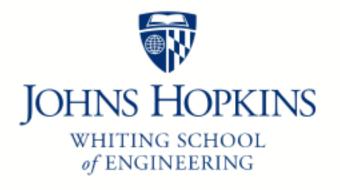
Johns Hopkins Engineering

625.464 Computational Statistics

Cross Validation for Smoothing and Fitting

Module 8 Lecture 8B



Fitting Y using X

Assume we have some so served data

set SCXXY. Consider the problem

of fitting I using I, ie the problem

of Letermining a function GIT(X)

Such that In AZI(X)

For a given point (Xo, y.) how well does gry (Xo) match yo?

How well will our fitted model grig penform at new points? How useful is it as a predictor?

How useful is our model as a predictor?

Let R(y, q) be the error between y= g. Ex (y-g)² To answer our question he read Lpyx (R (Yo, gxr (Xo)) want to minimize this, but don't know PyIX PENIX (R(Vo, gx (Xi)) = 1 5 R(yi, gx (Xi))

apparent

evror

evror

evror For observed (Xin4i) to Can we get a better estimate of the true error?

Consider partitioning our data set s'into two parts si & si est-training or estimating set and will be used to get the fit give Sig - validation or test set and can be used to estimate the expected error $E_{\hat{P}_{1}}V_{1X}(R(Y_{1},g_{XY}(X_{1})) = \frac{1}{4(65)} \sum_{i \neq 5} R(y_{i},g_{1,XY}(X_{i}))$ We can switch the roles of Site Setoobtain 92ky balances

rate

rate

Sometro

FOVIX (R(Volgy(Xol)) = 1 [ZR(yi,gly(Xi)) + 2 R(gi,gly(Xi)) + 2 R(gi,gly

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.