18.4 Model Selection

Consider predicting a new observation Y^* for covariates X^* and let $S \subset J$ denote a subset of the covariates in the model, where $|S| = k$ and $|J| = n$.
Issues

- Underfitting: too few covariates yields high bias
- Overfitting: too many covariates yields high variance

Procedure

1. Assign a score to each model
2. Search through all models to find the one with the highest score

Hypothesis testing

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0 \quad \forall j \in J$$

Mean squared prediction error (MSPE)

$$\text{MSPE} = \mathbb{E}\left[(\widehat{Y}(S) - Y^*)^2\right]$$

Prediction risk

$$R(S) = \sum_{i=1}^n \text{MSPE}_i = \sum_{i=1}^n \mathbb{E}\left[(\widehat{Y}_i(S) - Y_i^*)^2\right]$$

Training error

$$\widehat{R}_{tr}(S) = \sum_{i=1}^n (\widehat{Y}_i(S) - Y_i)^2$$

R^2

$$R^2(S) = 1 - \frac{\text{RSS}(S)}{\text{TSS}} = 1 - \frac{\widehat{R}_{tr}(S)}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (\widehat{Y}_i(S) - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2}$$

The training error is a downward-biased estimate of the prediction risk.

$$\mathbb{E}\left[\widehat{R}_{tr}(S)\right] < R(S)$$

$$\mathsf{bias}(\widehat{R}_{tr}(S)) = \mathbb{E}\left[\widehat{R}_{tr}(S)\right] - R(S) = -2\sum_{i=1}^n \text{Cov}\left[\widehat{Y}_i, Y_i\right]$$

Adjusted R^2

$$R^2(S) = 1 - \frac{n-1}{n-k}\frac{\text{RSS}}{\text{TSS}}$$

Mallow's C_p statistic

$$\widehat{R}(S) = \widehat{R}_{tr}(S) + 2k\widehat{\sigma}^2 = \text{lack of fit} + \text{complexity penalty}$$

Akaike Information Criterion (AIC)

$$AIC(S) = \ell_n(\widehat{\beta}_S, \widehat{\sigma}_S^2) - k$$

Bayesian Information Criterion (BIC)

$$BIC(S) = \ell_n(\widehat{\beta}_S, \widehat{\sigma}_S^2) - \frac{k}{2}\log n$$

Validation and training

$$\widehat{R}_V(S) = \sum_{i=1}^m (\widehat{Y}_i^*(S) - Y_i^*)^2 \qquad m = |\{\text{validation data}\}|, \text{ often } \frac{n}{4} \text{ or } \frac{n}{2}$$

Leave-one-out cross-validation

$$\widehat{R}_{CV}(S) = \sum_{i=1}^n (Y_i - \widehat{Y}_{(i)})^2 = \sum_{i=1}^n \left(\frac{Y_i - \widehat{Y}_i(S)}{1 - U_{ii}(S)}\right)^2$$

$$U(S) = X_S(X_S^T X_S)^{-1}X_S \text{ (“hat matrix”)}$$

19 Non-parametric Function Estimation

19.1 Density Estimation

Estimate $f(x)$, where $f(x) = \mathbb{P}[X \in A] = \int_A f(x)\,dx$.
Integrated square error (ISE)

$$L(f, \widehat{f}_n) = \int \left(f(x) - \widehat{f}_n(x)\right)^2 dx = J(h) + \int f^2(x)\,dx$$

Frequentist risk

$$R(f, \widehat{f}_n) = \mathbb{E}\left[L(f, \widehat{f}_n)\right] = \int b^2(x)\,dx + \int v(x)\,dx$$

$$b(x) = \mathbb{E}\left[\widehat{f}_n(x)\right] - f(x)$$
$$v(x) = \mathbb{V}\left[\widehat{f}_n(x)\right]$$

19.1.1 Histograms

Definitions

- Number of bins m
- Binwidth $h = \frac{1}{m}$
- Bin B_j has ν_j observations
- Define $\widehat{p}_j = \nu_j/n$ and $p_j = \int_{B_j} f(u)\,du$

Histogram estimator

$$
\begin{aligned}
\widehat{f}_n(x) &= \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) \\
\mathbb{E}\left[\widehat{f}_n(x)\right] &= \frac{p_j}{h} \\
\mathbb{V}\left[\widehat{f}_n(x)\right] &= \frac{p_j(1-p_j)}{nh^2} \\
R(\widehat{f}_n, f) &\approx \frac{h^2}{12}\int \left(f'(u)\right)^2\,du + \frac{1}{nh} \\
h^* &= \frac{1}{n^{1/3}}\left(\frac{6}{\int \left(f'(u)\right)^2}\,du\right)^{1/3} \\
R^*(\widehat{f}_n, f) &\approx \frac{C}{n^{2/3}} \qquad C = \left(\frac{3}{4}\right)^{2/3}\left(\int \left(f'(u)\right)^2\,du\right)^{1/3}
\end{aligned}
$$

Cross-validation estimate of $\mathbb{E}\left[J(h)\right]$

$$
\widehat{J}_{CV}(h) = \int \widehat{f}_n^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{(-i)}(X_i) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m}\widehat{p}_j^2
$$

19.1.2 Kernel Density Estimator (KDE)

Kernel K

- $K(x) \geq 0$
- $\int K(x)\,dx = 1$
- $\int xK(x)\,dx = 0$
- $\int x^2K(x)\,dx \equiv \sigma_K^2 > 0$

KDE

$$
\begin{aligned}
\widehat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-X_i}{h}\right) \\
R(f, \widehat{f}_n) &\approx \frac{1}{4}(h\sigma_K)^4\int (f''(x))^2\,dx + \frac{1}{nh}\int K^2(x)\,dx \\
h^* &= \frac{c_1^{-2/5}c_2^{-1/5}c_3^{-1/5}}{n^{1/5}} \qquad c_1 = \sigma_K^2,\ c_2 = \int K^2(x)\,dx,\ c_3 = \int (f''(x))^2\,dx \\
R^*(f, \widehat{f}_n) &= \frac{c_4}{n^{4/5}} \qquad c_4 = \underbrace{\frac{5}{4}(\sigma_K^2)^{2/5}\left(\int K^2(x)\,dx\right)^{4/5}}_{C(K)}\left(\int (f'')^2\,dx\right)^{1/5}
\end{aligned}
$$

Epanechnikov Kernel

$$
K(x) = \begin{cases} \frac{3}{4\sqrt{5}(1-x^2/5)} & |x| < \sqrt{5} \\ 0 & \text{otherwise} \end{cases}
$$

Cross-validation estimate of $\mathbb{E}\left[J(h)\right]$

$$
\widehat{J}_{CV}(h) = \int \widehat{f}_n^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{(-i)}(X_i) \approx \frac{1}{hn^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0)
$$

$$
K^*(x) = K^{(2)}(x) - 2K(x) \qquad K^{(2)}(x) = \int K(x-y)K(y)\,dy
$$

19.2 Non-parametric Regression

Estimate $f(x)$ where $f(x) = \mathbb{E}\left[Y \mid X = x\right]$. Consider pairs of points $(x_1, Y_1), \ldots, (x_n, Y_n)$ related by

$$
\begin{aligned}
Y_i &= r(x_i) + \epsilon_i \\
\mathbb{E}\left[\epsilon_i\right] &= 0 \\
\mathbb{V}\left[\epsilon_i\right] &= \sigma^2
\end{aligned}
$$

k-nearest Neighbor Estimator

$$
\widehat{r}(x) = \frac{1}{k}\sum_{i:x_i \in N_k(x)} Y_i \qquad \text{where } N_k(x) = \{k \text{ values of } x_1, \ldots, x_n \text{ closest to } x\}
$$

Nadaraya-Watson Kernel Estimator

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \quad \in [0,1]$$

$$R(\widehat{r}_n, r) \approx \frac{h^4}{4}\left(\int x^2 K^2(x)\,dx\right)^4 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \int \frac{\sigma^2 \int K^2(x)\,dx}{nhf(x)}\,dx$$

$$h^* \approx \frac{c_1}{n^{1/5}}$$

$$R^*(\widehat{r}_n, r) \approx \frac{c_2}{n^{4/5}}$$

Cross-validation estimate of $\mathbb{E}\left[J(h)\right]$

$$\widehat{J}_{CV}(h) = \sum_{i=1}^{n} (Y_i - \widehat{r}_{(-i)}(x_i))^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{r}(x_i))^2}{\left(1 - \frac{K(0)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}\right)^2}$$

19.3 Smoothing Using Orthogonal Functions

Approximation

$$r(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) \approx \sum_{j=1}^{J} \beta_j \phi_j(x)$$

Multivariate regression

$$Y = \Phi\beta + \eta$$

$$\text{where} \quad \eta_i = \epsilon_i \quad \text{and} \quad \Phi = \begin{pmatrix} \phi_0(x_1) & \cdots & \phi_J(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_n) & \cdots & \phi_J(x_n) \end{pmatrix}$$

Least squares estimator

$$\widehat{\beta} = (\Phi^T \Phi)^{-1} \Phi^T Y$$

$$\approx \frac{1}{n}\Phi^T Y \quad \text{(for equally spaced observations only)}$$

Cross-validation estimate of $\mathbb{E}\left[J(h)\right]$

$$\widehat{R}_{CV}(J) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{J} \phi_j(x_i) \widehat{\beta}_{j,(-i)}\right)^2$$

20 Stochastic Processes

Stochastic Process

$$\{X_t : t \in T\} \qquad T = \begin{cases} \{0, \pm 1, \dots\} = \mathbb{Z} & \text{discrete} \\ [0, \infty) & \text{continuous} \end{cases}$$

- Notations X_t, $X(t)$
- State space $\mathcal{X}$
- Index set T

20.1 Markov Chains

Markov chain

$$\mathbb{P}\left[X_n = x \mid X_0, \dots, X_{n-1}\right] = \mathbb{P}\left[X_n = x \mid X_{n-1}\right] \quad \forall n \in T, x \in \mathcal{X}$$

Transition probabilities

$$p_{ij} \equiv \mathbb{P}\left[X_{n+1} = j \mid X_n = i\right]$$

$$p_{ij}(n) \equiv \mathbb{P}\left[X_{m+n} = j \mid X_m = i\right] \quad \text{n-step}$$

Transition matrix $\mathbf{P}$ (n-step: $\mathbf{P}_n$)

- (i,j) element is p_{ij}
- $p_{ij} > 0$
- $\sum_i p_{ij} = 1$

Chapman-Kolmogorov

$$p_{ij}(m+n) = \sum_k p_{ij}(m) p_{kj}(n)$$

$$\mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n$$

$$\mathbf{P}_n = \mathbf{P} \times \cdots \times \mathbf{P} = \mathbf{P}^n$$

Marginal probability

$$\mu_n = (\mu_n(1), \dots, \mu_n(N)) \quad \text{where} \quad \mu_i(i) = \mathbb{P}\left[X_n = i\right]$$

$$\mu_0 \triangleq \text{initial distribution}$$

$$\mu_n = \mu_0 \mathbf{P}^n$$