

Stat 471: Lecture 30

Cross-Validation.

Moo K. Chung
mchung@stat.wisc.edu

November 21, 2003

1. Given models

$$M_\sigma : Y = g(x, \sigma^2) + \epsilon,$$

we wish to choose a best model. Suppose we estimate g by nonparametric regression such as kernel smoothing with bandwidth σ^2 :

$$\hat{g}(x_i, \sigma^2) = \sum_{i=1}^m w_i y_i$$

where

$$w(y_i, y) = \frac{\exp\left(-\frac{\|y_i - y\|^2}{2\sigma^2}\right)}{\sum_{i=0}^m \exp\left(-\frac{\|y_i - y\|^2}{2\sigma^2}\right)}.$$

Among all possible models, we choose a model that give the smallest possible prediction error. The residual is

$$e_i = y_i - \hat{y}_i = y_i - \hat{g}(x_i, \sigma^2).$$

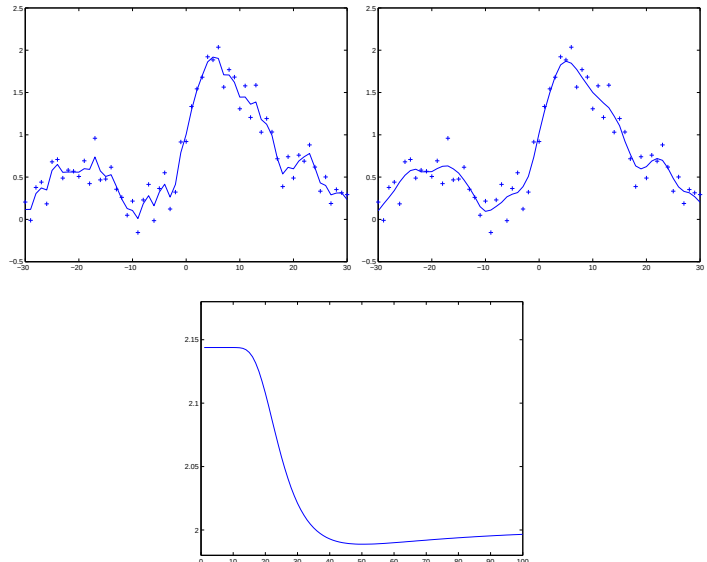
The residual sum of squares (RSS) of a model $\sum_{i=1}^n e_i^2$ gives a measure of predictive ability. However RSS is not a good model selector. Most complex model that overfits data will have the smallest RSS. It is due to the fact that $\hat{y}_i$ uses y_i as well as all other observations.

```
x=-30:30;
g=0.2*x.*exp(-x.^2/100) + cos(x/30).^2;
y=g+0.2*randn(1,61);
kernel=inline('exp(-x.^2/sigma^2)/
sum(exp(-x.^2/sigma^2))','x','sigma')
weight=kernel([-2:2],2);
yhat = convn(y,weight,'same');
>>sum((y-yhat).^2)
ans = 0.6445 %sigma=1
ans = 1.3638 %sigma=2
```

- To fix the above problem, we leave out y_i when we predict y_i and call the prediction $\hat{y}_{(i)}$ and let

$$e_{(i)} = y_i - \hat{y}_{(i)}.$$

Then the predicted residual errors (PRESS) is $\sum_{i=1}^n e_{(i)}^2$. It is also called the *cross-validation* (CV) statistic. The cross-validation splits data into two disjoint sets to fit a



model in one set and to validate the model in the other set. PRESS uses *leave-one-out* cross-validation. In kernel smoothing, the prediction depends on parameter σ^2 , so we choose σ^2 such that we minimize

$$CV(\sigma^2) = \sum_{i=1}^n (y_i - \hat{g}_{-i}(x_i, \sigma^2))^2.$$

where $\hat{g}_{-i}(x_i, \sigma^2)$ is the kernel smoothing estimator evaluated at x_i when (x_i, y_i) are removed.

```
for j=1:50
    sigma=j/20;
    for i=3:59 %forget boundary data
        weight=kernel([-2 -1 1 2],sigma);
        yi=[y(i-2) y(i-1) y(i+1) y(i+2)]';
        yhat(i)=weight*yi;
    end
    cv(j)=sum((y(3:59)-yhat(3:59)).^2);
end;
plot(cv)
>> find(cv==min(cv))
ans =
    50
```