Cross Validation

content layout

- Introduction
- Methods of Cross Validation
  - Test set method
  - Leave one out cross validation (LOOCV)
  - k-fold cross validation
- Example
- Spss programing

What is Cross Validation?

- Cross Validation is a technique which involves reserving a particular sample of a dataset on which you do not train the model. Later, you test your model on this sample before finalizing it.
- Here are the steps involved in cross validation:
- You *reserve* a sample data set
- Train the model using the remaining part of the dataset
- Use the reserve sample of the test (validation) set. This will help you in gauging the effectiveness of your model's performance. If your model delivers a positive result on validation data, go ahead with the current model. It rocks!

Methods of Cross Validation

- A few common methods used for Cross Validation

There are various methods available for performing cross validation. I've discussed a few of them in this section.

- The test set method
- Leave one out cross validation (LOOCV)
- k-fold cross validation

The test set method

- 1. Randomly choose 30% of the data to be in a test set
- 2. The remainder is a training set
- 3. Perform your regression on the training set
- 4. Estimate your future performance with the test

Leave one out cross validation (LOOCV)

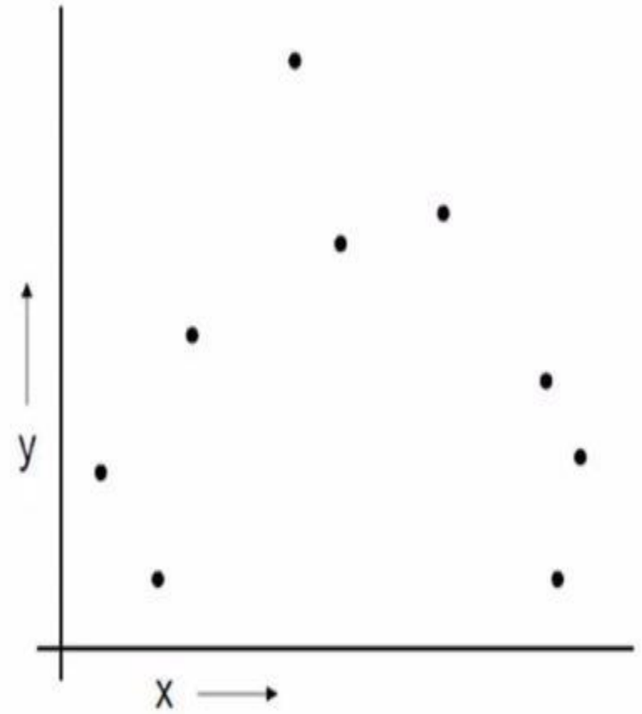
- In this approach, we reserve only one data point from the available dataset, and train the model on the rest of the data. This process iterates for each data point. This also has its own advantages and disadvantages. Let's look at them:
- We make use of all data points, hence the bias will be low
- We repeat the cross validation process n times (where n is number of data points) which results in a higher execution time
- This approach leads to higher variation in testing model effectiveness because we test against one data point. So, our estimation gets highly influenced by the data point. If the data point turns out to be an outlier, it can lead to a higher variation

k-fold cross validation

- Randomly split your entire dataset into k"folds"
- For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for *kth* fold
- Record the error you see on each of the predictions
- Repeat this until each of the k-folds has served as the test set
- The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the mode

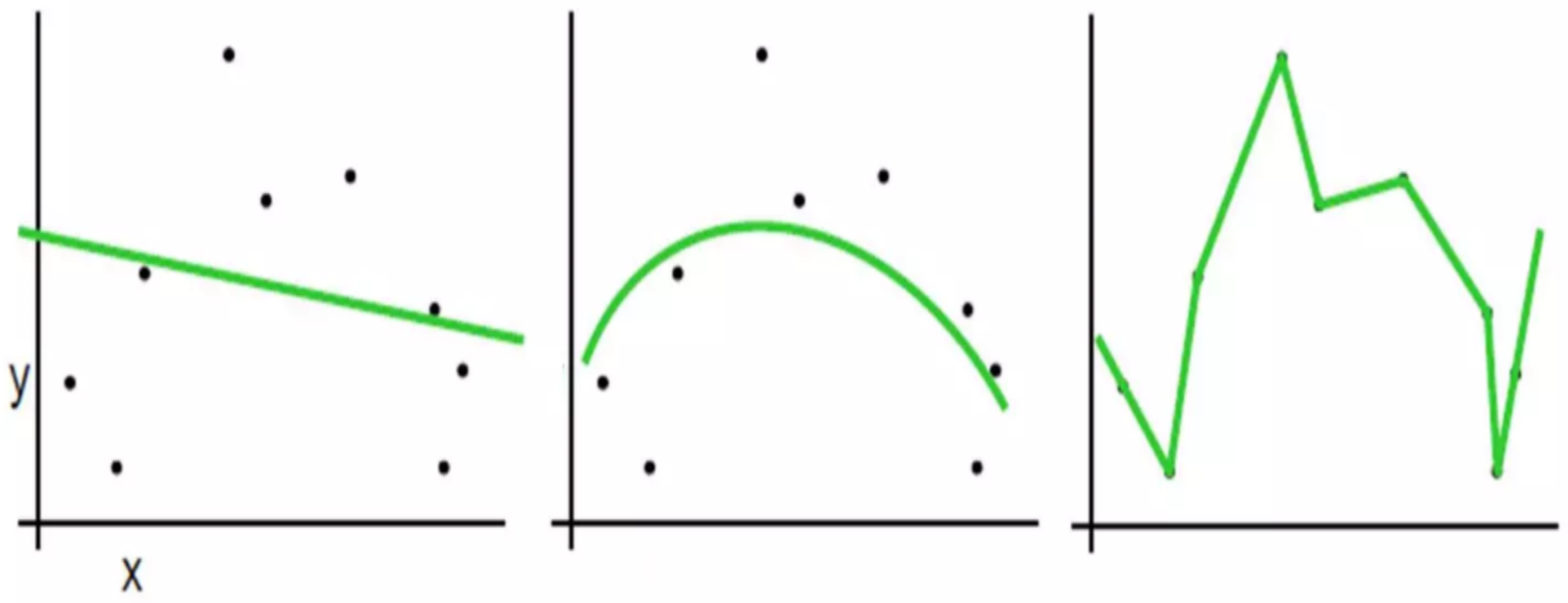
Example

- Let us find best representation of the data from a test for the relationship between water cement ration and the compressive strength of concrete as below:



X=w/c %	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55
Y=Compressive strength Mpa	15	10	20	40	25	30	18	16	10

Which is Best ?

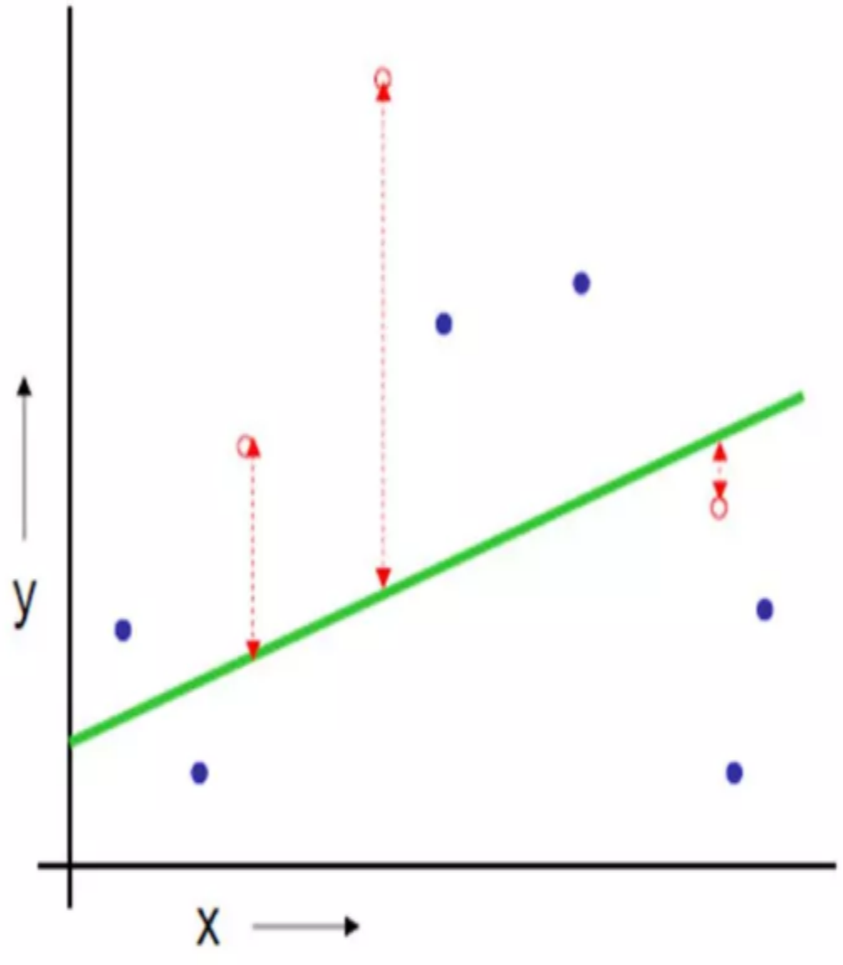


Linear regression

Quadratic regression

Join the dots

The test set method



1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

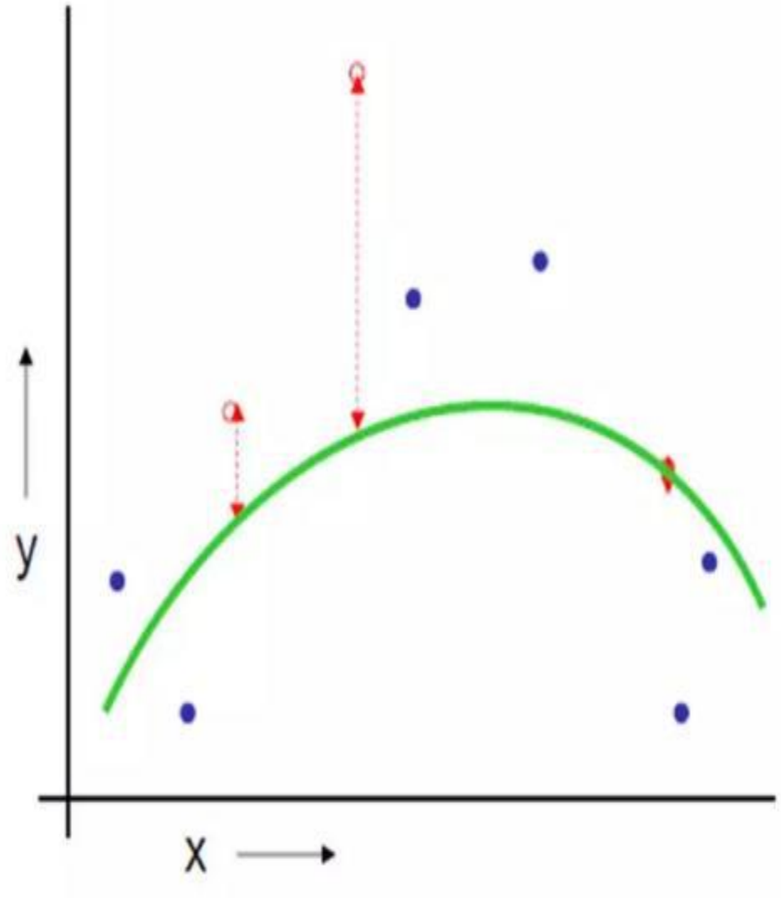
3. Perform your regression on the training set

(Linear regression example)

Mean Squared Error = 2.4

4. Estimate your future performance with the test set

The test set method



1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

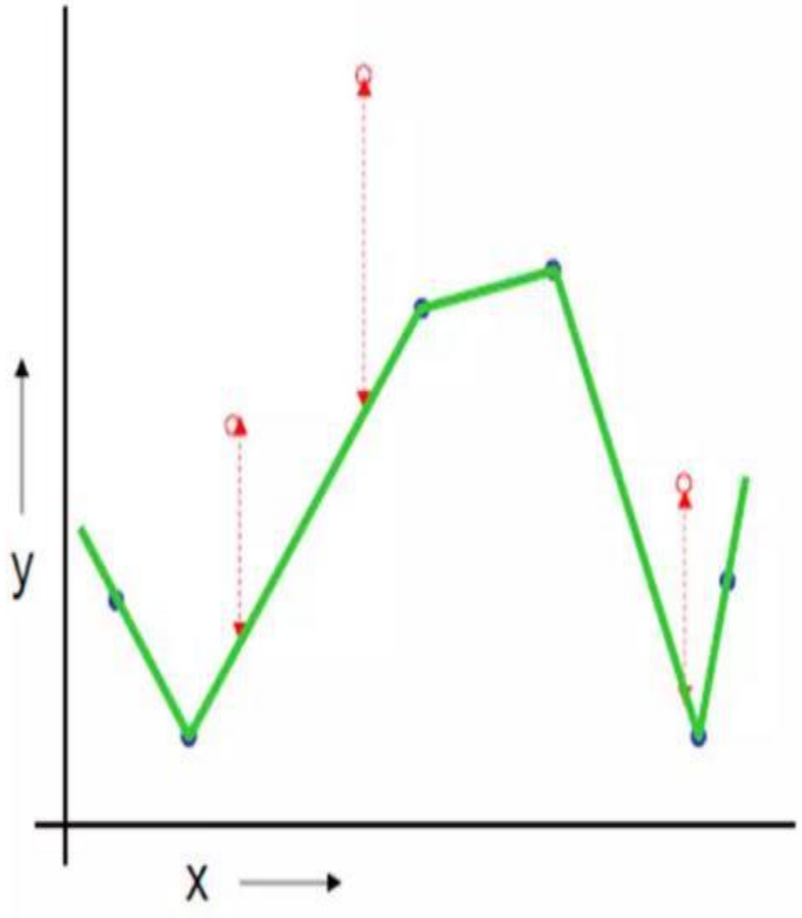
3. Perform your regression on the training set

(Quadratic regression example)

Mean Squared Error = 0.9

4. Estimate your future performance with the test set

The test set method



1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

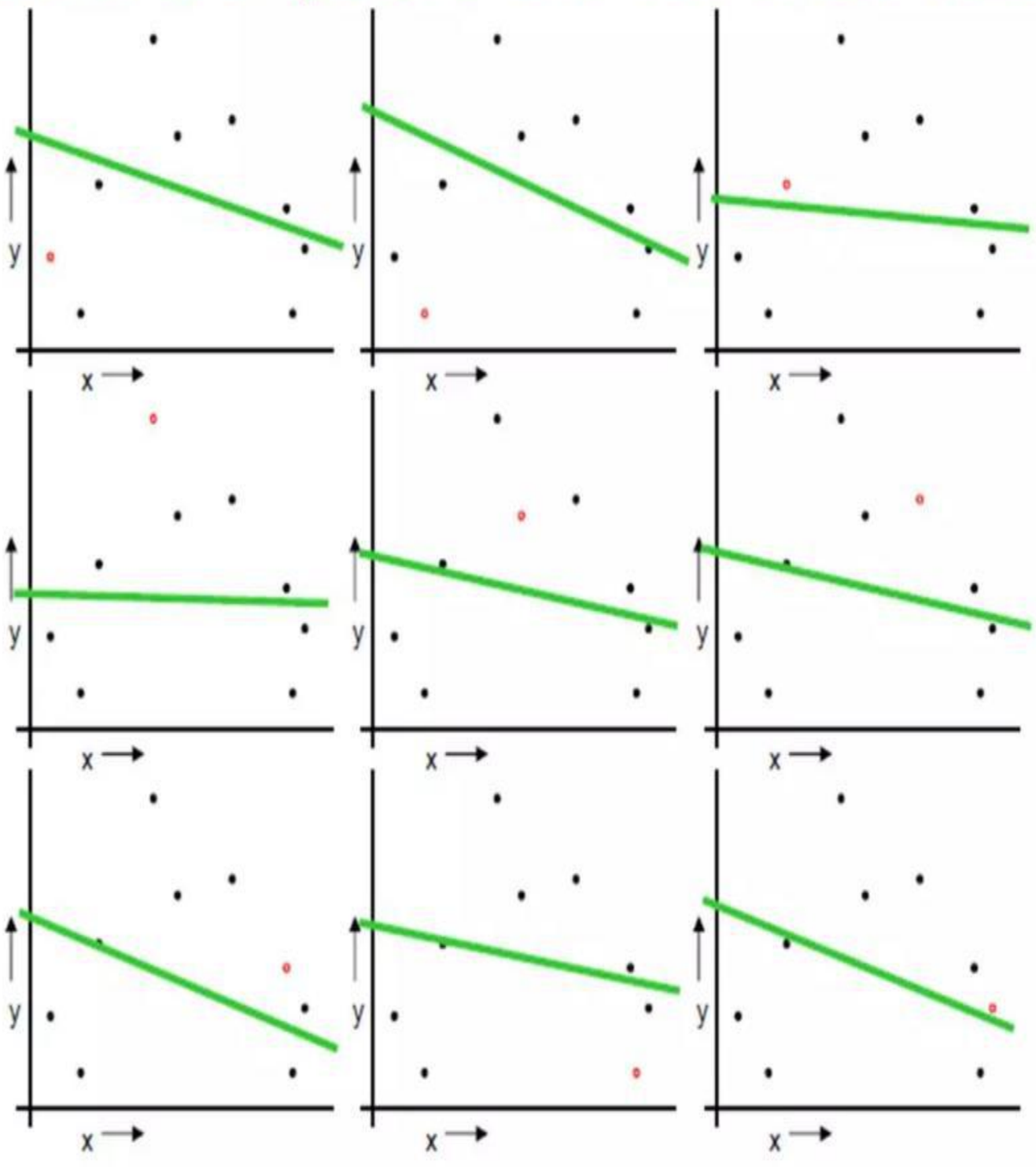
3. Perform your regression on the training set

(Join the dots example)

Mean Squared Error = 2.2

4. Estimate your future performance with the test set

LOOCV (Leave-one-out Cross Validation)



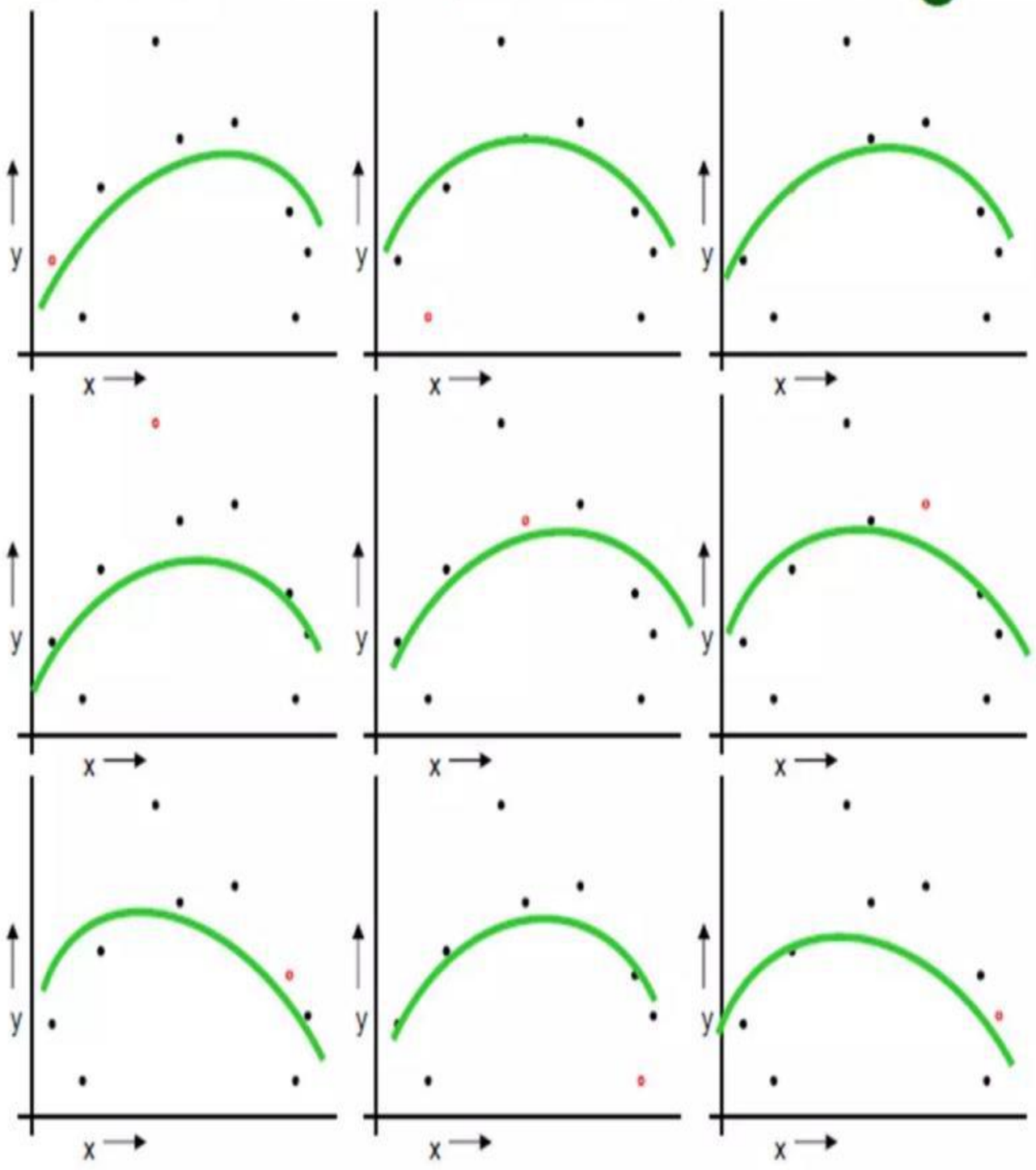
For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV} = 2.12$

LOOCV for Quadratic Regression



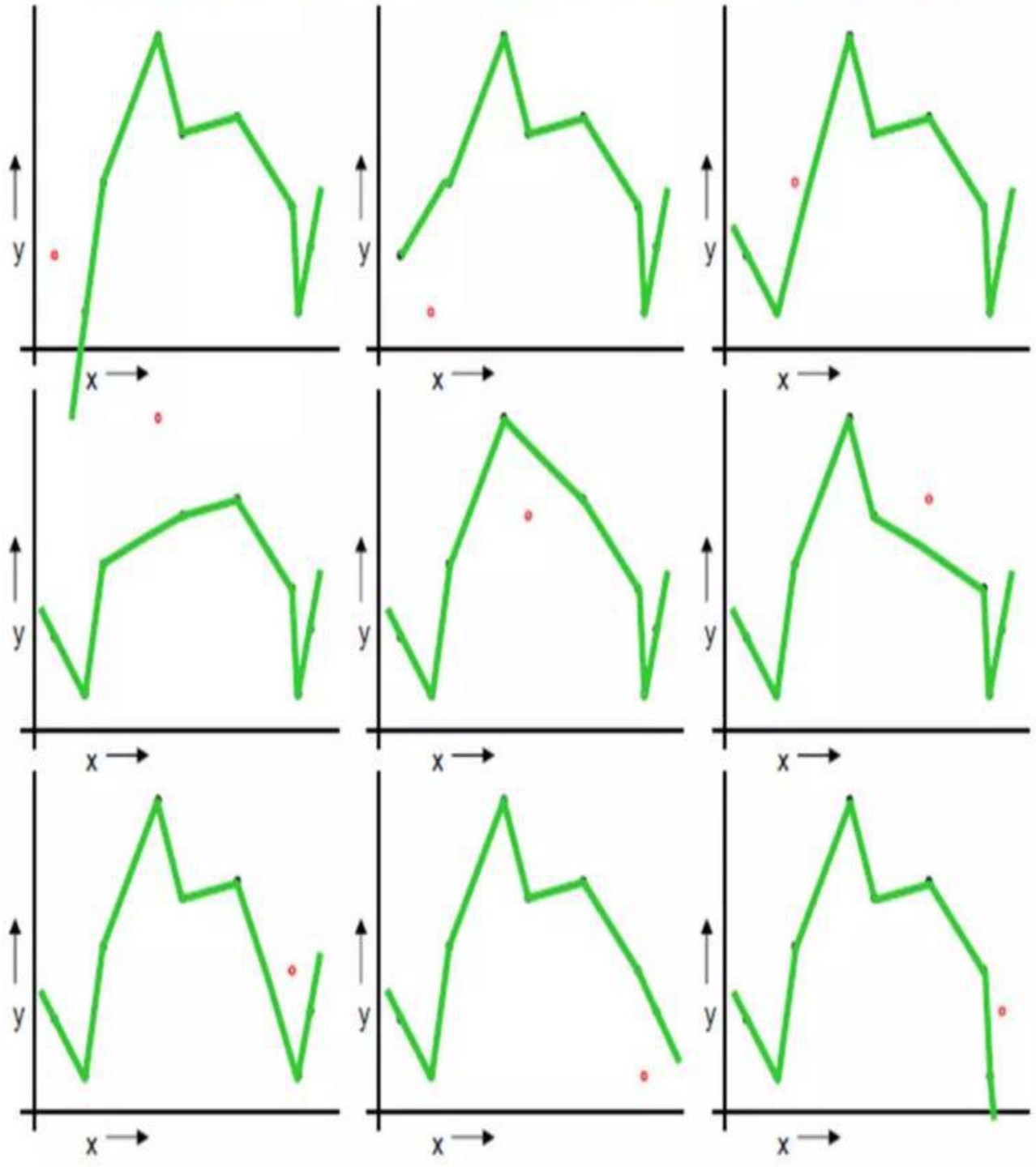
For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

When you've done all points, report the mean error.

$MSE_{LOOCV} = 0.962$

LOOCV for Join The Dots



For k=1 to R

1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)

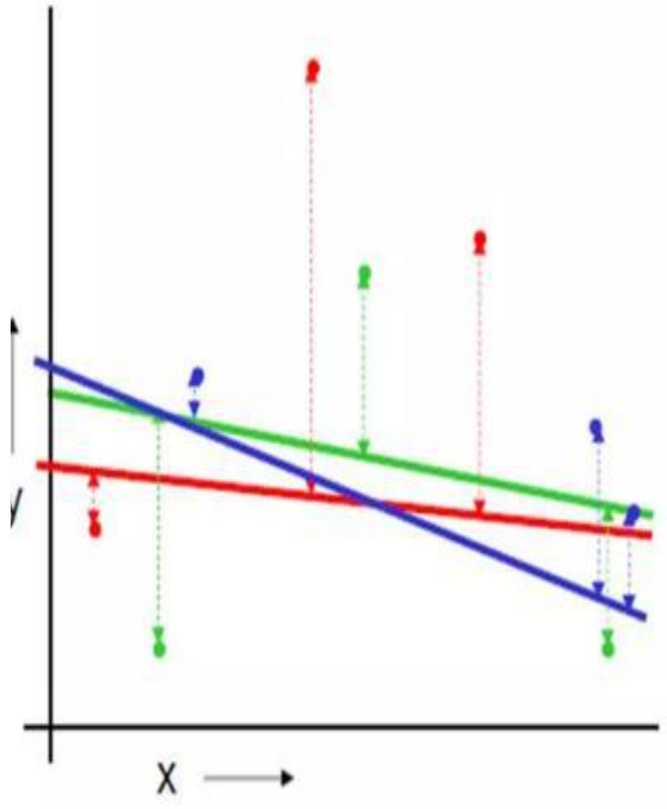
When you've done all points, report the mean error.

$MSE_{LOOCV} = 3.33$

k-fold Cross Validation

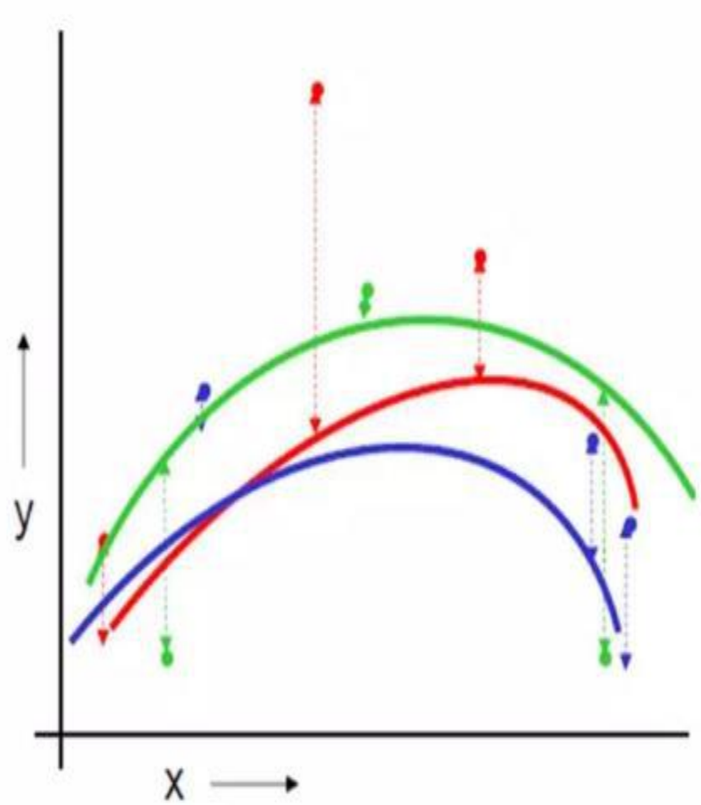
- Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue).
- For every partition: Train on all the points not in the same partition.
- Find the test-set sum of errors on the taken points.
- Then report the mean error

k-fold Cross Validation



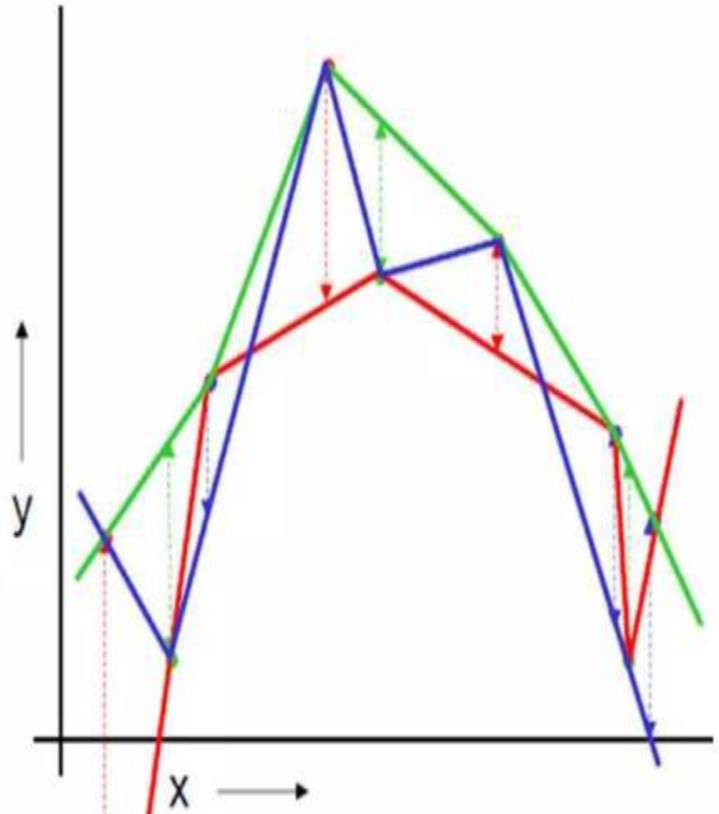
Linear Regression

$MSE_{3FOLD}=2.05$



Quadratic Regression

$MSE_{3FOLD}=1.11$



Joint-the-dots

$MSE_{3FOLD}=2.93$

Which kind of Cross Validation?

	downside	upside
Test - set	Variance: unreliable estimate of future performance	Cheap
Leave one-out	Expensive. Has some weird behavior	Doesn't waste data
K-fold	Waster than LOOCV Expensive than Test set	Slightly better than Test set