What is Machine Learning?

### What is Machine Learning?

# Machine learning "gives computers the ability to learn without being explicitly programmed"

- Arthur Samuel

	Supervised learning	Unsupervised learning	UNSUPERVISED	SUPERVISED
Discrete	Classification	Clustering		
Continuous	Regression	Density Estimation		Class 1 Class 2 Undrome

### Supervised Learning (a.k.a. Function approximation)

- Learning from labeled data
- This labeled data is called training data, and consists of observations and corresponding labels (*a.k.a.* ground truth).
- The goal of a supervised learning algorithm is to approximate a function that generalizes well to unseen examples.
- Supervised algorithms are typically evaluated on test data, which is usually held out, and thus distinct from training data.

### **Supervised Learning**

Given a labeled set of training data  $D = \{(x_i, y_i) \mid i = 1, ..., n\}$ , where each  $x_i$  is an instance and each  $y_i$  is a label, the goal of a supervised learner is to learn a function from instances to labels, so that it can appropriately label instances in the test data



### **Supervised Learning**

• Classification is a supervised learning problem

The goal of a classifier is to classify:
i.e., assign a category to each observation.

- Regression is also supervised learning problem
  - The labels in a regression problem are quantitative, while the labels in classification are categorical

The goal of supervised learning is generalization: to learn a general rule from many specific examples (in-sample data) that works well when applied to new examples (out-of-sample data).

## Supervised Learning CLASSIFICATION VS REGRESSION



- Supervised learning when labels are categorical
- **Binary** classification: Spam or not? Malignant or benign?
- Multiclass classification: Part-of-speech tagging, Face recognition, Spell checking.

#### If it Walks/Swims/Quacks Like a Duck ..... Then It Must Be a Duck







### **Classification Algorithms**

- There are many, many, many classification algorithms (and many regression algorithms, as well)
  - <u>https://en.wikipedia.org/wiki/Category:Classification\_algorithms</u>
- Why? Because there is no "best" algorithm
- Popular algorithms:
  - Logistic Regression
  - Naive Bayes
  - Decision trees (easily adopted to regression, as well)
  - *k* nearest neighbor (easily adopted to regression, as well)

### Regression

- Supervised learning when labels are continuous
- Earliest methods date back to Legendre and Gauss (1800s)
- Classical regression:
  - A statistical method to model relationships between variables
  - E.g., linear regression involves finding a line that "best" fits data
- Bayesian regression: still an area of active research

### Unsupervised learning

- Unsupervised learning finds hidden structure in unlabeled data
- Canonical discrete application is clustering: i.e., grouping objects based on their similarity to one another
  - Popular algorithms:
    - k means, k medians (Expectation-Maximization)
    - Hierarchical clustering
- Canonical continuous application is density estimation: i.e., estimating the probability distribution of a data set
  - Its skewness, (the number and location of) its modes, its tail behavior, etc.

### **Unsupervised** learning





#### **Density Estimation**

### Slovakian Youth Survey Research Q's

- Classification
- Regression
- Clustering (Dimensionality Reduction)
- Density Estimation

### Slovakian Youth Survey Research Q's

- Classification: Can we classify individuals as healthy or not?
- **Regression**: Can we predict spending habits from personality traits, etc.?
- Clustering: Can we cluster youth by their smoking and drinking habits? (Can we characterize people using only a small number of variables?)
- Density Estimation: Can we characterize how youth spend their free time?