UCL AI SOC
The Future is Calling

UCL AI Society – Tutorials for ML
Live Session 1

Introduction to Machine Learning

By Kamen Brestnichki,

November 5th, 2020

CONTENTS

01

Introduction to Machine Learning

From Statistics to Artificial Intelligence

A Brief History

- Statistical Analysis: early 20 century
- Data Mining: 1950s
- Machine Learning: 1980-1990 booming, to present
- Big Data: 2008
- Artificial Intelligence: 1956-present

A Brief History

- Statistical Analysis

- Random variable (independent, joint), statistics of random variable: x, categorical, continuous, expected value, variance ...

- Probability

- Distribution, z-score, confidence interval: Gaussian, Poisson...

- Statistical modeling: simple linear regression, multiple regression

- Bayes' theorem

A Brief History

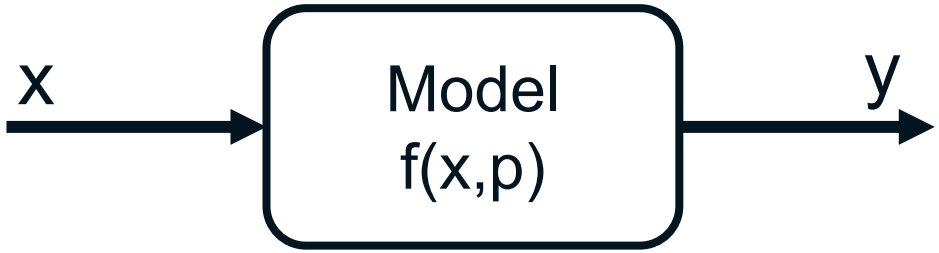
- Data Mining/ Machine Learning/ Big Data

  - Step 1: Learning (Training)

  x,y known, estimate *p'*

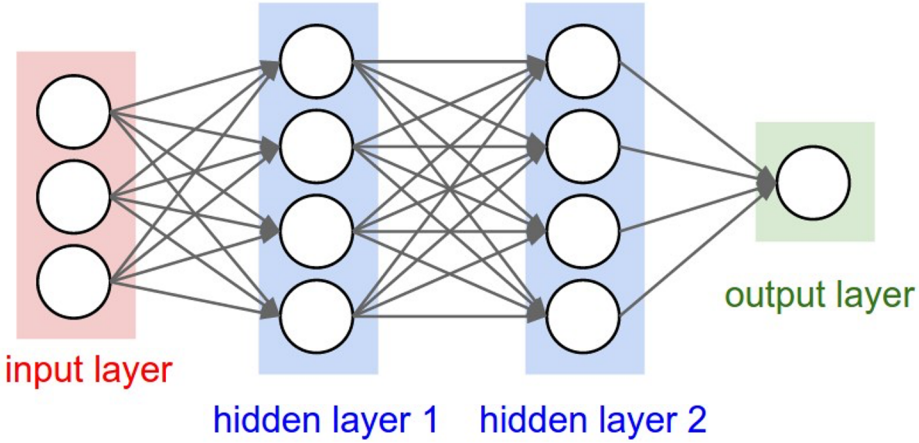
  - Step 2: Predicting

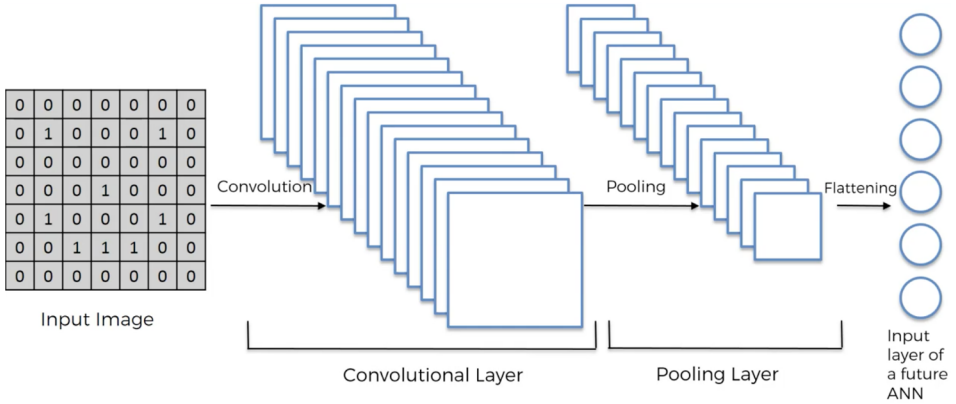
  x, p known, predict *y*



A Brief History

- Artificial Intelligence (ANN-based)



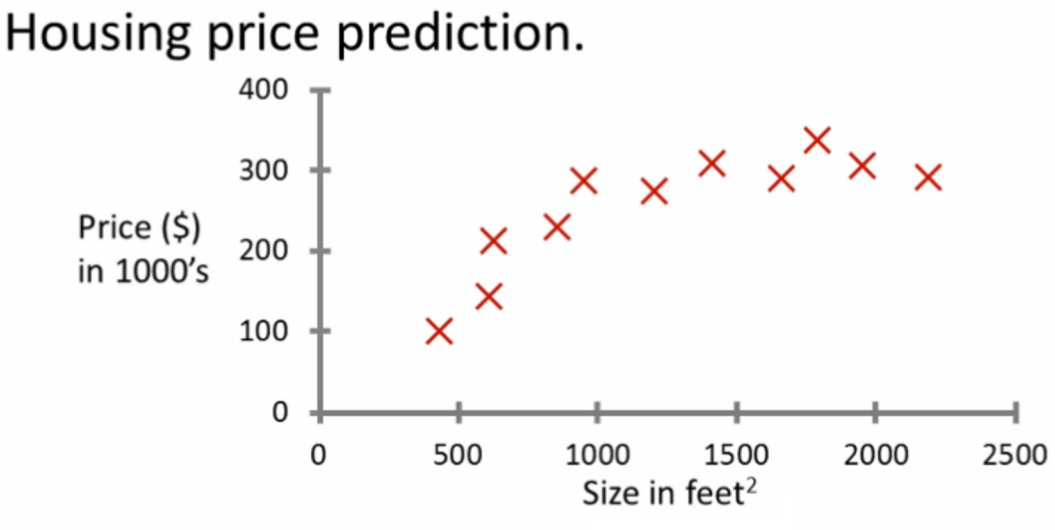


The Process of Machine Learning

- Understand the data (e.g., categorical/numerical, distribution)
- Split data: training vs validation
- Define performance metrics (e.g., mean squared error, accuracy)
- Build candidate models (e.g. linear regression)
- Train models
- Evaluate models
- Explain models and perform model selection
- Use the best model to predict!

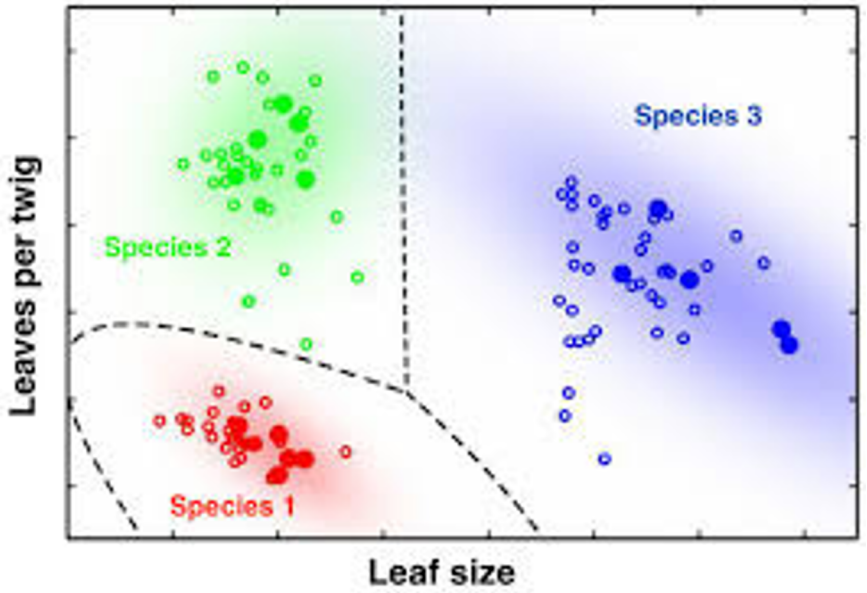
Four Types of Problems to Solve (Supervised/Unsupervised)

- Regression
- Classification
- Clustering
- Causality



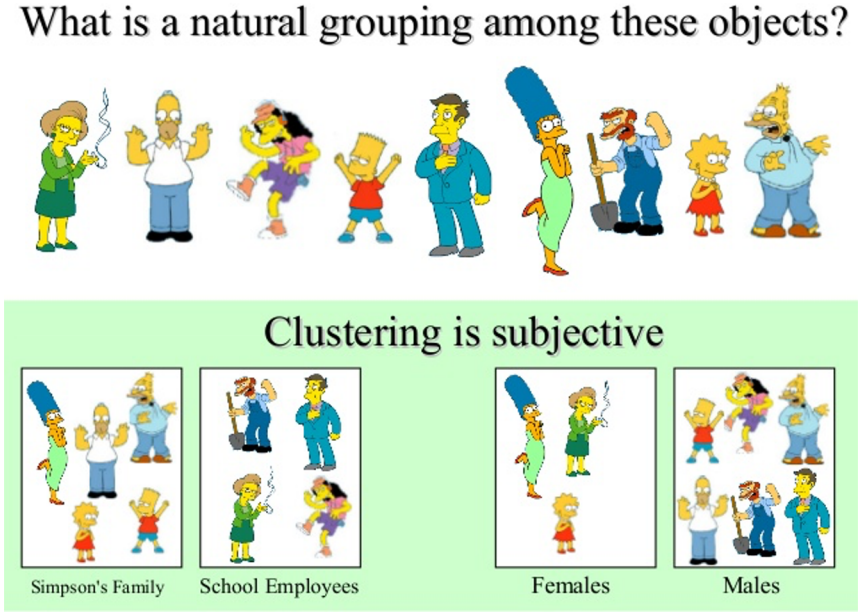
Four Types of Problems to Solve (Supervised/Unsupervised)

- Regression
- Classification
- Clustering
- Causality



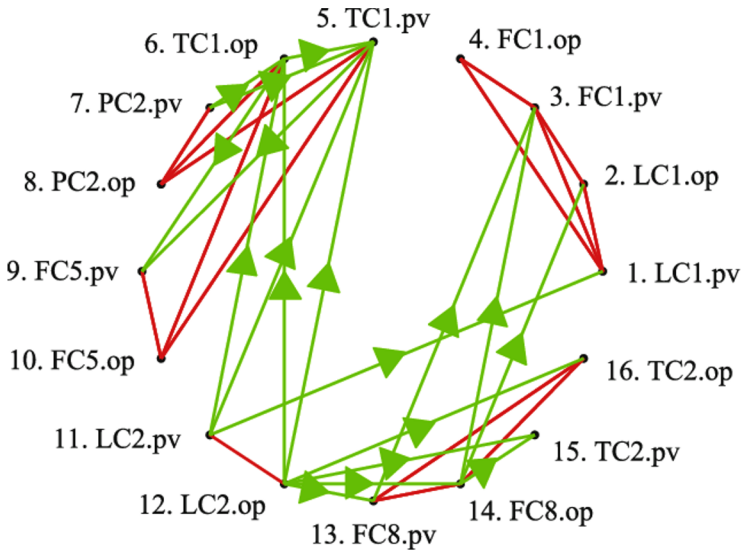
Four Types of Problems to Solve (Supervised/Unsupervised)

- Regression
- Classification
- Clustering
- Causality



Four Types of Problems to Solve (Supervised/Unsupervised)

- Regression
- Classification
- Clustering
- Causality



02

Data Pre-processing

One of the Most Important Steps in ML

Data Pre-processing - Rescaling

- Standardization
  - Centering: Subtract mean
  - Scaling: Divide by its standard deviation

$$y_i = \frac{x_i - \bar{x}}{s}$$

- Normalization
  - Centering: Subtract min
  - Scaling: Divide by [max-min]

$$y_i = \frac{x_i - \min_{1 \le j \le n} \{x_j\}}{\max_{1 \le j \le n} \{x_j\} - \min_{1 \le j \le n} \{x_j\}}$$

Data Pre-processing - Resolve Skewness

- Definition:

$$skewness = \frac{\sum (x_i - \bar{x})^3}{(n-1)v^{3/2}}, v = \frac{\sum (x_i - \bar{x})^2}{(n-1)}$$

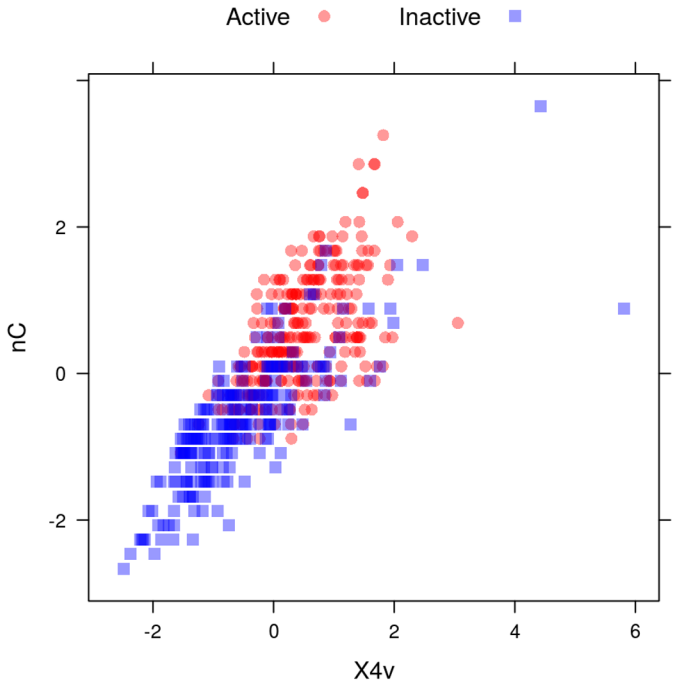
- Resolve:

$$x = \begin{cases} \frac{x^{\lambda} - 1}{\lambda}, if\ \lambda \neq 0 \\ \log(x), if\ \lambda = 0 \end{cases}$$

Ref. - Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

Data Pre-processing - Resolve Outliers

- Remove
- Spatial sign transformation



Data Pre-processing - Resolve missing value

- Various Techniques:
  - Random number
  - Removing predictors
  - mean value of the feature
  - Interpolation
  - Advanced: multiple imputation

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

Data Pre-processing - Transformation

- From continuous to discrete
  - Artificial division
  - Histogram
  - Information Gain
- From non-numerical to numerical
  - One-hot encoding
  - Embedding

Data Pre-processing - Transformation

- One-hot encoding V.S Embedding

```
books = ["War and Peace", "Anna Karenina",
         "The Hitchhiker's Guide to the Galaxy"]
books_encoded = [[1, 0, 0],
                 [0, 1, 0],
                 [0, 0, 1]]
```

```
books = ["War and Peace", "Anna Karenina",
         "The Hitchhiker's Guide to the Galaxy"]
books_encoded_ideal = [[0.53,  0.85],
                       [0.60,  0.80],
                       [-0.78, -0.62]]
```

03

Mathematical Fundamentals

Linear algebra | Probability

Linear Algebra

- A succinct way to represent operations on many quantities
  - E.g. $\sum_{n=1}^{N} w_n x_n = w_1 x_1 + w_2 x_2 + \ldots + w_N x_N$ $\quad \mathbf{w}^T \mathbf{x} = \mathbf{w} \cdot \mathbf{x}$
- Core components are vectors $\mathbf{x}$ and matrices A
  - Each has associated operations we can perform on them

Vectors

- A stack of D numbers $\mathbf{w} = \overrightarrow{w} = [w_1, \ldots, w_D]^T$
  - Each number w_d is real (continuous), so we say $\mathbf{w} \in \mathcal{R}^D$
  - Each vector has a magnitude $||\mathbf{w}||_2 = \sqrt{w_1^2 + w_2^2 + \ldots + w_D^2}$

Matrices 1

- A stack of N vectors $A = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N]$
  - Each element $a_{nd} = (\mathbf{w}_n)_d$ is real, so we say $A \in R^{N \times D}$
  - Matrices can be multiplied together

$$\begin{matrix} & \vec{b_1} & \vec{b_2} \\ & \downarrow & \downarrow \\ \vec{a_1} \rightarrow & 1 & 7 \\ \vec{a_2} \rightarrow & 2 & 4 \end{matrix} \cdot \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} \vec{a_1} \cdot \vec{b_1} & \vec{a_1} \cdot \vec{b_2} \\ \vec{a_2} \cdot \vec{b_1} & \vec{a_2} \cdot \vec{b_2} \end{bmatrix}$$

$A \qquad B \qquad C$

Matrices 2

- There are special matrices
  - Identity matrix $I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ which means $\forall A \in R^{D \times 3}, AI = A$
  - Matrix inverse A^{-1} such that $A^{-1}A = AA^{-1} = I$
    - ❑ This requires $A \in R^{D \times D}$ and **invertible**

Probability 1

- Describes the randomness of events

  - If I flip a coin X, $P(X = H) = P(X = T) = \frac{1}{2}$

- Probabilities have to follow rules

  - All probabilities are non-negative

  - The sum of probability of all outcomes is 1

Probability 2

- Conditional probability

  - Captures the idea $P(\text{rains}|\text{clear sky}) = 0$

    - Thought of as “updating our belief due to prior knowledge” $P(A|B)P(B) = P(A \text{ and } B)$

- Product rule

Probability 3

- Bayes' Theorem
  - Derived from product rule $P(A|B)P(B) = P(B|A)P(A)$
    - Gives us a way to represent uncertainty about our model

$$\underbrace{P(\text{parameters}|\text{data})}_{posterior} = \frac{\overbrace{P(\text{data}|\text{parameters})}^{likelihood}\overbrace{P(\text{parameters})}^{prior}}{\underbrace{P(\text{data})}_{model\ \ evidence}}$$

04

Linear Regression

The “Hello World” of Machine Learning

Problem setup

- I have data $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathcal{R}^{N \times D}$ and quantities of interest $\mathbf{y} = [y_1, y_2, ..., y_N] \in \mathcal{R}^N$
- I want to learn a function $f_W(\mathbf{x}) = W\mathbf{x} + \mathbf{b}$, such that $f_W(\mathbf{x}_i) \approx y_i$, in the hopes that given a new datapoint $\mathbf{x}_*$, I can predict what y_* will be

Example

- I know that **2** apples and **1** banana weigh **300** grams and **3** apples and **3** bananas weigh **500** grams.

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \end{bmatrix} = \begin{bmatrix} [2 & 1] \\ [3 & 3] \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 300 \\ 500 \end{bmatrix}$$

- How much do **5** apples and **3** bananas weigh?

$$\mathbf{x}_*^T = [5\ 3], y_* = ?$$

How to solve the above problem using linear regression will be shown in the notebooks

Thank you!