The Elements of Statistical Learning

Chapter 2.

predictors

responses

(categorical) factors   ordered categorical var: chapter 4

components (predictor j) X_j ith observation x_i prediction of the output $Y \in \mathbb{R}$: $\hat{Y} \in \mathbb{R}$

2.3 Two simple approaches to Prediction:

2.3.1 Least Squares

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j$$

$\hat{\beta}_0$: bias in machine learning

$$X^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}_{n \times p}$$

(an inner product) $\hat{Y} = X^T \hat{\beta}$

where x_i is $1 \times p$

Note: X is a column vector $(p \times n)$?   column vector: $p \times 1$ (p columns 1 row)

In general, Y can be a K-vector, the β would be $p \times K$ matrix of coefficients.

Solution: minimize the residual sum of squares:

$$\underset{\min \beta}{RSS(\beta)} = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

quadratic function: minimum always exists but may not be unique.

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \qquad X \ (N \times p)$$

$\Rightarrow$ Differentiating $RSS(\beta)$ w.r.t β: $X^T(y - X\beta) = 0 \quad X^T y = X^T X \beta \Rightarrow \beta = (X^T X)^{-1} X^T Y$ ☆

Fitted value of ith input: $\hat{y}_i = x_i^T \beta$

2.3.2 Nearest-Neighbor Methods

fitted $\hat{Y}$: $\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$ average or sometimes: majority vote (categorical response)

2.3.3 comparison

① Least squares: smoothness & stability 平滑且稳定 low-variance & high-bias

Guassian distr. assumption ☆

② KNN: not rely on any stringent assumptions, can adapt to any situation $\frac{\text{high-var}}{\text{\& low-bias}}$

$\boxed{2.4}$ Statistical Decision Theory EPE: expected (squared) prediction error

① Quantitative response (Y)

$$EPE(f) = E(Y - f(X))^2 = \int [y - f(x)]^2 Pr(dx, dy) = \int_x \int_Y (y - f(x))^2 P_r(y|x) P_r(x)$$

Conditioning on input X: (joint density $Pr(X,Y) = Pr(Y|X) \cdot Pr(X)$)

Therefore $EPE(f) = E_X E_{Y|X}([Y - f(x)]^2 | X)$

To minimize EPE: $f(x) = \underset{c}{\operatorname{argmin}}\, E_{Y|X}([Y - c]^2 | X = x)$

solution: $f(x) = E(Y | X = x)$ the conditional expectation

KNN: $\hat{f}(x) = Ave(y_i | x_i \in N_k(x))$

$f(x) \approx x^T\beta$ plug into $EPE(f) = E(Y - x^T\beta)^2$

differentiate w.r.t $\beta \Rightarrow \beta = [E(X^TX)]^{-1} E(XY)$

② Qualitative response (G) → K levels

$$EPE = E[L(G, \hat{G}(X))] \qquad Pr(G, X)$$

$$EPE = E_X \sum_{k=1}^{K} L[G_k, \hat{G}(X)] Pr(G_k | X)$$

Loss function: $K \times K$ matrix L where $L(k, \ell)$ is the price paid for classifying an observation belonging to class G_k to class G_ℓ.

To minimize EPE:

$$\hat{G}(x) = \underset{g \in G}{\operatorname{argmin}} \sum_{k=1}^{K} L(G_k, g) Pr(G_k | X = x)$$

With 0-1 loss function: When $g = G_k$

$$\hat{G}(x) = \underset{g \in G}{\operatorname{argmin}} \, (1 - Pr(g|X=x))$$

or simply $\hat{G}(x) = G_k$ if $Pr(G_k|X=x) = \max_{g \in G} Pr(g|X=x)$ Bayes Classifier

Example: if G only has 2 levels and be denoted as $Y=0$ or 1.

Then $\hat{f}(x) = E(Y|X) = Pr(G = G_1|X)$ if G_1 corresponds to $Y=1$.

$\boxed{2.5}$ Local Methods in High Dimensions

※ Curse of dimensionality

Example: $Y = f(x) = e^{-8\|X\|^2}$ → true relationship predict y_0 at the test point $x_0 = 0$.

Training set $\mathcal{T}$.

true value

MSE for estimating $f(0)$:

estimator

$$MSE(x_0) = E_\mathcal{T}[f(x_0) - \hat{y}_0]^2 = E_\mathcal{T}[f(x_0) - E_\mathcal{T}(\hat{y}_0) + E(\hat{y}_0) - \hat{y}_0]^2$$

$$= E_\mathcal{T}[\hat{y}_0 - E_\mathcal{T}(\hat{y}_0)]^2 + (E_\mathcal{T}[\hat{y}_0] - f(x_0))^2$$

$$= Var_\mathcal{T}(\hat{y}_0) + Bias^2\,\hat{y}_0$$

(2.27) $EPE(x_0) = E_{y_0|x_0} E_\mathcal{T}(y_0 - \hat{y}_0)^2$

$= E_{y_0|x_0} E_\mathcal{T}(x_0^T\beta$

2.6

$T = \{x_i, y_i\}$ $i = 1, 2, \cdots, N$ T : training set

① linear basis expansions: $f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k$

3.4 Shrinkage Methods

3.4.1 Ridge

exclude β_0 from the penalty term

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{} \right\} \qquad \lambda > 0$$

normally standardize the inputs before solving the above formula

In matrix form: $RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$ $\qquad -2X^T(y - X\beta) + 2\lambda\beta = 0$

$-X^T y + X^T X\beta + \lambda\beta = 0$

$(X^T X + \lambda I)\beta = X^T y$

Then $\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$

For orthonormal inputs, $\hat{\beta}^{ridge} = \dfrac{\hat{\beta}}{1+\lambda}$

$$df(\lambda) = tr[X(X^T X + \lambda I)^{-1} X^T] = tr(H\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Exercise 3.6 prior dist. $\beta_j \overset{i.i.d}{\sim} N(0, \tau^2)$

$$f(\beta|y) = \frac{f(\beta, y)}{f(y)} = \frac{f(y|\beta)f(\beta)}{\underbrace{\int_{\beta} f(y|\beta)f(\beta)d\beta}_{\text{Constant}}} \propto f(y|\beta)f(\beta) \leftarrow$$

$f(\beta|y)$ is a function of $f(y|\beta)f(\beta)$ or like follow the same (scale different) distribution

$$\| \quad N(\beta_0 + x_i^T\beta, \sigma^2) \cdot N(0, \tau^2)$$

$$\int_{\beta|y} c f(\beta|y) d\beta = 1$$

Thin SVD of X matrix.

$$\underset{N\times p}{X} = \underset{N\times p}{U} \ \underset{p\times p}{D} \ \underset{p\times p}{V^T}$$

only calculated the p column vectors which correspond to p row vectors of V^T

Then $X\hat{\beta} = X(\underbrace{X^T X}_{I})^{-1} X^T y = UU^T y$

$UDV^T (VD^T U^T U DV^T)^{-1} VD^T U^T y$

$UD\underbrace{V^T (V^T)^{-1}}_{1} D^{-1} (D^T)^{-1} \underbrace{V^{-1} V}_{1} D^T U^T y$

Then ridge $X\hat{\beta}^{ridge} = X(X^TX+\lambda I)^{-1}X^Ty$

$$= X(VD^2V^T+\lambda VV^T)^{-1}VDU^Ty$$
$$= X(V(D^2+\lambda I)V^T)^{-1}VDU^Ty$$
$$= X(V^T)^{-1}(D^2+\lambda I)^{-1}\underbrace{V^{-1}V}_{I}DU^Ty$$
$$= X(V^T)^{-1}(D^2+\lambda I)^{-1}DU^Ty$$
$$= UD\underbrace{V^T(V^T)^{-1}}_{I}(D^2+\lambda I)^{-1}DU^Ty$$
$$= UD(D^2+\lambda I)^{-1}DU^Ty$$
$$= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y$$

$u_j \downarrow$ column of U

sample covariance matrix: $S = X^TX/N$

where $X^TX = \underline{VD^2V^T}$ → eigen decomposition

first principle component direction v_1: $z_1 = Xv_1$ has the largest sample variance.

$$Var(z_1) = Var(Xv_1)$$
$$= v_1^T Var(X) v_1$$
$$= v_1^T \frac{X^TX}{N} v_1$$
$$= v_1^T \frac{VD^2V^T}{N} v_1 = \frac{d_1^2}{N}$$

where $v_1^TV = [1\ 0\ 0 \cdots]$

$$Var(z_1) = \frac{[1\ 0\ 0\cdots]\begin{bmatrix} d_1^2 & & 0 \\ & d_2^2 & \\ 0 & & \ddots\ d_p^2 \end{bmatrix}\begin{bmatrix}1\\0\\0\\\vdots\end{bmatrix}}{N} = \frac{d_1^2}{N}$$

3.4.2 Lasso

$$\hat{\beta}^{lasso} = \underset{\beta}{\arg\min} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

equals

$$\hat{\beta}^{lasso} = \underset{\beta}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|} \right\}$$

nonlinear in the y_i.
no closed form expression as in ridge

eg: $t_0 = \sum_1^p |\beta_j|$ if $t = \frac{t_0}{2}$, LSE coefficients are shrunk by about 50% on average.

Appendix. - Terminology

Chapter 2.

环绕空间

hyperplane (超平面): a subspace whose dimension is one less than that of its ambient space.

An affine set (仿射集) affine space (仿射空间)

condimension:

Bivariate Guassian (/normal) distribution