Lecture 1: Overview of Supervised Learning

Aiying Zhang

January 12

Terminology and Variable Types

- **Inputs**: predictors, features
- **Outputs**: responses
- Variable types:
  - Quantitative (measurements)
  - Qualitative / categorical
    - Discrete
    - Ordered categorical
- Output types:
  - Regression $\rightarrow$ quantitative
  - Classification $\rightarrow$ categorical

Two Simple Approaches to Prediction

- Least Squares (global methods)
- Nearest Neighbors (local methods)

These two ideas underlie many modern learning algorithms.

Linear Models and Least Squares

Model:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} X_j \beta_j = \boldsymbol{x}^T \boldsymbol{\beta}$$

Least Squares objective:

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

Solution (if $X^T X$ is nonsingular):

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

- Strong assumptions
- Low variance, potentially high bias

Nearest Neighbor Methods

k-NN regression:

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

- $N_k(x)$: neighborhood of k closest points
- Effective number of parameters $\approx N/k$
- Adaptive, minimal assumptions
- Low bias, high variance

Extensions of These Ideas

Many popular methods are variants of:

- Least squares
- Nearest neighbors

Examples:

- Kernel methods
- Local regression
- Linear models with basis expansion
- Projection pursuit
- Neural networks

Statistical Decision Theory

Let:

$$X \in \mathbb{R}^p, \quad Y \in \mathbb{R}$$

Loss function (squared error):

$$L(Y, f(X)) = (Y - f(X))^2$$

Expected Prediction Error (EPE):

$$\text{EPE}(f) = \mathbb{E}(Y - f(X))^2$$

Optimal Predictor

Minimizing EPE pointwise yields:

$$f(x) = \mathbb{E}(Y \mid X = x)$$

- Conditional expectation = regression function
- Fundamental target of regression methods

k-NN as Approximation

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Under regularity conditions:

$$k \to \infty, \quad \frac{k}{N} \to 0$$

$$\hat{f}(x) \to \mathbb{E}(Y \mid X = x)$$

- Curse of dimensionality slows convergence

Classification Problems

Output:

$$G \in \{G_1, \ldots, G_K\}$$

0–1 loss:

$$L(G, \hat{G}) = \mathbb{I}(G \neq \hat{G})$$

Bayes classifier:

$$\hat{G}(x) = \arg\max_g \Pr(G = g \mid X = x)$$

Statistical Models

Additive error model:

$$Y = f(X) + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0$$

- $f(X)$: systematic component
- ε: noise, unmeasured variables

Extensions:

- Nonconstant variance
- Generalized linear models

Function Approximation View

Data:

$$(x_i, y_i) \in \mathbb{R}^{p+1}$$

Examples:

- Linear models: $f(x) = \boldsymbol{x}^T \beta$
- Basis expansions:

$$f(x) = \sum_{k=1}^{K} h_k(x)\theta_k$$

Estimated by:

- Least squares
- Maximum likelihood

Curse of Dimensionality

- Local neighborhoods become sparse in high dimensions
- k-NN requires exponentially more data
- Structured models improve efficiency

Classes of Restricted Estimators

To obtain useful solutions with finite data:

- Roughness penalties / Bayesian methods

$$\text{PRSS}(f) = \text{RSS}(f) + \lambda J(f)$$

- Kernel methods and local regression
- Basis function and dictionary methods