 22 April 04

<u>Cross-validation</u>. The practice of partitioning a sample of data into subsamples such that analysis is initially performed on a single subsample, while further subsamples are retained "blind" in order for subsequent use in confirming and validating the initial analysis.

Tied in with the idea of <u>prediction error</u>.

Loss function $L(y, \hat{y})$

y: future response, $\hat{y}$: prediction from model

e.g. $(y-\hat{y})^2$

<u>prediction error</u>, Δ

$$E(y-\hat{y})^2$$

in classification problem

$$\text{Prob}\{\hat{y} \neq y\}$$

$E\,L(y,\hat{y})$ <u>error rate</u>

22 April 04

e.g. <u>multiple regression</u>

Data $y_i, \underset{\sim}{x}_i^T$

Future observation $Y_0, \underset{\sim}{x}_0^T$

$$Y_0 = \underset{\sim}{x}_0^T \underset{\sim}{\beta} + \epsilon_0$$

Aggregate prediction error, Δ

$$\frac{1}{n} \sum' E(Y_0 - \hat{Y}_0)^2 \qquad (*)$$

$$\hat{Y}_0 = \underset{\sim}{x}_0^T \hat{\underset{\sim}{\beta}}$$

$$Y_0 - \hat{Y}_0 = \underset{\sim}{x}_0^T (\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}}) + \epsilon_0$$

$$E(Y_0 - \hat{Y}_0)^2 = \underset{\sim}{x}_0^T (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_0 \sigma^2 + \sigma^2$$

$\sum'$ is over $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$

$$(*) = \frac{1}{n} \sum_i' \mathrm{tr}\{(\underset{\sim}{X}^T \underset{\sim}{X})^{-1} x_i x_i^T\} \sigma^2 + \sigma^2$$

$$= \left(\frac{p}{n} + 1\right) \sigma^2$$

Estimate by $\left(\frac{p}{n} + 1\right) s^2$

22 April 04

Estimate, deleting cases in turn

$$\hat{\Delta}_{CV} = \frac{1}{n} \sum_i (y_i - \underset{\sim}{x}_i^T \hat{\beta}_{-i})^2$$

$$\underset{\sim}{\hat{\beta}} - \hat{\beta}_{-i} = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{x}_i (y_i - x_i^T \hat{\beta}) / (1 - H_{ii})$$

H_{ii} : leverage $\quad H = X(X^T X)^{-1} X^T$

$$\hat{\Delta}_{CV} = \frac{1}{n} \sum_i (y_i - \underset{\sim}{x}_i^T \underset{\sim}{\hat{\beta}})^2 / (1 - H_{ii})^2$$

Here $m(\underset{\sim}{x}_i) = \underset{\sim}{x}_i^T \underset{\sim}{\beta}$

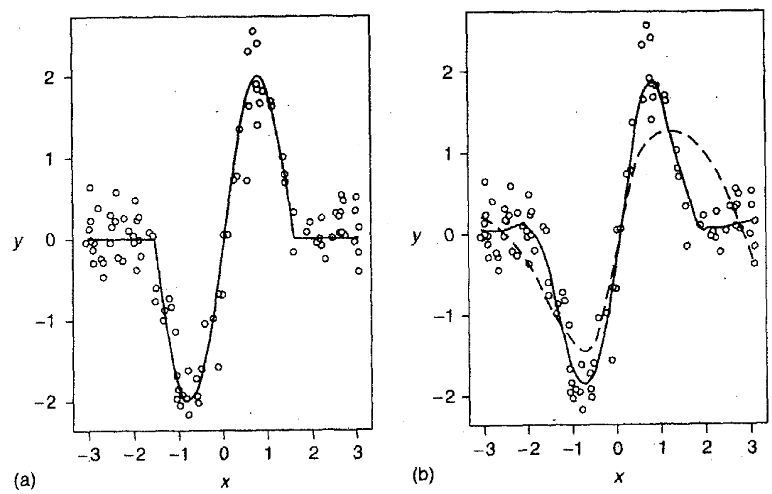
In case of $EY = m(\underset{\sim}{x}, \alpha)$ and

$$\hat{Y} = \hat{m}(\underset{\sim}{x}, \alpha) = H(\alpha) y$$

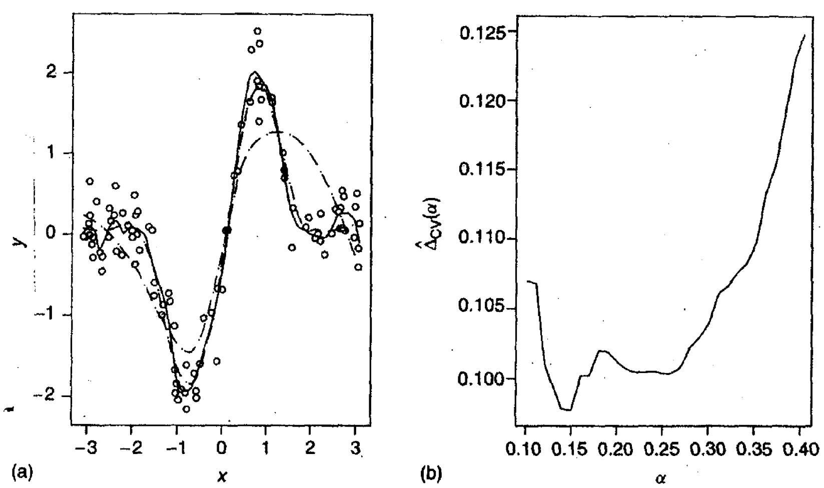
have

$$\hat{\Delta}_{CV} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{m}(x_i, \alpha))^2 / (1 - H_{ii}(\alpha))^2$$

Can be used to choose α.



(a) Simulated data with underlying model $m(x)$; (b) LOESS fits with $\alpha = \frac{1}{2}, \frac{2}{3}$



(a) LOESS fits based on $\alpha \in \left\{\frac{1}{2}, \frac{2}{3}\right\}$ (dashed) and on cross-validated $\alpha_{CV}(S) = 0.15$ (solid); (b) $\hat{\Delta}_{CV}(\alpha)$ for .11, ..., 0.4

22 April 04

Cross-validation in

Splus.

cv.tree( )
supsmu(, span = "cv")
ucv MASS
bcv MASS
smooth.spline()

R.

validate.tree(Design)
bcv (MASS)
ucv (MASS)
cv.glm (boot)
knn.cv (class)
(fun Fits)
xpred.rpart (rpart)
crossval (supclust)
cv. tree (tree)

mgcv ?