Introduction to Machine Learning

Shaobo Li

University of Kansas

You should take this course if...

- You like playing with data
- You want to be a data scientist
- You want to proceed an advanced degree in data science
- You live in the 21st century

An article on who should purse a master's degree in BA.

How much salary can a data scientist earn? (a report from indeed)

Who are looking for people with DM and ML skills...

Almost all industries.

What is Data Mining and Machine Learning?

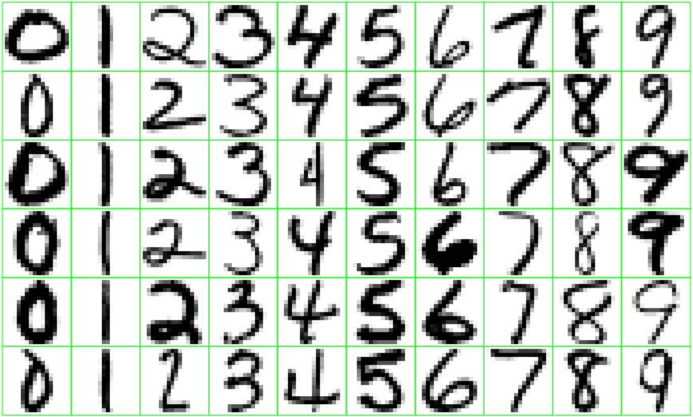
- Data mining focuses on discovering patterns and relationships in a given data
- Machine learning focuses on training models and predicting future
- A large overlap, but have different focuses
- No need to distinguish them conceptually for our course
- A good article to read

Machine Learning in Different Fields

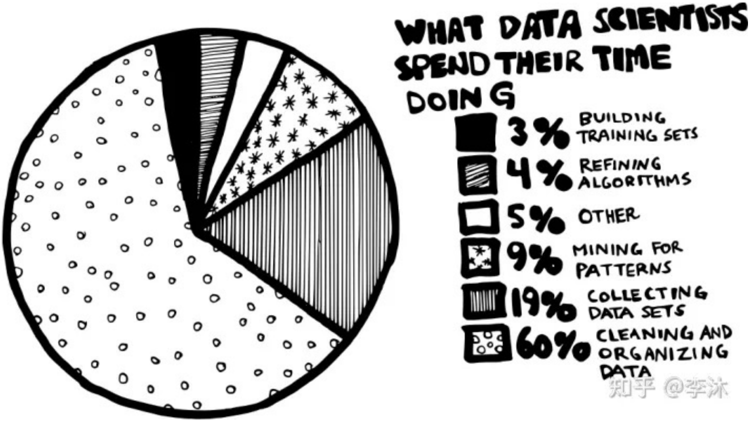
Learning from data is essential in different scientific disciplines

- Predict stock returns in next six months based on historical data;
- Predict the probability of a loan default based on customer's information and historical records;
- Identify certain diseases based on medical image;
- Identify handwritten digits from image;
- Facial recognition;
- Natural language processing;
- Cluster customers based on their purchase behavior and other information

Examples



The reality


WHAT DATA SCIENTISTS SPEND THEIR TIME DOING
3% BUILDING TRAINING SETS
4% REFINING ALGORITHMS
5% OTHER
9% MINING FOR PATTERNS
19% COLLECTING DATA SETS
60% CLEANING AND ORGANIZING DATA
知乎 @李沐

Learning Types

- **Supervised learning**
  - There is specific response you need to predict
- **Unsupervised learning**
  - No response, instead, you need to create response based on some patterns
- **Semi-supervised learning**
  - Mixture of both

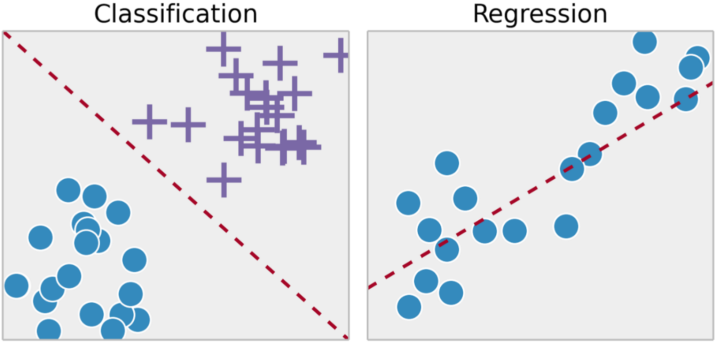
Supervised Learning

- Suppose we observe data Y_i and $\mathbf{x}_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$.
- Y is the outcome (or response, dependent variable, target), and $\mathbf{x}$ is predictor (or independent variables, covariates, features, inputs)
- The learning problem can be modeled as

$$Y_i = f(\mathbf{x}_i) + \epsilon_i$$

where $f(\cdot)$ is unknown function, and ϵ is random error.

Regression and Classification



Regression and Classification

Regression:

- Response variable is continuous
- e.g., stock return, housing price, temperature

Classification:

- Response variable is categorical
- e.g., $\{A, B, C\}$, $\{dog, cat\}$, $\{0, 1\}$
- an example of neural networks: [link]

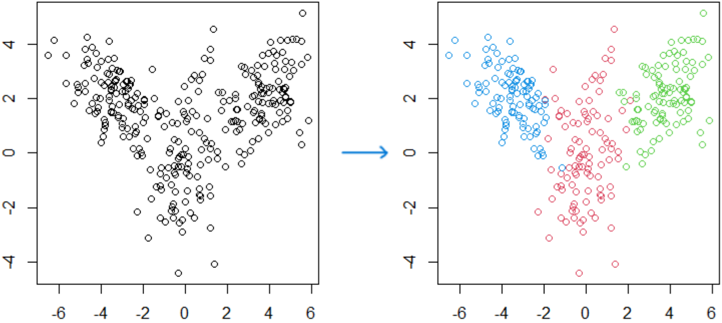
Learning methods:

- Parametric methods, e.g., maximum likelihood
- Nonparametric methods, e.g., decision tree, neural network

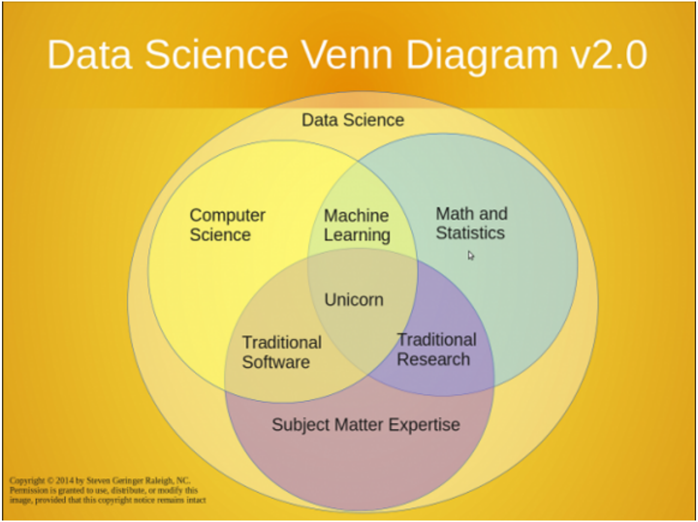
Unsupervised Learning

- Data is **unlabeled** (no Y's)
- Uncover patterns, groups among X's
- Subjective, no simple goal such as prediction
- Examples: Recommendation systems, clustering, principle component analysis (PCA), association rules

K-means Clustering



In practice, you need...



Computer programming

- **R** – we will learn R in this class
  Download R, and install
  Download RStudio, and install
- **Python** – explore by yourself
  To start, download and install Anaconda
  Get started with Anaconda

Learning Resources

- Data
  - Most commonly used public data sets
  - Textbook data (James, et al.): install R package `ISLR`
  - UCI Machine Learning Repository
  - Kaggle
  - KDD Nuggets
- Lectures and other tutorials
  - Videos of textbook (ISLR)
  - DataCamp
  - Coursera