Information Criterion

Ways to quantify error in a given estimator.

One idea

$$\overline{err} = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(x_i))\right]$$

In general underestimates error due to training and testing on $\{Y_i\}$.

Create a new dataset of responses $Y_i^{(new)}$ to get a better estimate of the error $Err_{in} = \mathbb{E}_Y\left[\frac{1}{N}\sum_{i=1}^{N} L(\hat{f}(x_i), Y_i)\right]$

View $Y_i^{(new)}$ as R.V. coming from the underlying model.

Def: The difference between the training error and the theoretical error is given by the optimisism

$$op = Err_{in} - \overline{err}$$

and the true optimism is given by

$$\mathbb{E}_\tau[op] = \frac{2}{N}\sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$$

Which yields $\mathbb{E}_\tau[Err_{in}] = \overline{err} + \frac{2}{N}\sum_{i=1}^{n} Cov(Y_i, \hat{Y}_i)$

We normally say $\mathbb{E}_{\tau}(\text{Err}_{\text{in}})$ is the CP information criterion.

VC Dimension Criterion

Rmk: Optimism related to how complex the space $f \in \mathcal{F}$.

Suppose $f(x) = I_A$ (classification) e.g. $A = \{x : \beta^T x > 0\}$

Def: The family $\mathcal{F}$ of indicator functions is said to shatter a collection of points $C \subset \mathcal{F}$. If for every subset $C_1 \subset C$. there exists a function $f_n = I_A \in \mathcal{F}$ s.t. $I_A(x) = \begin{cases} 1 & x \in C_1 \\ 0 & x \in C \setminus C_1 \end{cases}$

Def: $\mathcal{F}$ has VC dimension if it can shatter any collection of points

Thm: $\mathbb{E}[L(\hat{f}(X,Y))] = \text{Err}_{\tau} \le \dfrac{\overline{\text{err}}}{(1-\sqrt{\varepsilon})_+}$ $\qquad \varepsilon = \dfrac{h(\log N/h + 1) - \log \eta/4}{N}$

$h = \text{VC}$ dimension with probability $1-\eta$.

Bootstrapping

Define $Z = \begin{pmatrix} x_1^T & y_1 \\ \vdots & \vdots \\ x_n^T & y_n \end{pmatrix}$ with a goal of estimating $S(Z)$.

Estimate the density $\hat{P}$ and resample from it to estimate

$\{\hat{S}(z)\}_{i=1}^{B}$