Ch => 5

Resampling Methods

Data is currency of applied Machine learning. Therefore, it is important that it is both collected & used safely.

Data resampling refers to method of economically using a collected dataset to improve the estimate of the population parameter & help to quantify the uncertainity of the estimate.

Resampling methods are an indispensible tool in modern statistics.

For example, in order to estimate the variability of a linear regression fit, we can repeatedly draw different samples from training data, fit a linear regression to each new sample, and then examine the extent to which the resulting fits differ. Such an approach may allow us to obtain information that would not be available from fitting the model only once using training samples.

Cross-Validation:

We had a simple approach to test our data on the test set by dividing data into 2 sets: One for training & another for testing. This though requires a lot of data & that is not the case a lot of times. In those times resampling methods comes into the pictures.

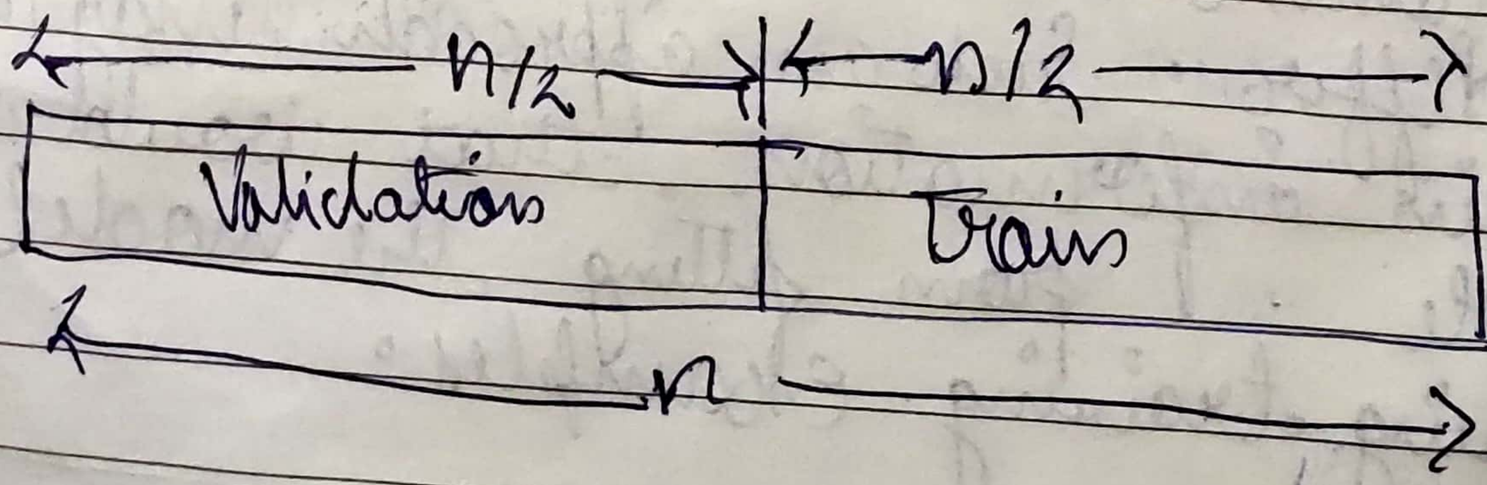
→ We hold out some random data of our training set & estimate our testing error on that subset of data & repeat this.

(1) The validation Set Approach:

→ Here, we randomly divides available set of samples into 2 parts, training set & validation set.

→ The model is fit on training set & the fitted model is used to predict the response for the observations of validation set.

→ The resulting validation error provides an estimate of the test error.



This has 2 potential drawbacks.

1. The validation test set error estimate have high variability on the estimates of test error. Because, the more the data, the better is the training & this approach kindof looses the because of this reason.

2. The validation set error tend to overestimate the test error because of the same reason that, statistical methods tends to perform worse when trained on fewer examples.

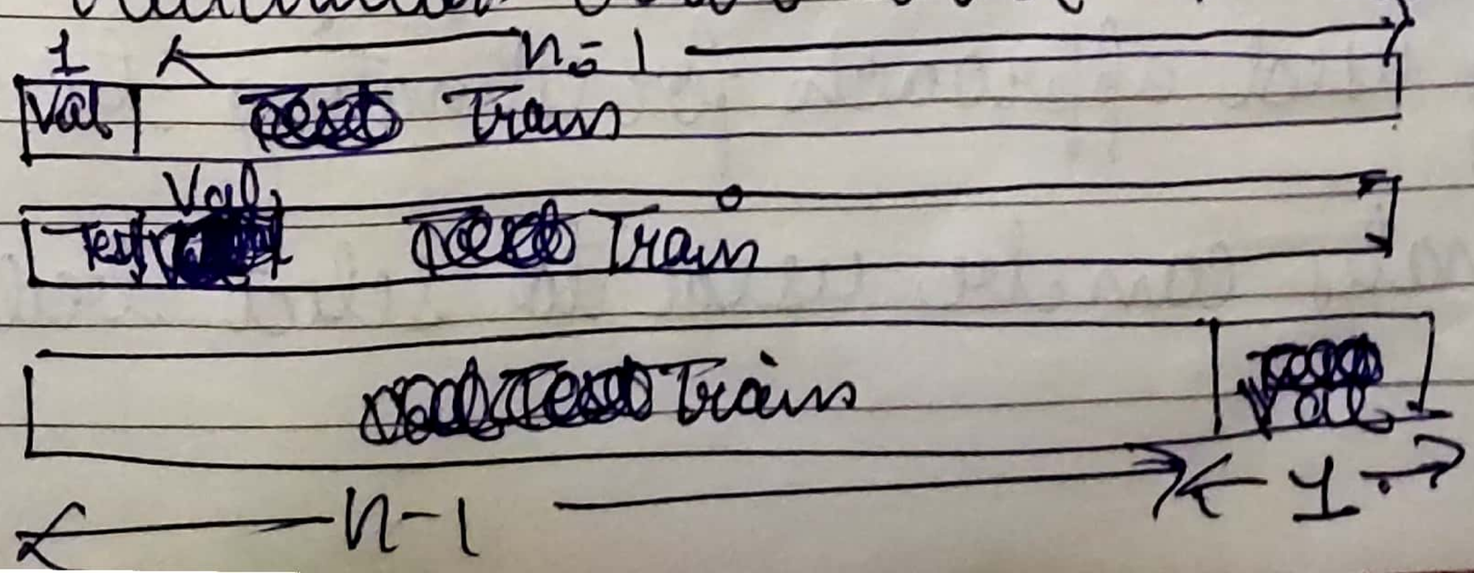
(2.) LOOCV (Leave One Out Cross-Validation)

This is closely related to validation set approach.

→ Here we split data into 2 set, one set contains n-1 values while other one contains only one value.

→ we apply this method n times, on n different validation set.

→ The validation error is calculated by averaging the validation error over "n validation" test.



Date :
Page :

Advantages Over Validation Set Approach

1. Firstly, there is far less bias because it fits data in the over a dataset of "n-1" values.
2. It doesn't overestimate the test error due to lack of training or test examples. This will almost yield same results everytime.

$$CV_n = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(1-h_i)^2}$$

where h_i is , high leverage points learnt in ch-3. Pg 98

Disadvantage:

1. High Computational expensive.
2. Estimates of each fold is highly correlated hence their average can have high variance.

(3) K-Fold Cross Validation :

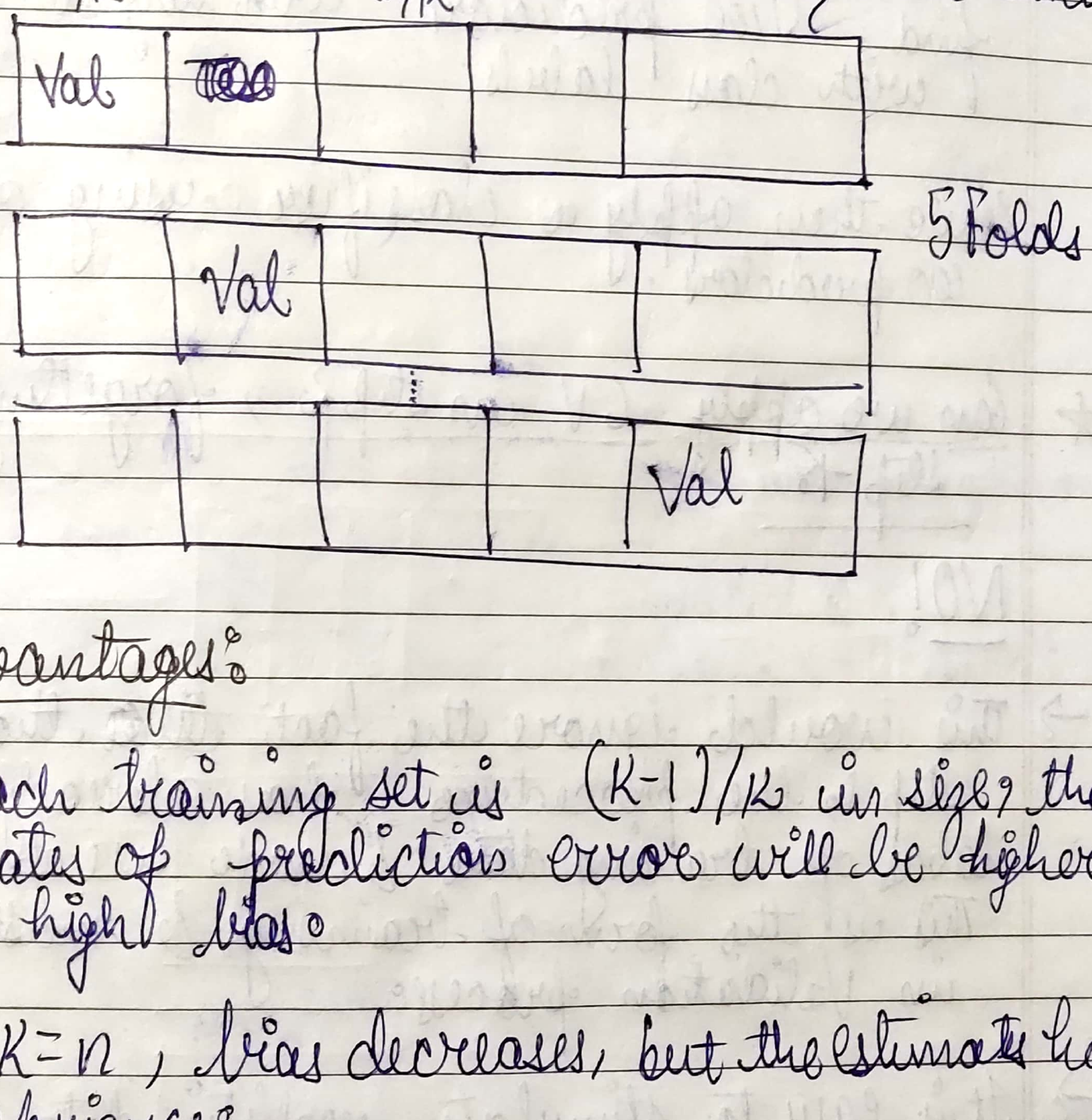
An alternative to LOOCV is K-fold CV.

→ Widely used approach for estimating test error.

→ Estimates can be used to select best model.

→ Idea is to randomly divide the data into K-equal sized parts. We leave out part K, fit the model to other K-1 parts (combined) & then obtains for left out Kth part.

→ This is done in turns & then results are combined

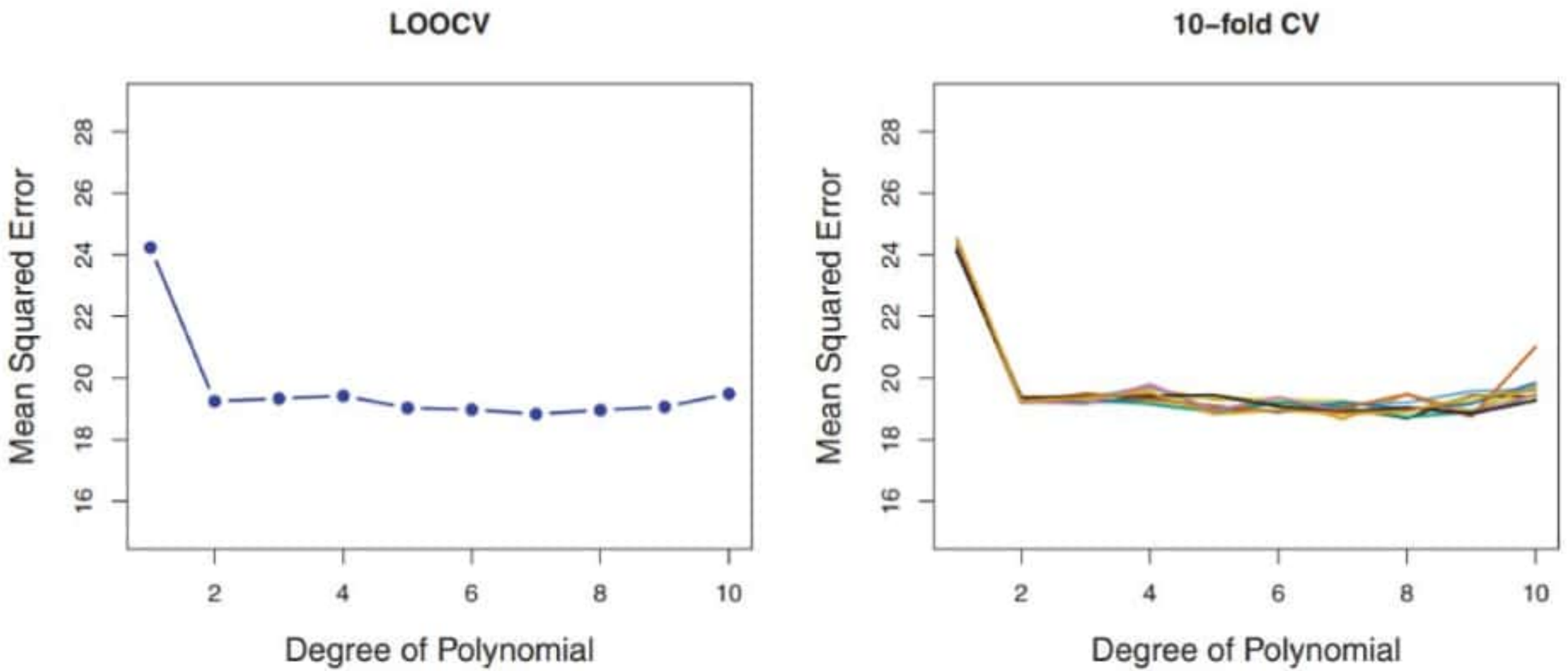


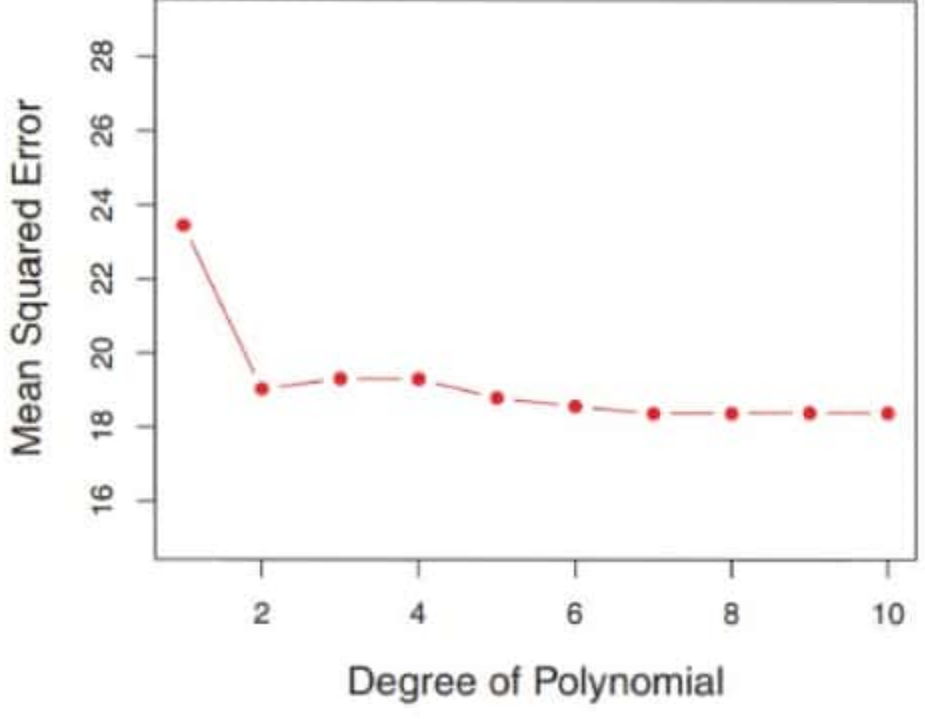
Disadvantages:

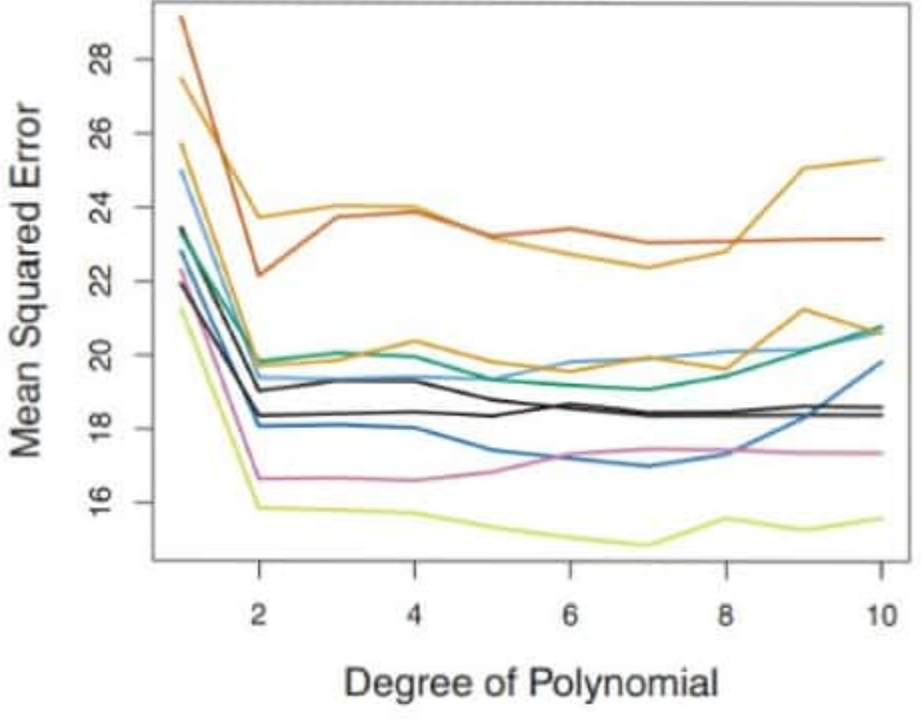
→ Since each training set is (K-1)/K in size, the estimates of prediction error will be higher & have high bias.

→ When K=n, bias decreases, but the estimate have high variance.

→ K=5 to 10 provides a good compromise for bias variance tradeoff.


LOOCV
Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial
10-fold CV
Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial


Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial


Mean Squared Error
16
18
20
22
24
26
28
2
4
6
8
10
Degree of Polynomial

Cross Validation : right and wrong

→ Consider a simple classifier applied to some two class data :
1. Starting with 5000 predictors & 50 samples find 100 predictors with largest correlations with class labels.
2. We then apply a classifier using only these 100 predictors.

★ Can we apply CV in step 2, forgetting about step-1?

NO!

→ This would ignore the fact that in Step 1, the procedure has already seen the labels of training data. & made use of them. This is the form of training & must be included in Validation process.

→ It is easy to simulate realistic data with the class labels independent of outcomes so true test error = 50%, but if we ignore Step-1, CV error estimate will be 0%.

Right Way :

Apply, CV to both steps.

THE BOOTSTRAP:

The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainity associated with a given estimator or statistical learning method.

For example, it can provide an estimate of the Standard Error of a coefficient, or a confidence interval for that coefficient.

EXAMPLE:

→ Suppose that you wish to invest a fixed sum of money in 2 finances that yields X & Y.

→ let α be fraction of money invested @ X.

→ So, we'll try to minimize $(Var(\alpha X + (1-\alpha)Y))$

To minimize the risk α is given by: $\frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y}$

But, σ_X, σ_Y & σ_{XY} are unknown. So, we'll compute estimates for them.

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

→ To estimate $\hat{\alpha}$, will simulate 100 paired samples of X & Y 1000 times.

Now, $\hat{\alpha}$ will be very close to α.

$$\bar{\alpha} = \frac{1}{1000}\sum_{i}^{1000} \hat{\alpha}_i$$

We perform Bootstrap to calculate the Standard Error of these $\hat{\alpha}$ estimates.

$$SE = \sqrt{\frac{1}{1000-1} \sum_{j=1}^{1000} (\hat{\alpha}_j - \bar{\alpha})^2} = 0.0855$$

So, this gives very good estimate of accuracy.

HOW DO WE DO THIS IN REAL WORLD?

Rather than Repeatedly sampling independent datasets from population we instead obtain distinct datasets by repeated sampling of original data with replacement.

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Z_1^* →

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

→ $\hat{\alpha}_1^*$

Z_2^* →

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

→ $\hat{\alpha}_2^*$

Z_B^* →

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

→ $\hat{\alpha}_B^*$

Date :
Page :

BOOTSTRAP IN GENERAL:

In complex data situations, figuring out appropriate way to generate bootstrap can require some thought.

FOR EXAMPLE: if a data is in time-series, we can simply sample the population with replacement.
Because, we ~~@~~ our data points are dependent on previous datapoints, we expect them to be correlated.

* So, rather than doing that, we can divide data in blocks & than sample that.

OTHER USES OF BOOTSTRAP:

→ Primary use is to calculate Standard Error of the estimate.

→ Can be used to calculate the confidence interval of our data. (Coz by estimates & Random Sampling we can create an empirical Sampling Distribution & than problem solved)

Can Bootstrap estimate predict error?

No, because of a lot of overlap in train & validation data: we can rather use K-fold CV.