Introduction to Statistical Learning

Chapter 1 : Introduction.

doubt :: Where to use KNN? Classify KNN? in short.

stepwise, ridge, PCA, LASSO.
↑ Non linear.

doubt: KNN when to classify?

Statistical learning
(Tools to understand data)
→ set of approaches for estimating f.
→ minimize reducible error

linear ← Regression

logistic regression

Classification

Supervised

Unsupervised.
→ Clustering
→ K Means
→ Hierarchical
→ PCA.

$X \rightarrow |f(\cdot)| \rightarrow Y$.

s.t. $\hat{Y}$ & Y are close.

Find relationship, patterns & structures in data.
(no 'Y' involved).

linear discriminant analysis.

natural output is available.

Supervised Learning.

Unsupervised Learning. → natural output not available

(↳ clustering problem. (types of customers).

X \ Y	continuous quantitative	categorical qualitative
continuous/ quantitative	Regression (Wage)	Classification
doubt: categorical.	?	?

- No single approach will perform well in all possible applications.
→ model, intuition, assumption, trade off

Notation:

n := sample size.

p := no. of features.

x_{ij} := value of j^{th} variable for i^{th} observation.

$\Rightarrow i = 1(1)n \quad j = 1(1)p$.

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$x_i = \underset{(p \times 1)}{\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}} \quad \text{(column vector)}$$

Choosing best method:
① cross validation
② bootstrap.

data: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$