Introduction

Wednesday, August 25, 2021 6:02 PM

- Overview
  - Overall approach to understanding data
    - Supervised
    - Unsupervised

Supervised

X - Inputs
Y - Outputs } (Linear?) System

Unsupervised

X - Inherent/Latent structure

$\boxed{\text{* Tukey}}$

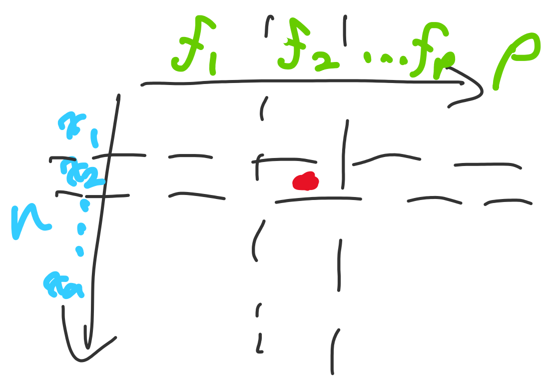
- History
  - Gauss : OLS → Linear Regression
  - Linear Discriminant Analysis (LDA) → 1930s
    - Classification
  - Logistic Regression → 1940s
    - Classification & Regression
  - Generalized Linear Models (GLM) → 1970s
    - Linear & Logistic & Others
  - Neural Networks → 1980s
    - Linear / Non-Linear
  - Support Vector Machines → 1990s
    - Non-Linear Classification

~2000s + → Deep Learning & Neural Networks
- CNN / RNN

- Themes
  - Statistical Learning is more relevant to applied areas beyond traditional stats/ml.
  - SL should not be used for black-box modeling.
  - Statistical Learning does not require deep methodical knowledge of proofs in order to be applied.
  - SL has to be applied to real world problems in order to be truly relevant.

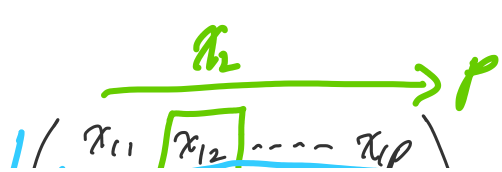
- Notation
  - Data in n observations, p dimensions

* Data Frame



* Heterogeneous Types!

* Matrix



$$X = \begin{array}{c} n \\ \downarrow \end{array} \begin{pmatrix} \boxed{x_{21} \quad x_{22} \quad \cdots \quad x_{2p}} \\ \vdots \qquad\qquad\qquad \vdots \\ x_{n1} \quad x_{n2} \quad \cdots \quad x_{np} \end{pmatrix}$$

* Homogeneous Types $\Rightarrow$ Real Numbers OR Integers (maybe)

$x_i \in R^p$ Row

$x_j \in R^n$ Col

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

$$\therefore \quad X = \left(\vec{x}_1 \quad \vec{x}_2 \quad \cdots \quad \vec{x}_p \right)$$

OR

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$