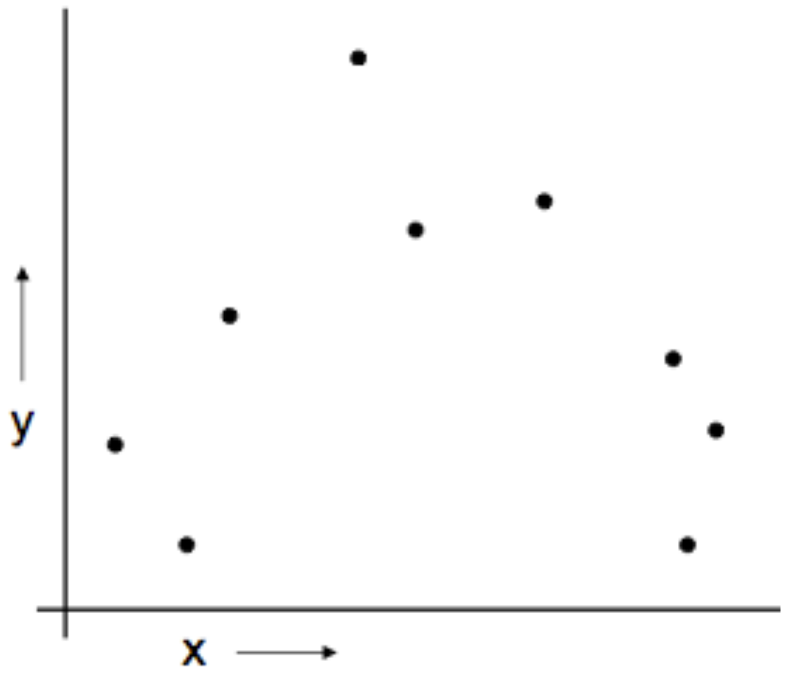
Cross-validation for model selection

Outline

- Test-set cross-validation
- Leave-one-out cross-validation
- k-fold cross-validation

A Regression Problem



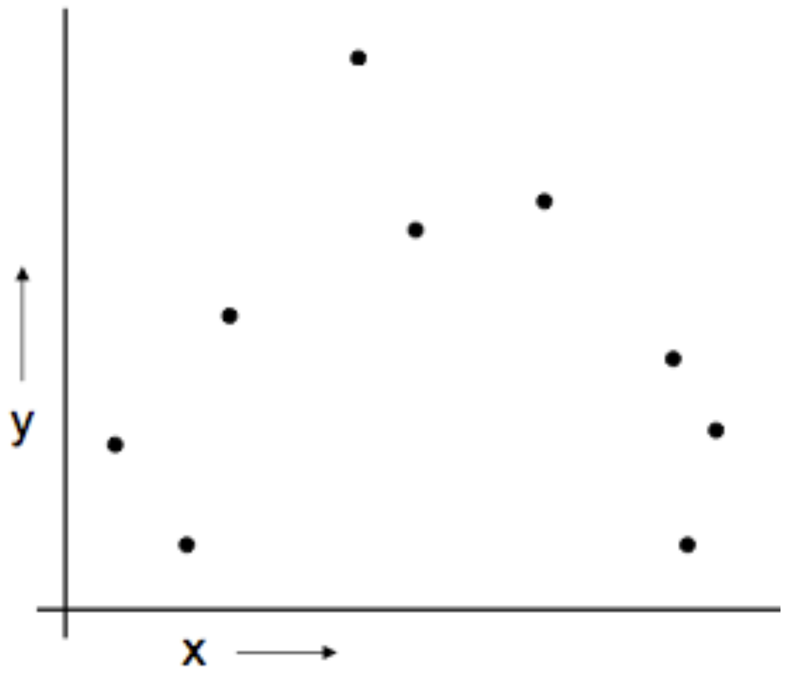
Regression

- a statistical process for estimating the relationships among variables.

Regression vs. classification

- Regression: the output variable takes continuous values.
- Classification: the output variable takes class labels.

A Regression Problem



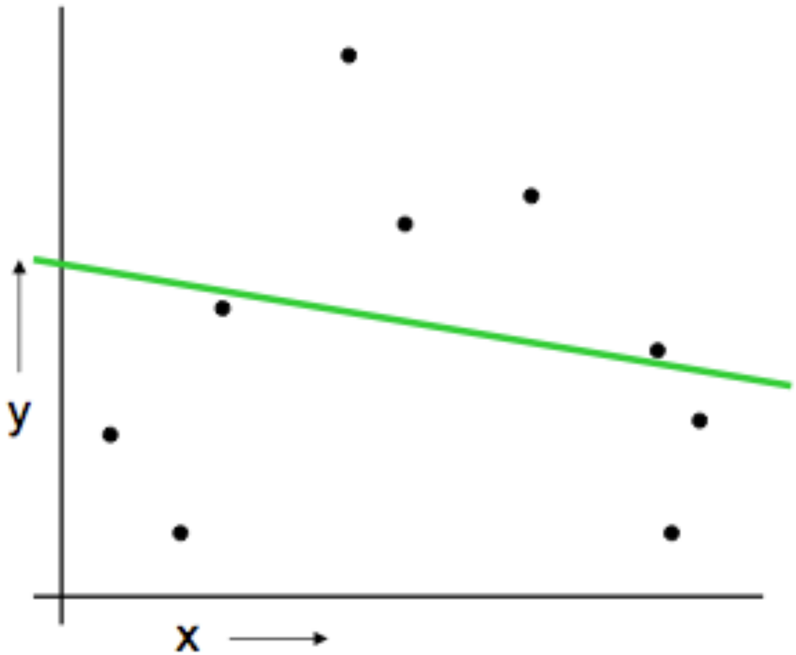
$y = f(x) + \text{noise}$

Can we learn **f** from this data?

Let's consider three methods

- Linear regression
- Quadratic regression
- Linear non-parametric regression

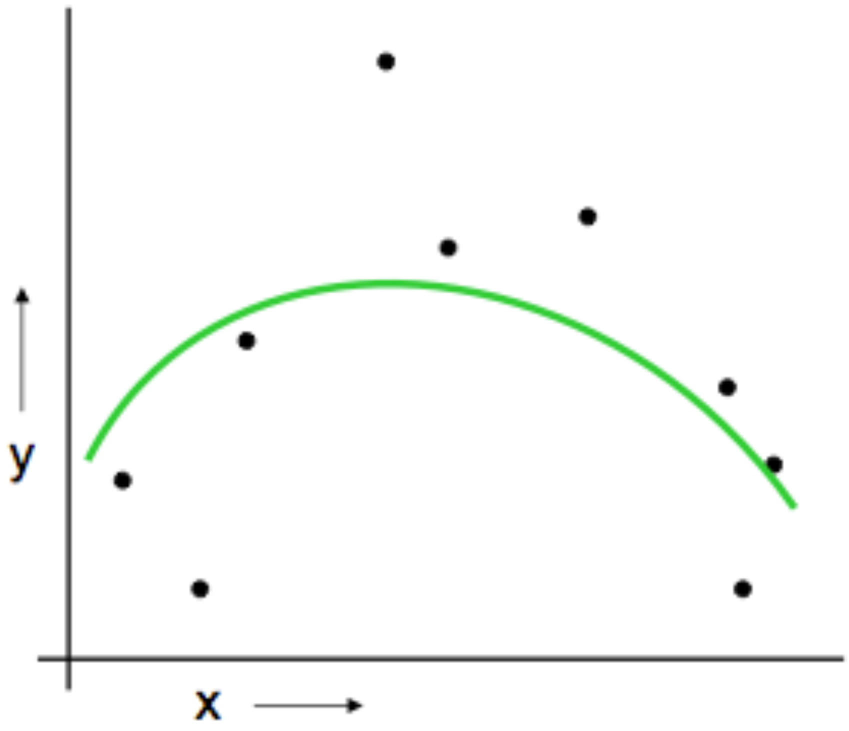
Linear Regression



Linear regression:

- an approach to model the relationship between a scalar dependent variable **y** and one or more explanatory variables denoted **X**

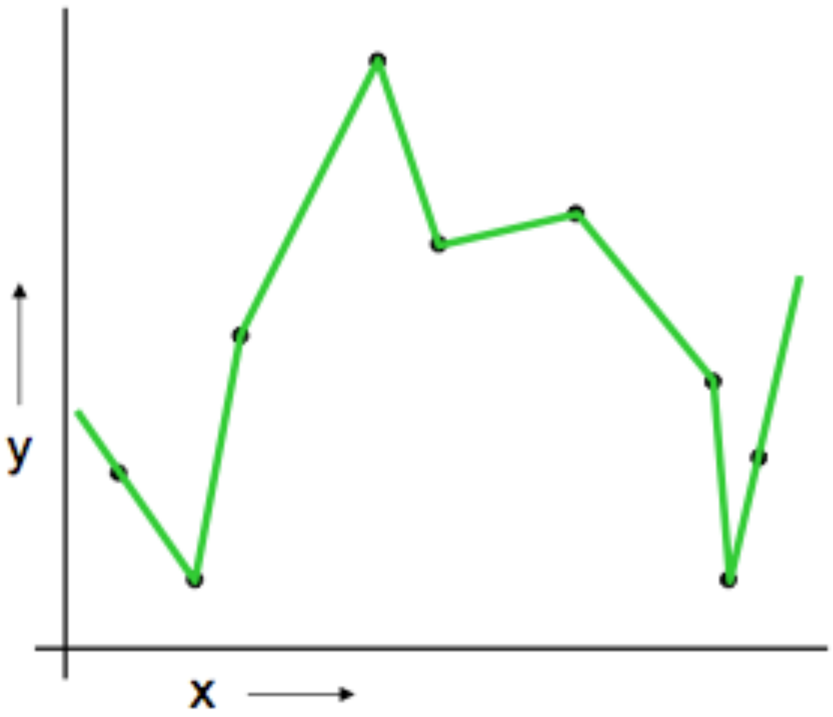
Quadratic Regression



Quadratic regression:

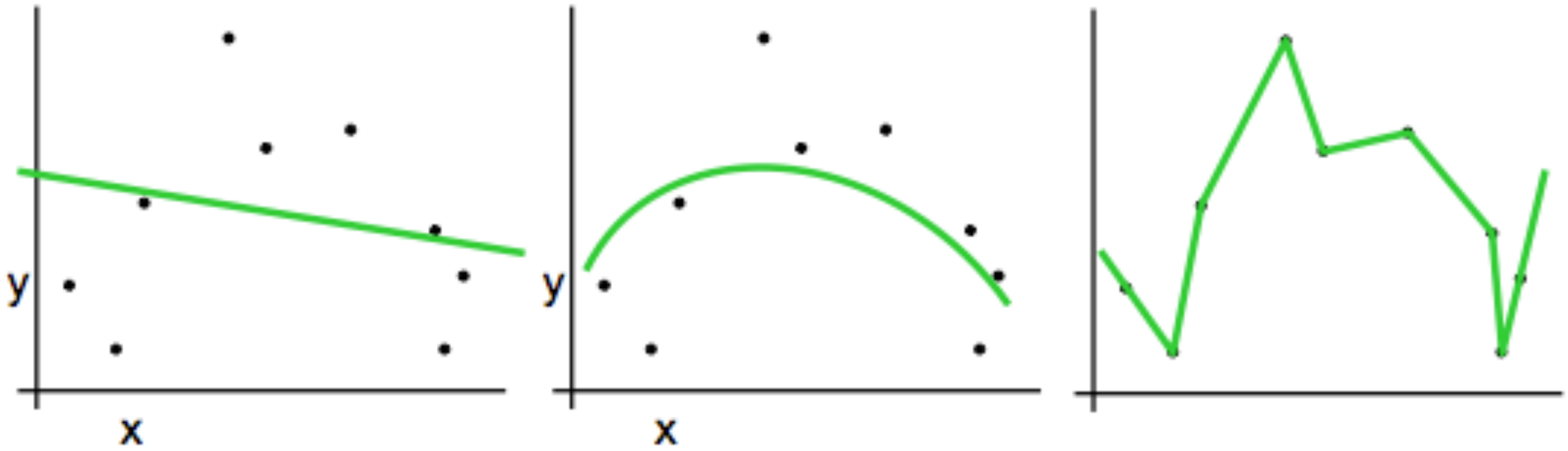
- the process of finding the equation of the parabola that fits best for a set of data

Join-the-dots



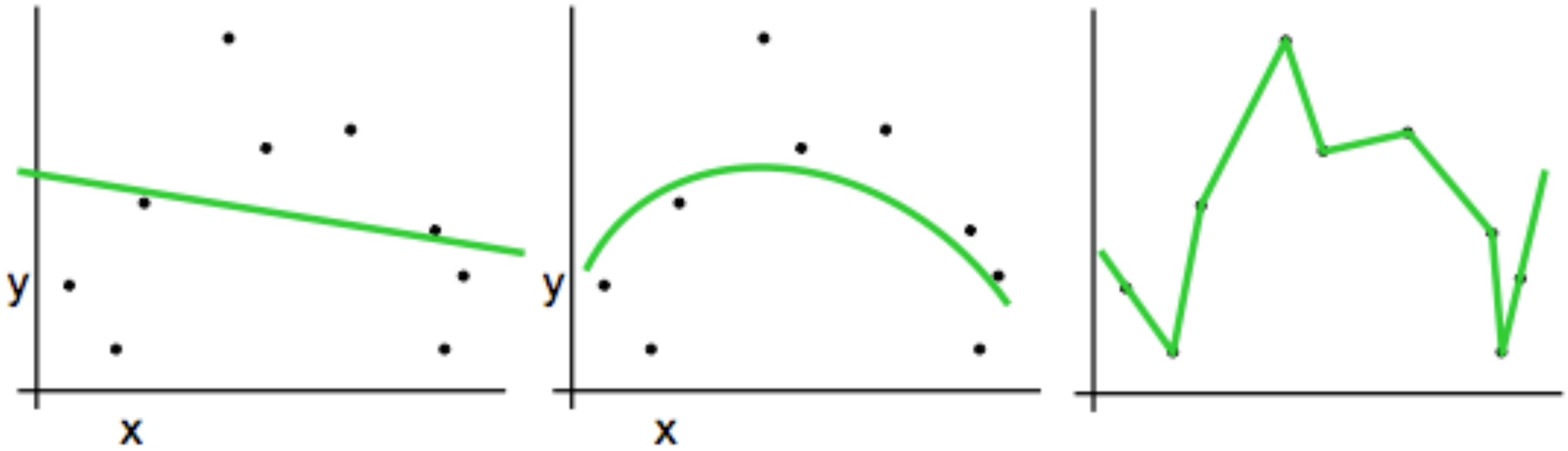
Also known as piecewise linear nonparametric regression

Which is best?



How to choose the method with the best fit to the data?

What do we really want?



How to choose the method with the best fit to the data?

How well a model is going to predict future data drawn from the same distribution?

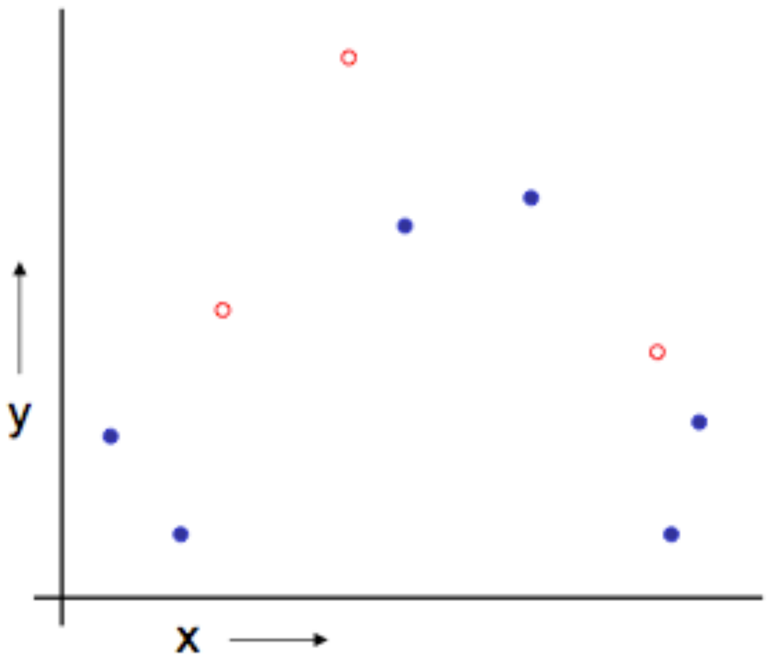
Mean Squared Error

Mean Squared Error (MSE)

- one of many ways to quantify the difference between values implied by a model (aka estimator) and the true values of the quantity being estimated
- Commonly used in regression analysis

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

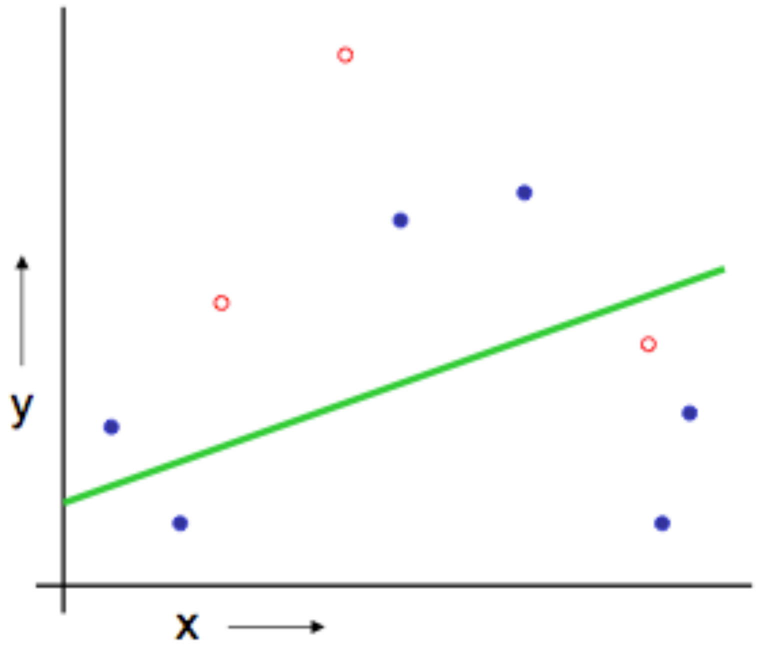
The test set method



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

2. The remainder is a training set

The test set method



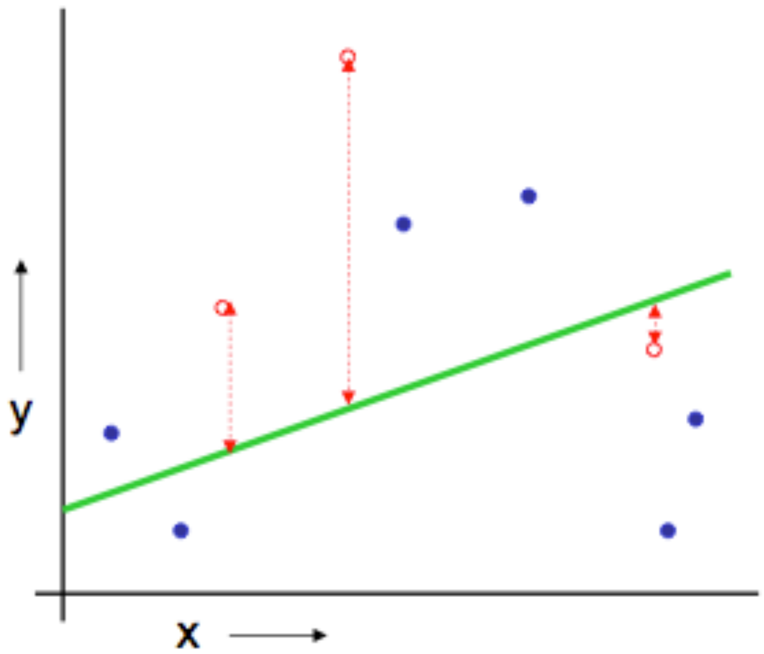
(Linear regression example)

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

The test set method



(Linear regression example)

Mean Squared Error = 2.4

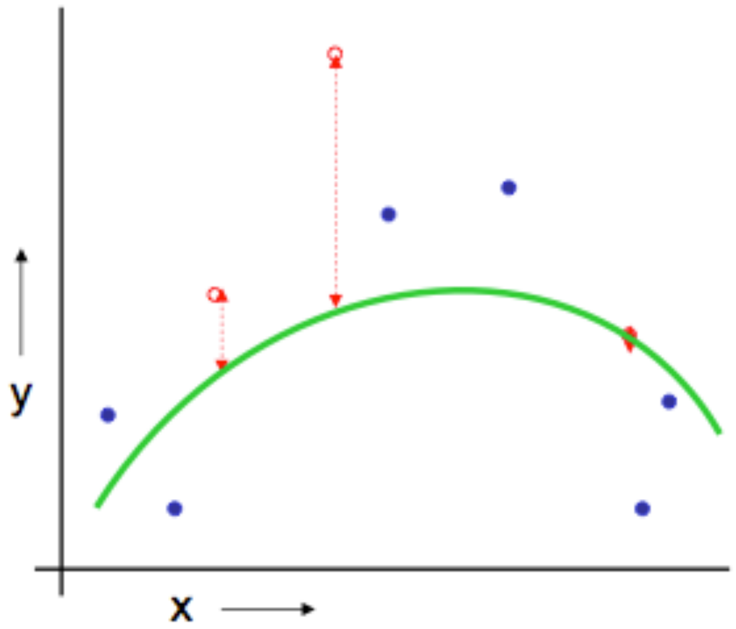
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

The test set method



(Quadratic regression example)

Mean Squared Error = 0.9

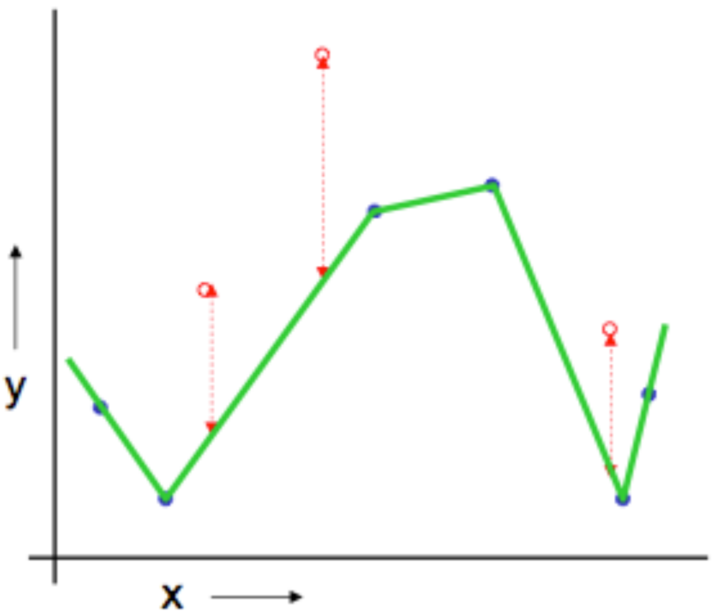
1. Randomly choose 30% of the data to be in a test set

2. The remainder is a training set

3. Perform your regression on the training set

4. Estimate your future performance with the test set

The test set method



(Join the dots example)

Mean Squared Error = 2.2

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 set

The test set method

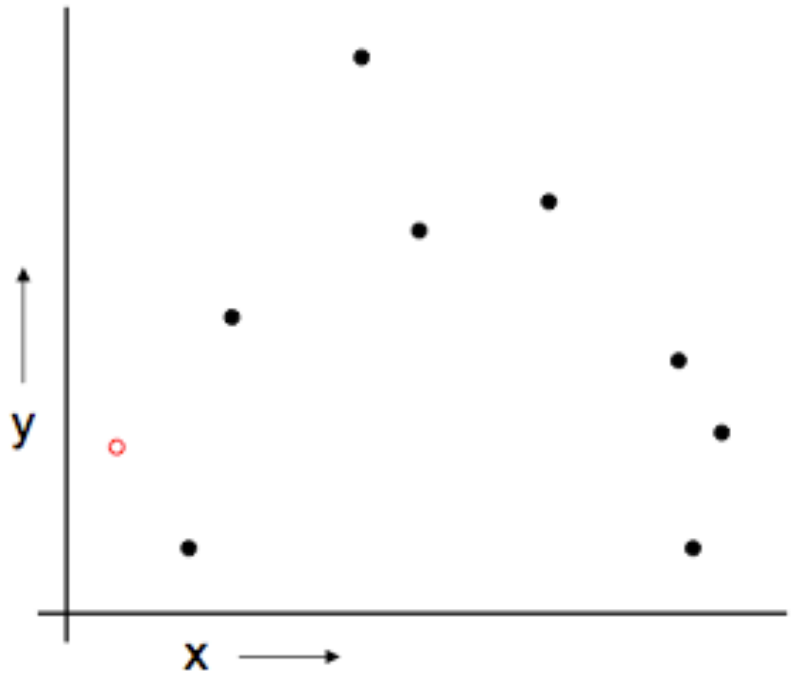
Good news

- Very very simple
- Can then simply choose the method with the best test-set score

Bad news

- Wastes data: we get an estimate of the best method to apply to 30% less data
- If we don't have much data, our test-set might just be lucky or unlucky

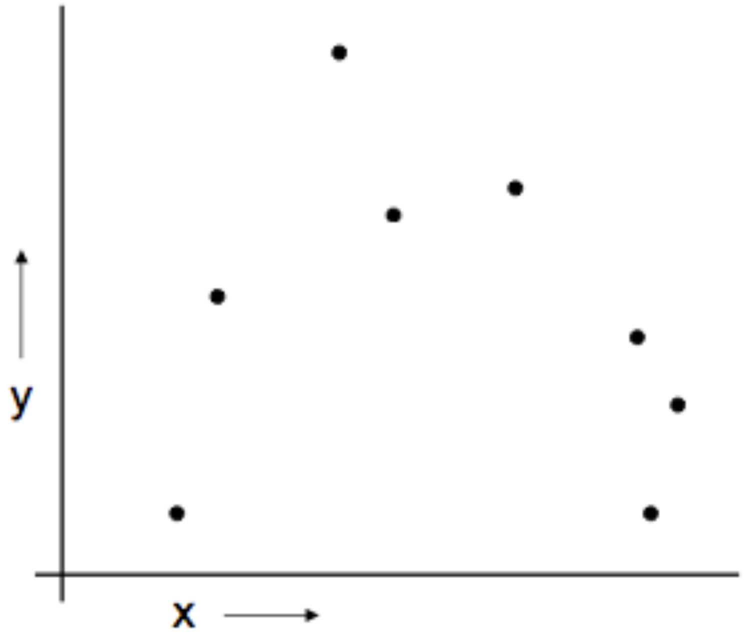
Leave-one-out Cross Validation(LOOCV)



For k=1 to R

1. Let (x_k,y_k) be the k^{th} record

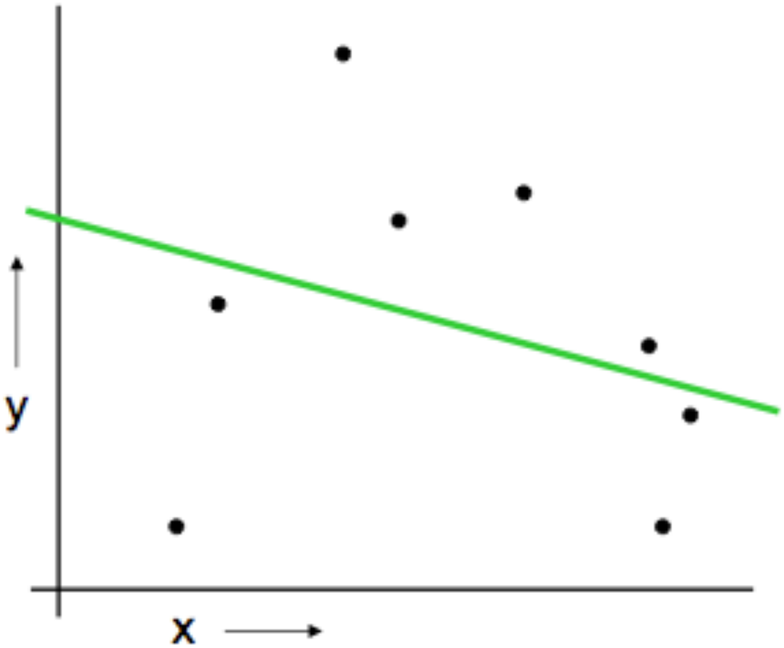
Leave-one-out Cross Validation(LOOCV)



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

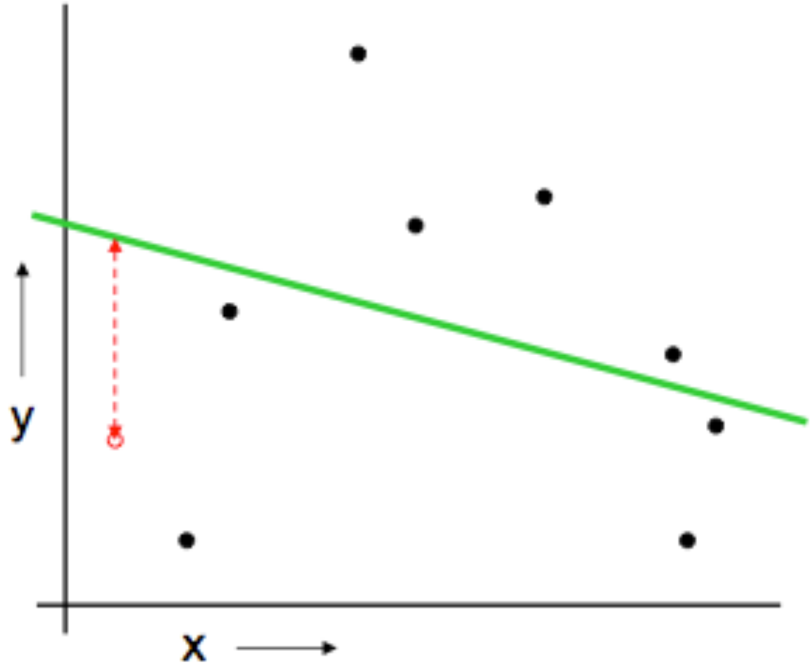
Leave-one-out Cross Validation(LOOCV)



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

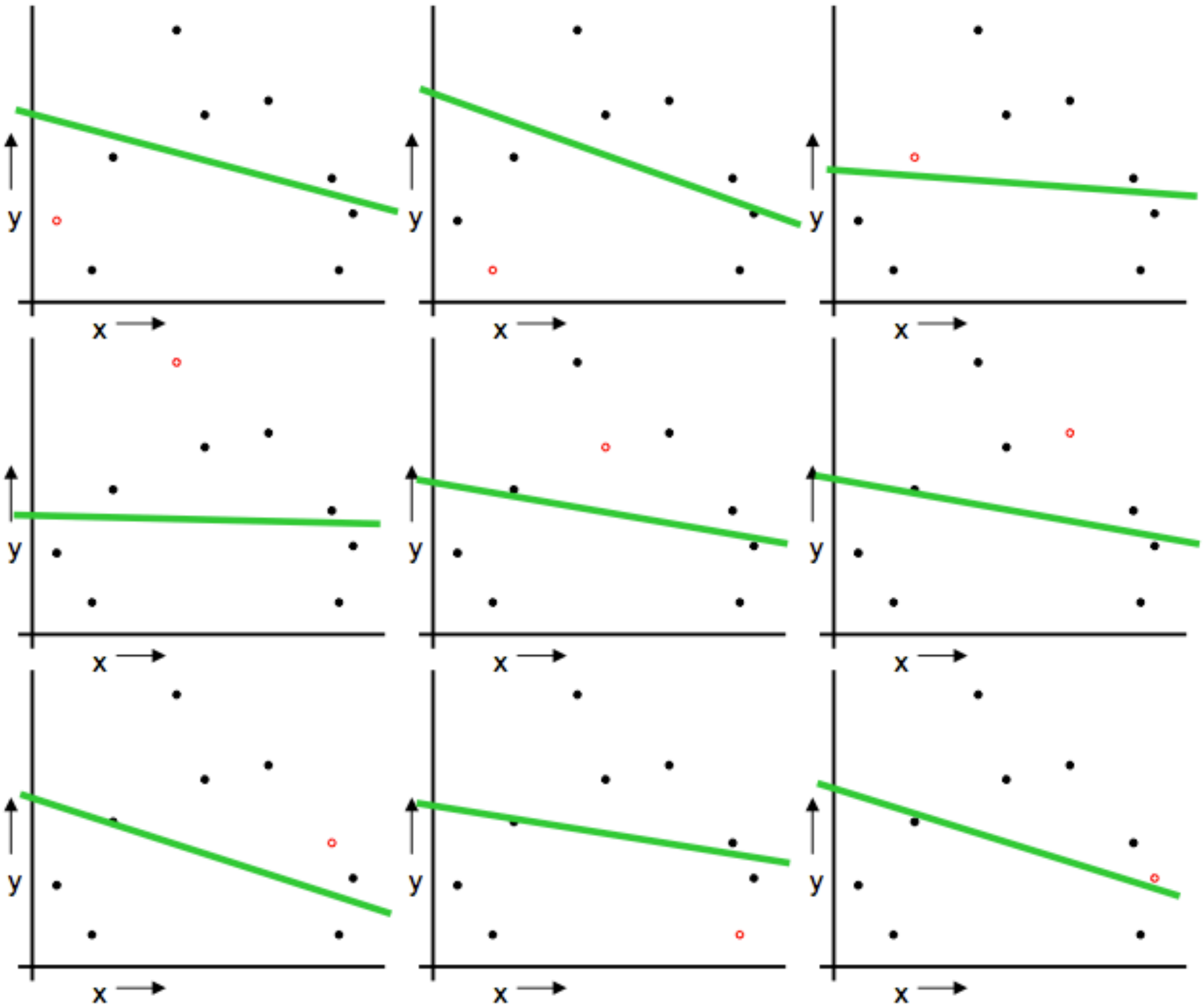
Leave-one-out Cross Validation(LOOCV)



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)

Leave-one-out Cross Validation(LOOCV)



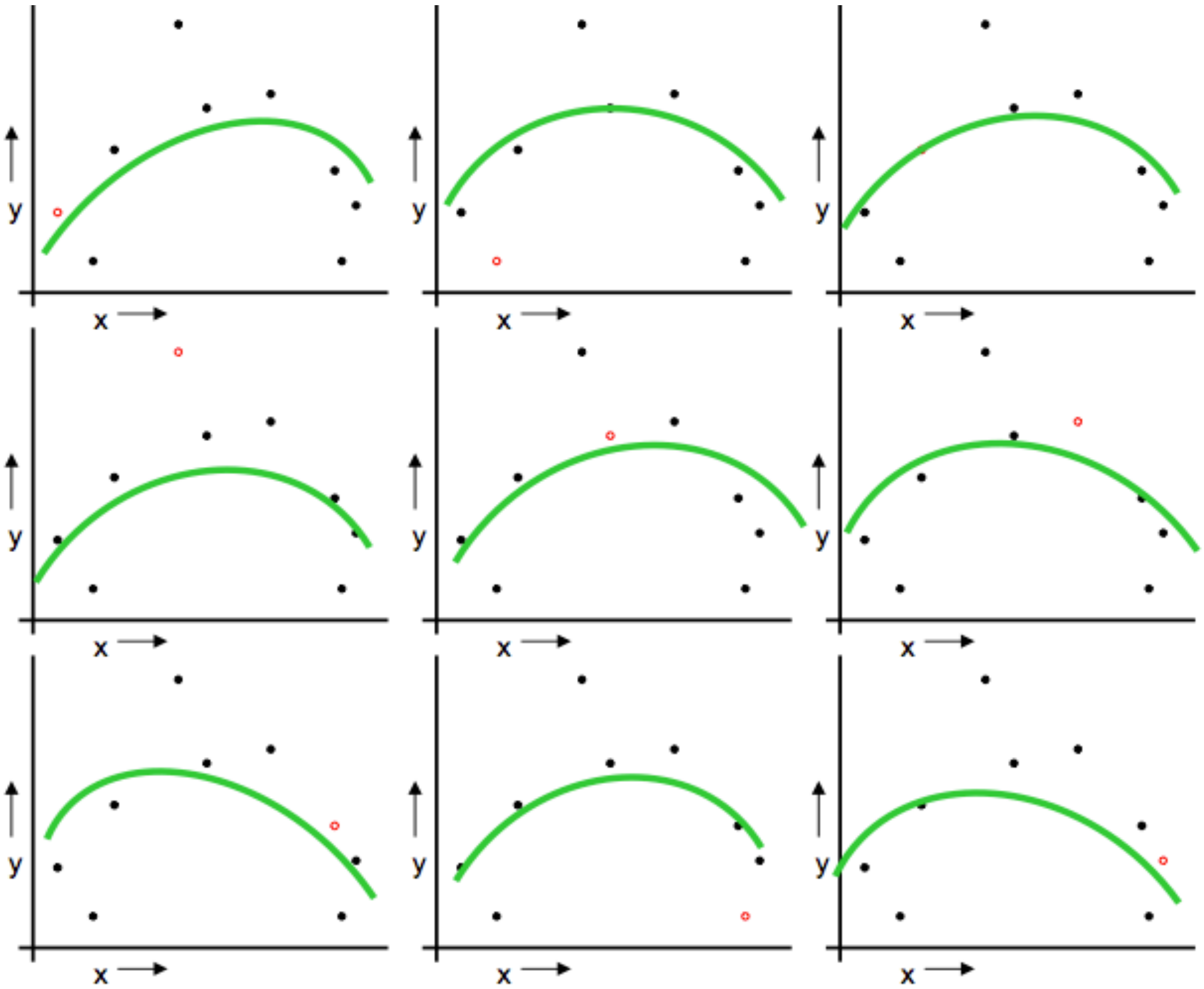
For k=1 to R

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



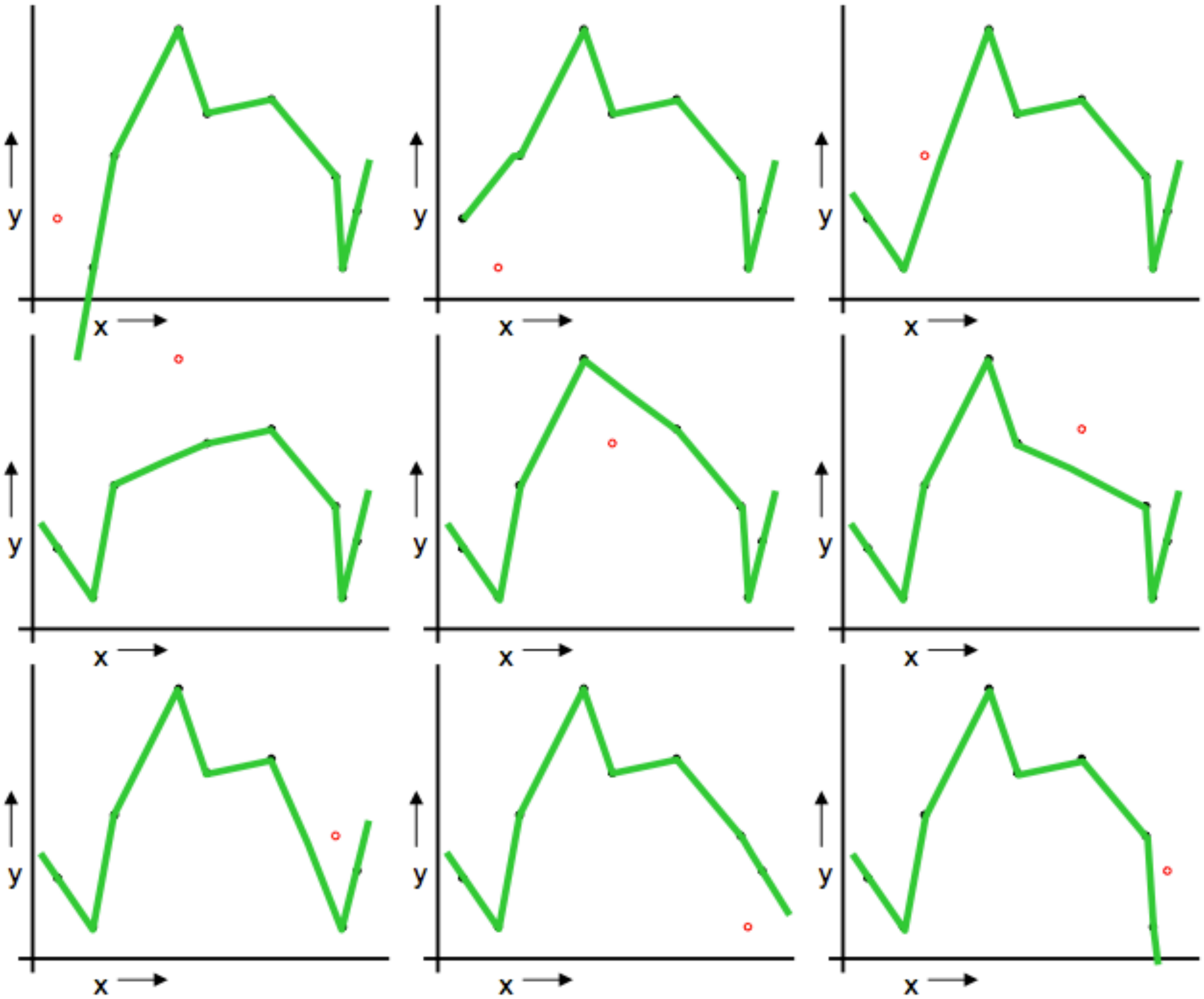
For k=1 to R

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 Non-Parametric 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)

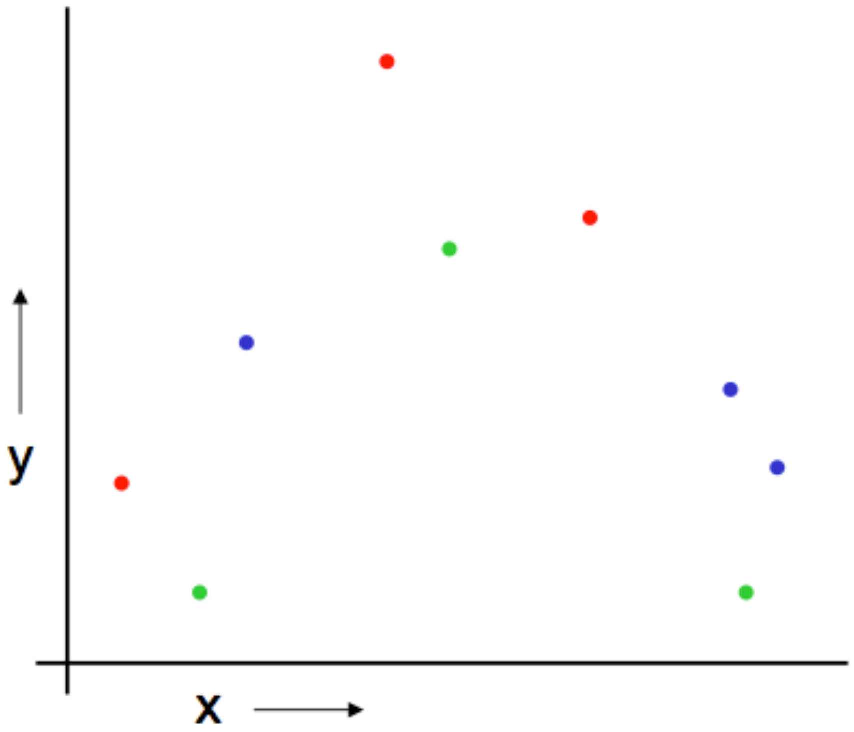
When you've done all points, report the mean error.

$$MSE_{LOOCV} = 3.33$$

Which kind of validation?

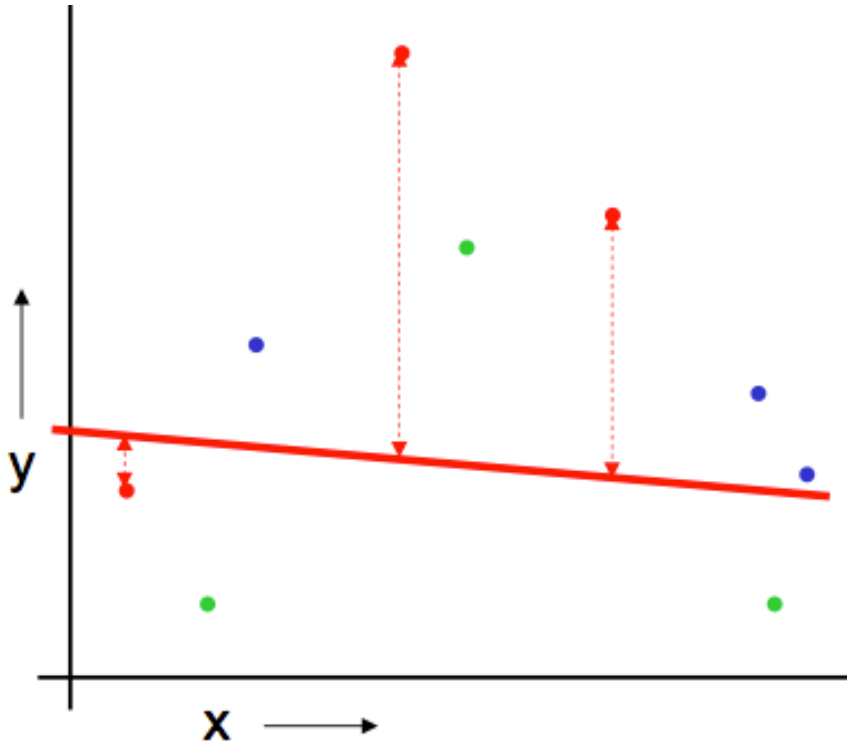
	Downside	Upside
Test-set	Variance: unreliable estimate of future performance	Cheap
Leave-one-out	Expensive.	Doesn't waste data

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)

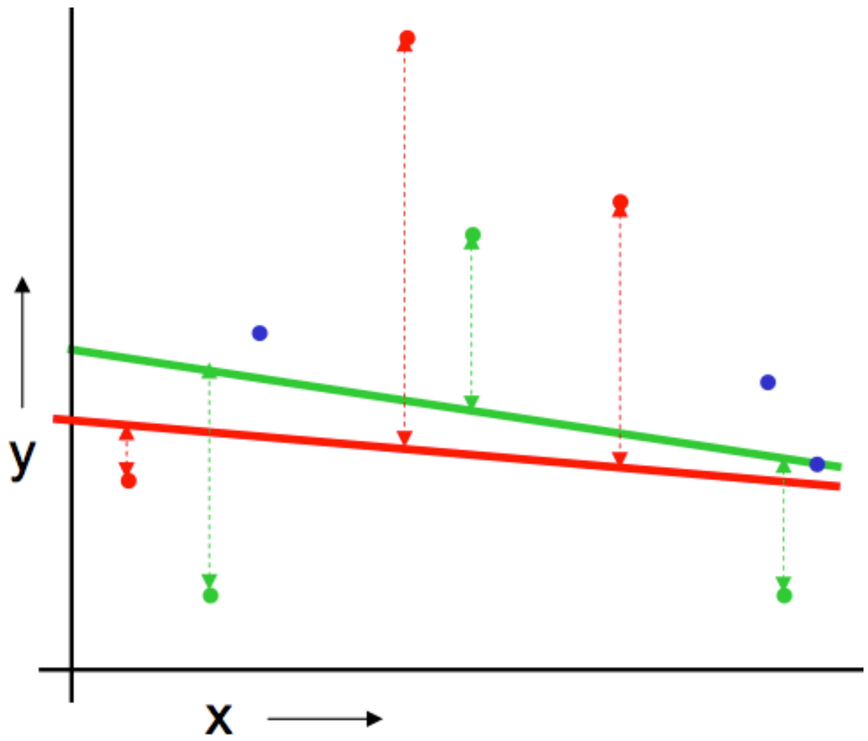
K-fold Cross validation



Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

K-fold Cross validation

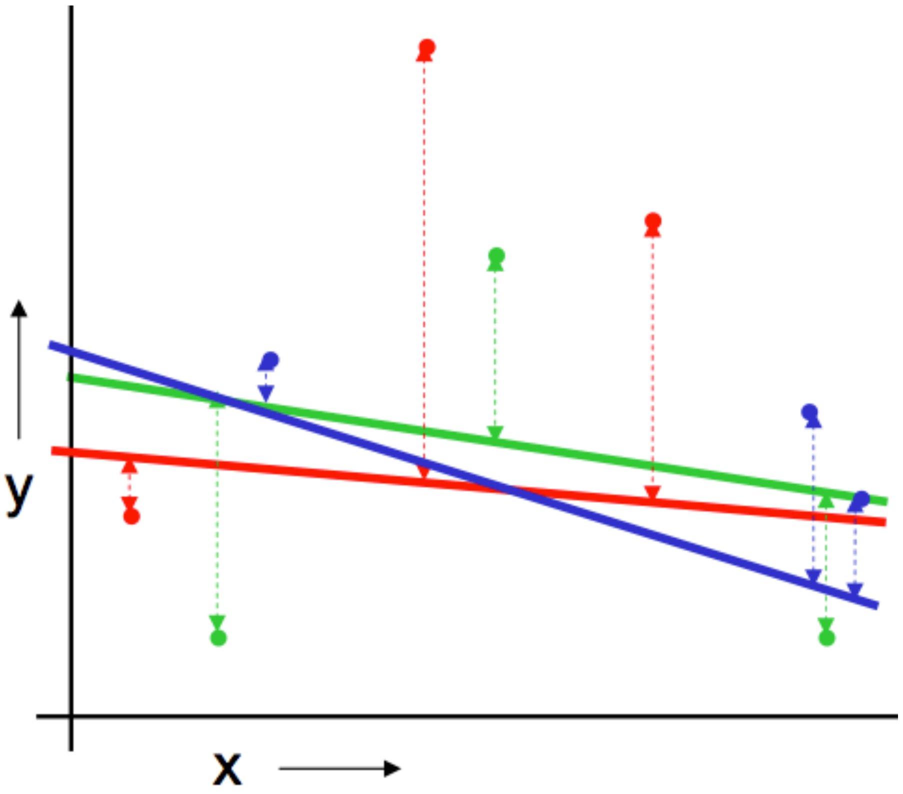


Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

K-fold Cross validation



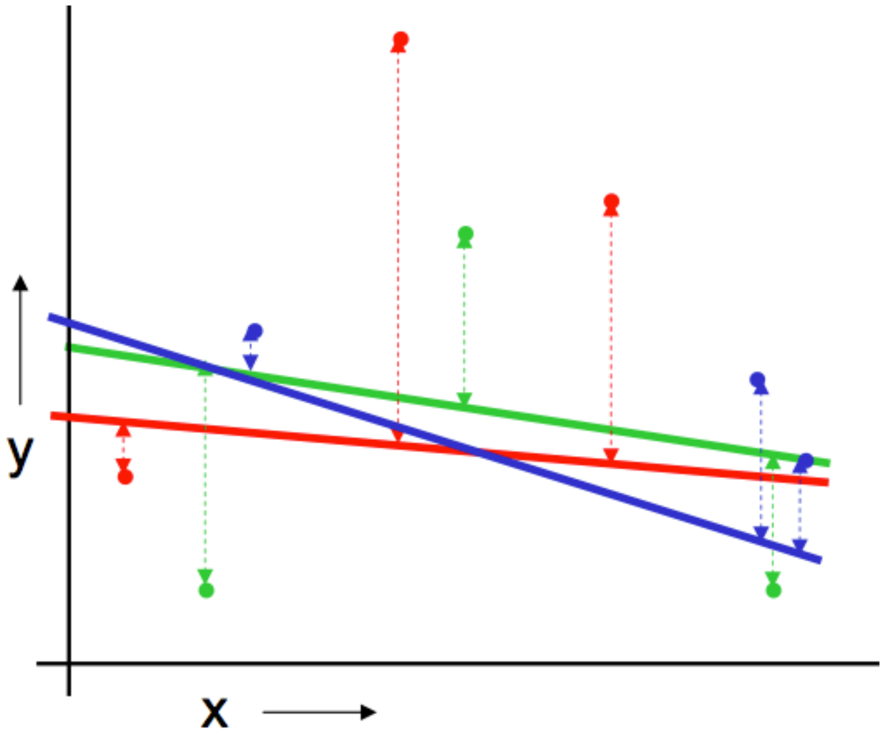
Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

K-fold Cross validation



Linear Regression
MSE_{3FOLD}=2.05

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

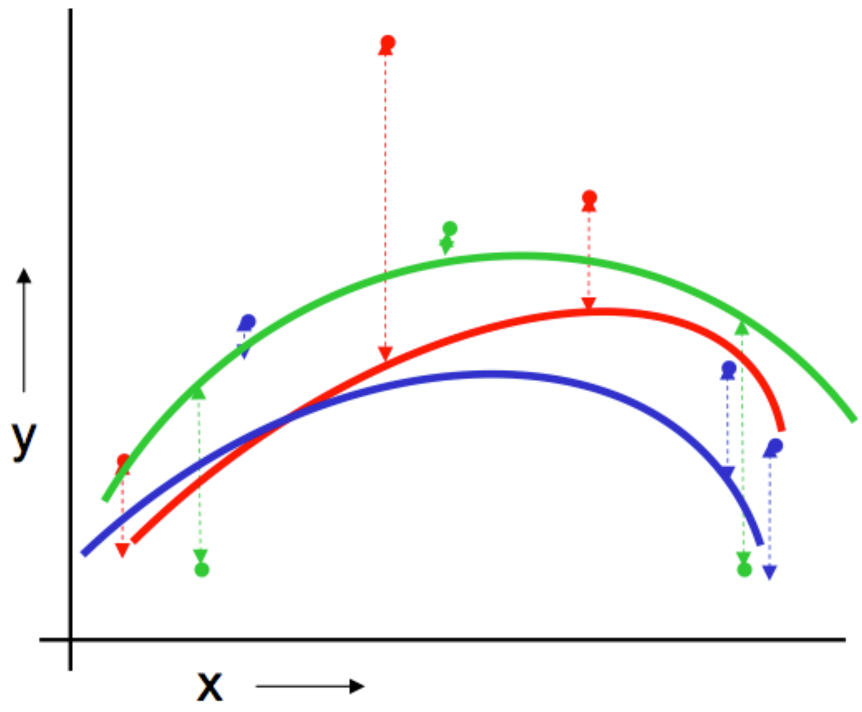
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

K-fold Cross validation



Quadratic Regression
$MSE_{3FOLD}=1.11$

Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red Green and Blue)

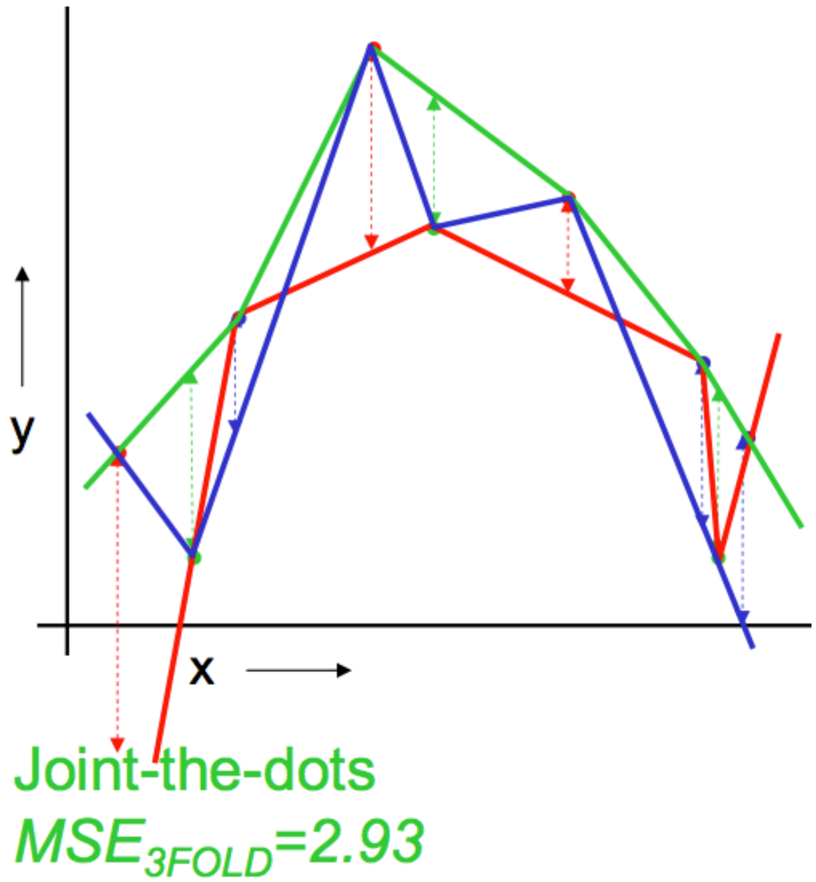
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

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 the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

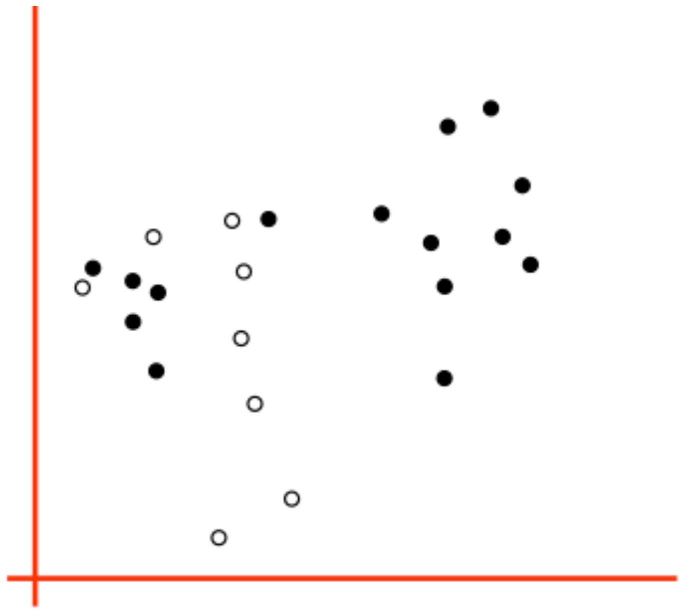
Then report the mean error

Which kind of validation

	Downside	Upside
Test-set	Variance: unreliable estimate of future performance	Cheap
Leave-one-out	Expensive.	Doesn't waste data
10-fold	Wastes 10% of the data. 10 times more expensive than test set	Only wastes 10%. Only 10 times more expensive instead of R times.
3-fold	Wastier than 10-fold. Expensivier than test set	Slightly better than test-set
R-fold	Identical to Leave-one-out	

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute
  - The total number of misclassifications on a testset.
  - E.g., the test set has 10 data points (4 data point -> positive, 6 data point -> negative)
  - Your classifier somehow predicted them all as positive …



What you should know

- Why you can’t use “training-set-error” to estimate the quality of your learning algorithm on your data, or to choose the learning algorithm
- Test-set cross-validation
- Leave-one-out cross-validation
- k-fold cross-validation