Johns Hopkins Engineering

625.464 Computational Statistics

Cross Validation for Smoothing and Fitting

Module 8 Lecture 8B

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Fitting Y using X

Assume we have some observed data set $\mathcal{S} \subseteq X \times Y$. Consider the problem of fitting $\underline{Y}$ using $\underline{X}$, ie the problem of determining a function $g_{\underline{XY}}(x)$ such that $\underline{Y} \sim g_{\underline{xy}}(x)$

For a given point (x_0, y_0) how well does $g_{\underline{xy}}(x_0)$ match y_0?

How well will our fitted model $g_{\underline{xy}}$ perform at new points? How useful is it as a predictor?

How useful is our model as a predictor?

Let $R(y, \hat{y})$ be the error between y & $\hat{y}$. Ex $(y-\hat{y})^2$

To answer our question we need

$$E_{P_{Y|X}}\left(R\left(Y_0, g_{XY}(X_0)\right)\right)$$

want to minimize this, but don't know $P_{Y|X}$

$$E_{\hat{P}_{Y|X}}\left[R\left(Y_0, g_{XY}(X_0)\right)\right] = \frac{1}{n}\sum_{i=1}^{n} R\left(y_i, g_{XY}(X_i)\right)$$

for observed $(X_i, y_i) \in S$

apparent error < true error

Can we get a better estimate of the true error?

Consider partitioning our data set S' into two parts S_1' & S_2'

S_1' - training or estimating set and will be used to get the fit g_{1xy}

S_2' - validation or test set and can be used to estimate the expected error

$$E_{\hat{P}_1 Y|X}\left(R(Y_1, g_{1xy}(X_1))\right) = \frac{1}{\#(S_2')} \sum_{i \in S_2'} R(y_i, g_{1xy}(X_i))$$

We can switch the roles of S_1' & S_2' to obtain g_{2xy}

balanced half sampling

$$E_{\hat{P} Y|X}\left(R(Y_0, g_{xy}(X_0))\right) = \frac{1}{n}\left[\sum_{i \in S_2} R(y_i, g_{1xy}(X_i)) + \sum_{i \in S_1'} R(y_i, g_{2xy}(X_i))\right]$$

Cross Validation

Cross Validation is forming multiple partial data sets with overlap and then comparing the fitted values with the observed values.

K-fold Cross Validation

--Divide the sample into k approximately equal subsets

--1 by 1 hold each subset back and generate a fit with the remaining k-1 subsets

--Measure the prediction error by using the subset held back

--This gives k estimates of the prediction error which can be averaged to find an overall error estimate.