Bayes risk

$$r(f,\widehat{\theta}) = \iint L(\theta,\widehat{\theta}(x)) f(x,\theta)\,dx\,d\theta = \mathbb{E}_{\theta,X}\left[L(\theta,\widehat{\theta}(X))\right]$$

$$r(f,\widehat{\theta}) = \mathbb{E}_{\theta}\left[\mathbb{E}_{X|\theta}\left[L(\theta,\widehat{\theta}(X))\right]\right] = \mathbb{E}_{\theta}\left[R(\theta,\widehat{\theta})\right]$$

$$r(f,\widehat{\theta}) = \mathbb{E}_{X}\left[\mathbb{E}_{\theta|X}\left[L(\theta,\widehat{\theta}(X))\right]\right] = \mathbb{E}_{X}\left[r(\widehat{\theta}\,|\,X)\right]$$

17.2 Admissibility

- $\widehat{\theta}'$ dominates $\widehat{\theta}$ if

$$\forall\theta : R(\theta,\widehat{\theta}') \le R(\theta,\widehat{\theta})$$

$$\exists\theta : R(\theta,\widehat{\theta}') < R(\theta,\widehat{\theta})$$

- $\widehat{\theta}$ is inadmissible if there is at least one other estimator $\widehat{\theta}'$ that dominates it. Otherwise it is called admissible.

17.3 Bayes Rule

Bayes rule (or Bayes estimator)

- $r(f,\widehat{\theta}) = \inf_{\widetilde{\theta}} r(f,\widetilde{\theta})$
- $\widehat{\theta}(x) = \inf r(\widehat{\theta}\,|\,x)\ \forall x \implies r(f,\widehat{\theta}) = \int r(\widehat{\theta}\,|\,x) f(x)\,dx$

Theorems

- Squared error loss: posterior mean
- Absolute error loss: posterior median
- Zero-one loss: posterior mode

17.4 Minimax Rules

Maximum risk

$$\bar{R}(\widehat{\theta}) = \sup_{\theta} R(\theta,\widehat{\theta}) \qquad \bar{R}(a) = \sup_{\theta} R(\theta,a)$$

Minimax rule

$$\sup_{\theta} R(\theta,\widehat{\theta}) = \inf_{\widetilde{\theta}} \bar{R}(\widetilde{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta,\widetilde{\theta})$$

$$\widehat{\theta} = \text{Bayes rule} \ \wedge\ \exists c : R(\theta,\widehat{\theta}) = c$$

Least favorable prior

$$\widehat{\theta}^f = \text{Bayes rule} \ \wedge\ R(\theta,\widehat{\theta}^f) \le r(f,\widehat{\theta}^f)\ \forall\theta$$

18 Linear Regression

Definitions

- Response variable Y
- Covariate X (aka predictor variable or feature)

18.1 Simple Linear Regression

Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \mathbb{E}\left[\epsilon_i \,|\, X_i\right] = 0,\ \mathbb{V}\left[\epsilon_i \,|\, X_i\right] = \sigma^2$$

Fitted line

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Predicted (fitted) values

$$\widehat{Y}_i = \widehat{r}(X_i)$$

Residuals

$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)$$

Residual sums of squares (RSS)

$$\text{RSS}(\widehat{\beta}_0,\widehat{\beta}_1) = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Least square estimates

$$\widehat{\beta}^T = (\widehat{\beta}_0,\widehat{\beta}_1)^T : \min_{\widehat{\beta}_0,\widehat{\beta}_1} \text{RSS}$$

$$\widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{X}_n$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\overline{X^2}}$$

$$\mathbb{E}\left[\widehat{\beta}\,|\,X^n\right] = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$$

$$\mathbb{V}\left[\widehat{\beta}\,|\,X^n\right] = \frac{\sigma^2}{n s_X^2}\begin{pmatrix} n^{-1}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1\end{pmatrix}$$

$$\widehat{\mathsf{se}}(\widehat{\beta}_0) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{s_X\sqrt{n}}$$

where $s_X^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ and $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2$ (unbiased estimate). Further properties:

- Consistency: $\widehat{\beta}_0 \xrightarrow{\text{P}} \beta_0$ and $\widehat{\beta}_1 \xrightarrow{\text{P}} \beta_1$

- Asymptotic normality:

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\mathsf{se}}(\widehat{\beta}_0)} \xrightarrow{\text{D}} \mathcal{N}(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\mathsf{se}}(\widehat{\beta}_1)} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$

- Approximate $1-\alpha$ confidence intervals for β_0 and β_1:

$$\widehat{\beta}_0 \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\beta}_0) \quad \text{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\beta}_1)$$

- Wald test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$: reject H_0 if $|W| > z_{\alpha/2}$ where $W = \widehat{\beta}_1/\widehat{\mathsf{se}}(\widehat{\beta}_1)$.

R^2

$$R^2 = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2} = 1 - \frac{\textsc{rss}}{\textsc{tss}}$$

Likelihood

$$\mathcal{L} = \prod_{i=1}^n f(X_i, Y_i) = \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i \,|\, X_i) = \mathcal{L}_1 \times \mathcal{L}_2$$

$$\mathcal{L}_1 = \prod_{i=1}^n f_X(X_i)$$

$$\mathcal{L}_2 = \prod_{i=1}^n f_{Y|X}(Y_i \,|\, X_i) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i \Big(Y_i - (\beta_0 - \beta_1 X_i)\Big)^2\right\}$$

Under the assumption of Normality, the least squares estimator is also the MLE but the least squares variance estimator is not the MLE.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2$$

18.2 Prediction

Observe $X = x_*$ of the covariate and want to predict their outcome Y_*.

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*$$

$$\mathbb{V}\left[\widehat{Y}_*\right] = \mathbb{V}\left[\widehat{\beta}_0\right] + x_*^2\mathbb{V}\left[\widehat{\beta}_1\right] + 2x_*\mathrm{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1\right]$$

Prediction interval

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2\left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_i (X_i - \bar{X})^2 j} + 1\right)$$

$$\widehat{Y}_* \pm z_{\alpha/2}\widehat{\xi}_n$$

18.3 Multiple Regression

$$Y = X\beta + \epsilon$$

where

$$X = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Likelihood

$$\mathcal{L}(\mu, \Sigma) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\textsc{rss}\right\}$$

$$\textsc{rss} = (y - X\beta)^T(y - X\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^N (Y_i - x_i^T\beta)^2$$

If the $(k \times k)$ matrix X^TX is invertible,

$$\widehat{\beta} = (X^TX)^{-1}X^TY$$

$$\mathbb{V}\left[\widehat{\beta} \,|\, X^n\right] = \sigma^2(X^TX)^{-1}$$

$$\widehat{\beta} \approx \mathcal{N}\left(\beta, \sigma^2(X^TX)^{-1}\right)$$

Estimate regression function

$$\widehat{r}(x) = \sum_{j=1}^k \widehat{\beta}_j x_j$$

Unbiased estimate for σ^2

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{\epsilon}_i^2 \qquad \hat{\epsilon} = X\widehat{\beta} - Y$$

MLE

$$\widehat{\mu} = \bar{X} \qquad \widehat{\sigma}^2 = \frac{n-k}{n}\sigma^2$$

$1-\alpha$ Confidence interval

$$\widehat{\beta}_j \pm z_{\alpha/2}\widehat{\mathsf{se}}(\widehat{\beta}_j)$$

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)

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

Prediction risk

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

Training error

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

R^2

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

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

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

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

Adjusted R^2

$$R^2(S) = 1 - \frac{n-1}{n-k}\frac{\textsc{rss}}{\textsc{tss}}$$

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

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

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