

2 Overview of Supervised Learning

$X \in \mathbb{R}^p$: Random vector input

$Y \in \mathbb{R}$: Random variable to predict

$f(X)$: function we seek to predict Y

$L(Y, f(X))$: *Loss function* to penalize error in prediction

$EPE(f) = \mathbb{E}[L(Y, f(X))]$: Expected Prediction Error

Square error loss (or L_2 *loss function*): $L(Y, f(X)) = (Y - f(X))^2$

In this case, the sol is $f(x) = \mathbb{E}[Y|X = x]$: *regression function*

Curse of dimension: if $p \uparrow$, local methods fail because density $O(N^{1/p})$

Mean Squared Error of the estimator \hat{f} at x_0 : $MSE(\hat{y}_0) := \mathbb{E}_\tau[(y_0 - \hat{y}_0)^2]$

τ (training set) is random, so is $\hat{y}_0 := \hat{f}(x_0)$. True $y_0 := f(x_0)$ is fixed

bias-variance decomposition: $MSE(\hat{y}_0) = Var_\tau(\hat{y}_0) + Bias(\hat{y}_0)^2$

Additive error model : $Y = f(X) + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $X \perp \epsilon$

We model f with parameters θ and try to determine f_θ

In *least squares* method, θ chosen to minimize *Residual Sum of Squares*

$RSS(\theta) := \sum_{i=1}^N (y_i - f_\theta(x_i))^2$

Maximum likelihood estimation: maximize \mathbb{P} to have those observations

$\theta = \operatorname{argmax}_\theta L(\theta) := \sum_{i=1}^N \log \mathbb{P}_\theta(Y = y_i | X = x_i)$

Least square=Max likelihood if $Y = f_\theta(X) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Penalized RSS: $PRSS(f, \lambda) := RSS(f) + \lambda J(f)$, J *roughness penalty*

If output is a categorical variable $G \in \mathcal{G}$, L is a $\operatorname{card}(\mathcal{G})$ -square matrix

zero-one loss function: $L(G, \hat{G}(X)) = 1_{G \neq \hat{G}(X)}$: (f noted \hat{G})

In this case, $\hat{G}(x) = \operatorname{argmax}_{g \in \mathcal{G}} \mathbb{P}(G = g | X = x)$, called *bayes classifier*

3 Linear Methods for Regression

3.2 Linear Regression and Least Squares

$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$

$RSS(\beta) = \sum_{i=1}^N (y_i - (\beta_0 + \sum_{j=1}^p X_j \beta_j))^2 = (y - X\beta)^T (y - X\beta)$

it's convex, solving $\frac{\partial RSS}{\partial \beta} = 0$, we get $\hat{\beta} = (X^T X)^{-1} X^T y$

Alternatively, we note that X generate a p -dim subspace of \mathbb{R}^N , so to minimize $RSS(\beta) = \|y - X\beta\|^2$, $\hat{y} = X\hat{\beta}$ is the \perp -projection of y

unbiased estimator of σ : $\hat{\sigma} = \frac{\|y - \hat{y}\|^2}{N - p - 1}$

If true model is $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$: $\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1} \sigma^2)$

(X considered deterministic, randomness comes from ϵ and Y)

With some linear algebra: $(N - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi_{N-p-1}^2$ and $\hat{\sigma} \perp \hat{\beta}$

3.3 Subset Selection