Cross Validation

Lesson

AI Academy

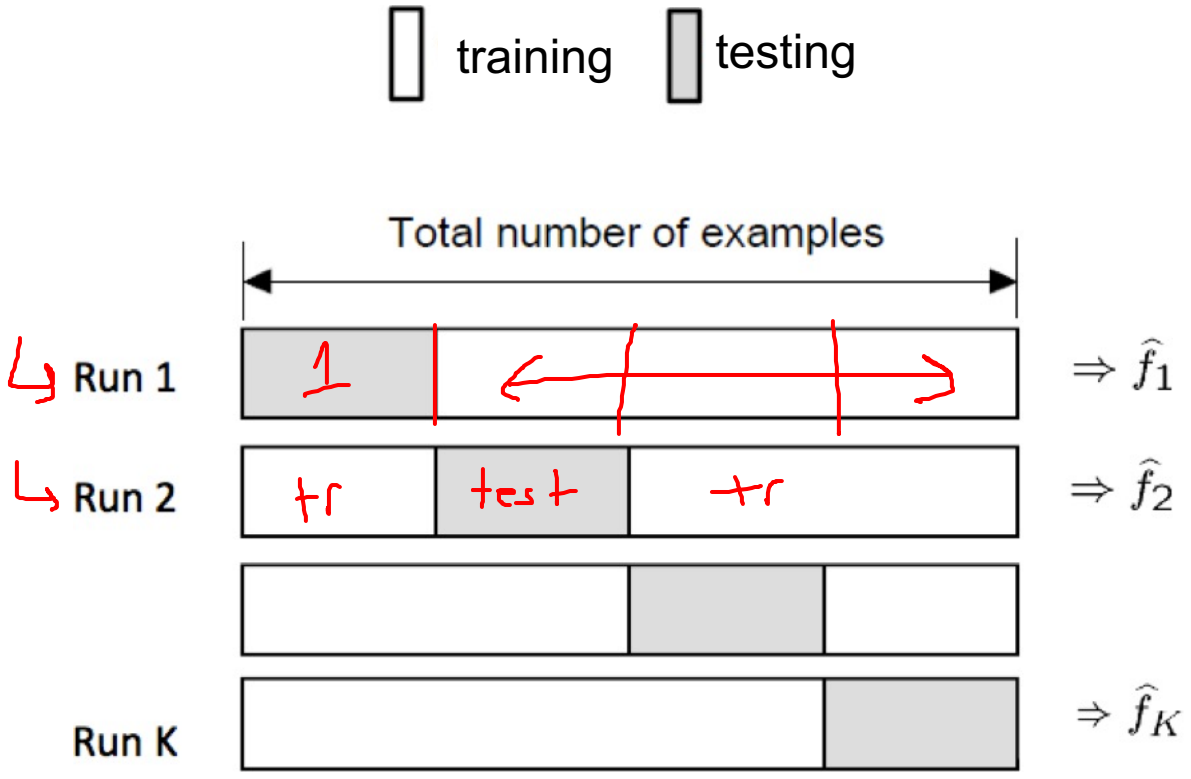
Holdout Method

- **Good news**
  - Very very simple.
  - We choose the method best test-set score.
- **Bad news**
  - **Wastes data**: Only 70% of data is used to train.
  - **Not Robust**: What if one model is just lucky on that test data?

Cross Validation (k-fold)

***k*-fold cross-validation (CV)**

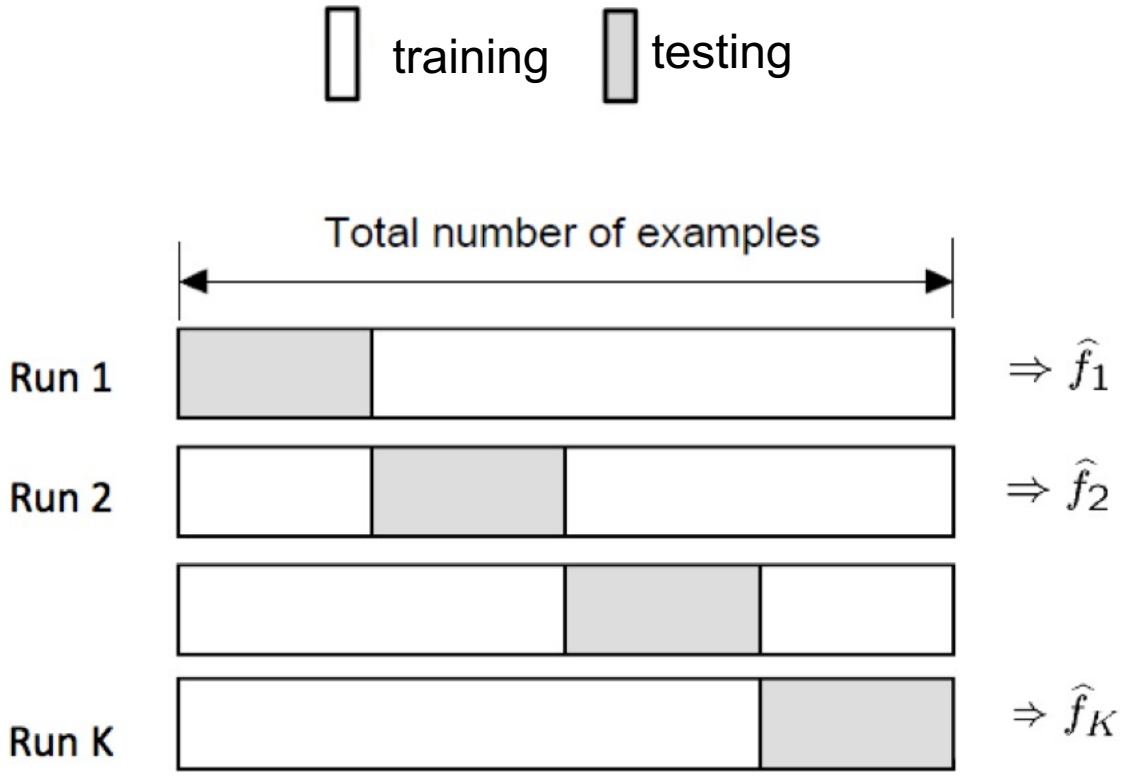
1. Create *k*-fold partition of the dataset.
2. For each fold t_i, train a model, f_i, on the *other* (k-1) folds.
3. Evaluate f_i using t_i as a test dataset. Average the results.



Cross Validation (LOOCV)

Leave one out CV (LOOCV)

1. Create *n*-fold partition of the dataset, i.e. *k* = *n*.
2. For each instance x_i, train a model, f_i, on the *other* (n-1) instances.
3. Evaluate f_i using $\{x_i\}$ as a test dataset. Average the results.



Cross Validation (LOOCV)

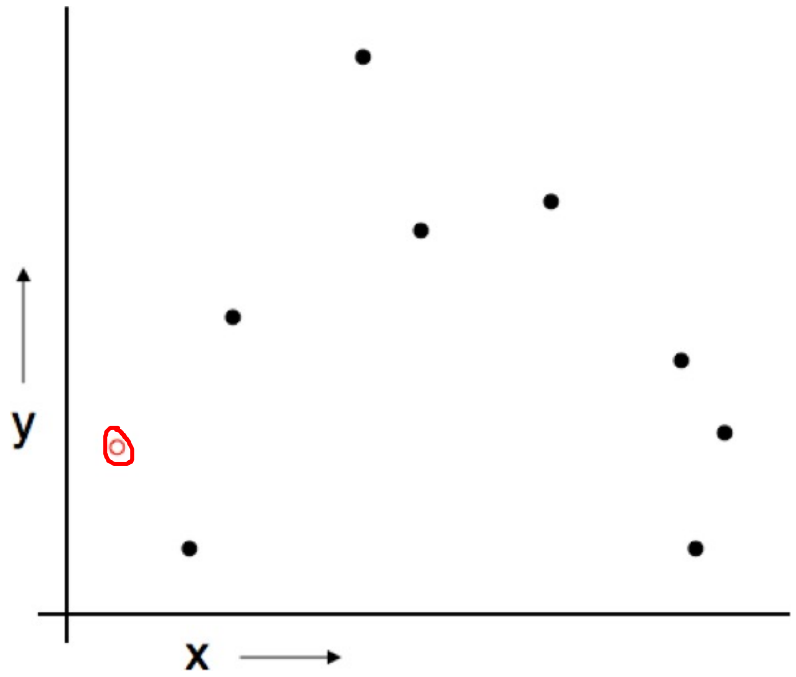
Advantages:

- Maximal use of training data.
- No sampling involved.
  - No randomness!

Disadvantages:

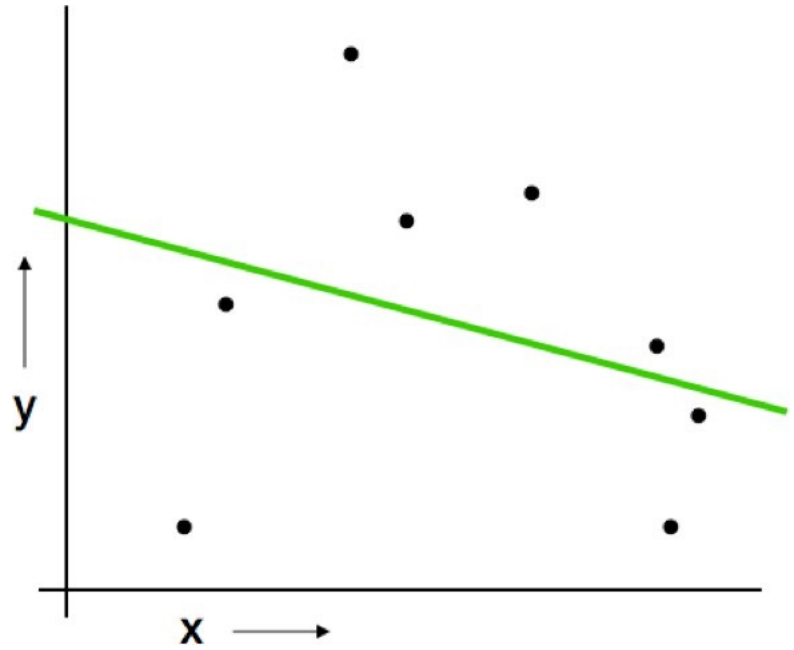
- Expensive on large datasets.
- Cannot be stratified (only one class in the test set).
- Some weird behavior.

LOOCV (Leave-one-out Cross Validation)



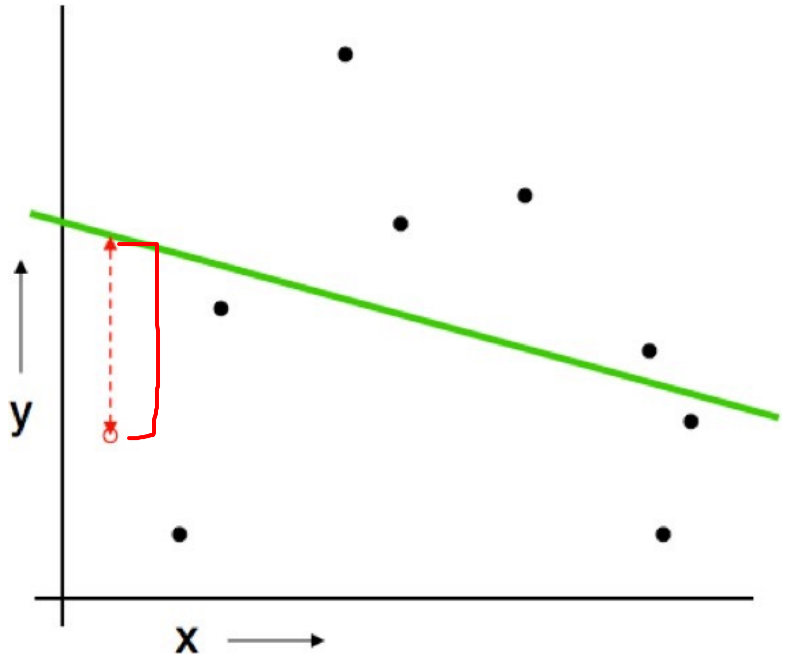
- For i=1 to N:
  1. Temporarily remove $\mathbf{d_i}$ from the dataset.

LOOCV (Leave-one-out Cross Validation)



- For i=1 to N:
  1. Temporarily remove $\mathbf{d_i}$ from the dataset.
  2. Train on the other N-1 data points: $\boldsymbol{f}_i(\boldsymbol{x})$.

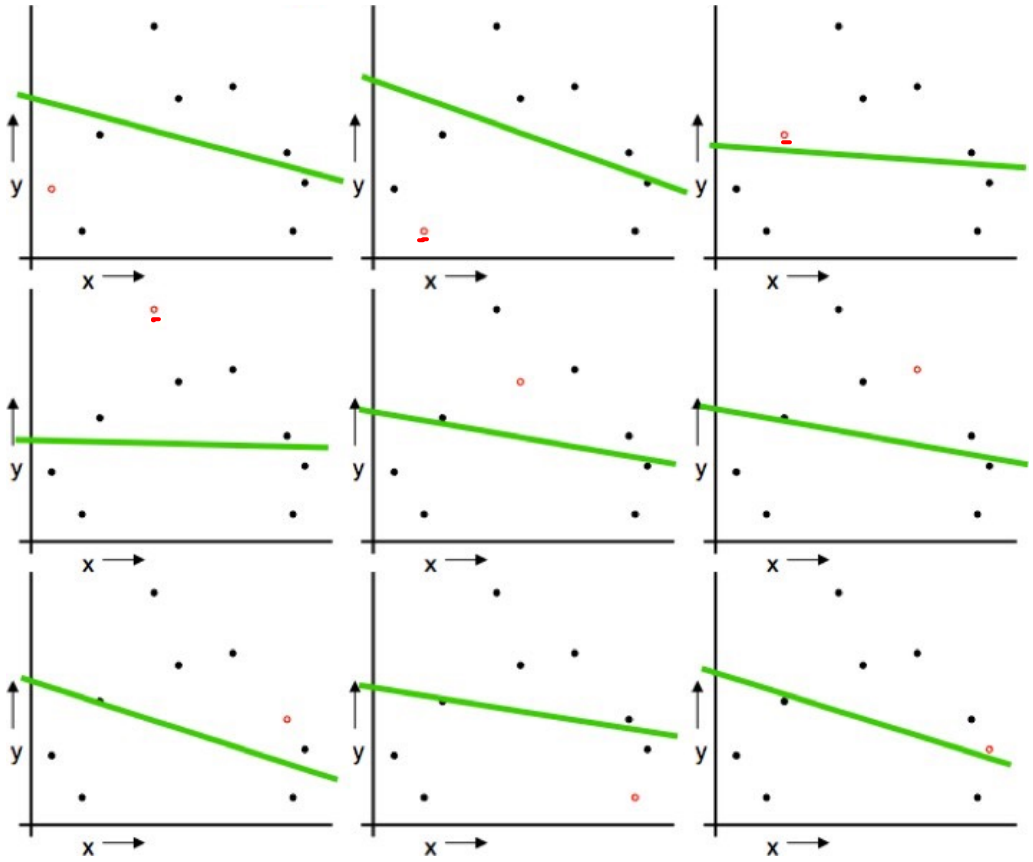
LOOCV (Leave-one-out Cross Validation)



- For i=1 to N:
  1. Temporarily remove $\mathbf{d_i}$ from the dataset.
  2. Train on the other N-1 data points: $f_k(x)$.
  3. Calculate error on $\mathbf{d_i}$.

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

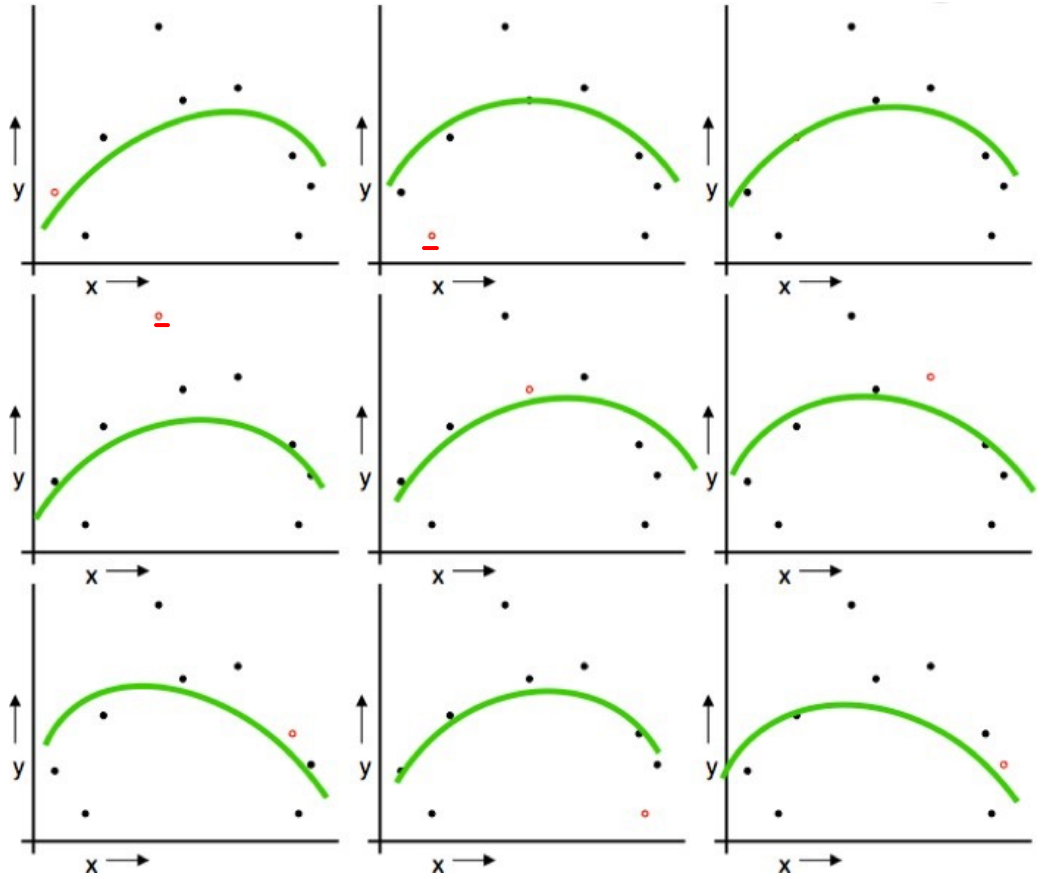
LOOCV for Linear Regression



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

$$MSE_{Loocv} = 2.12$$

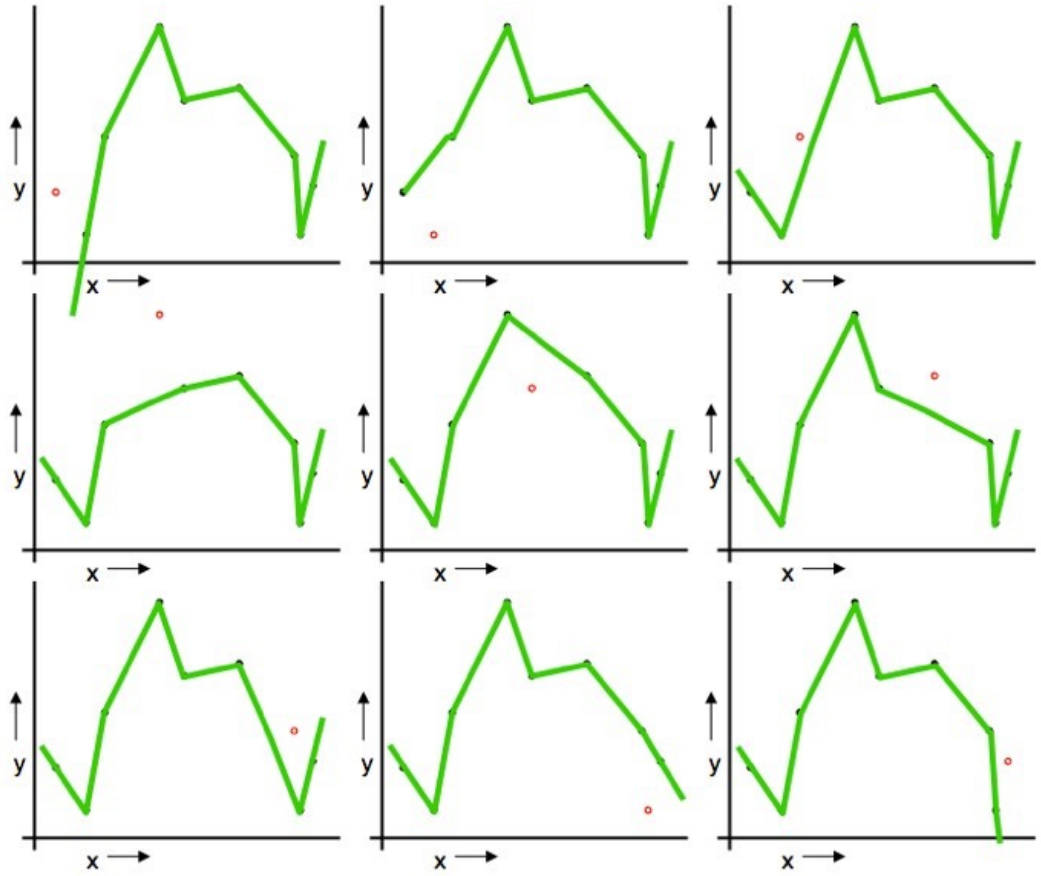
LOOCV for Quadratic Regression



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

$$MSE_{Loocv} = 0.962$$

LOOCV for Join the Dots



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

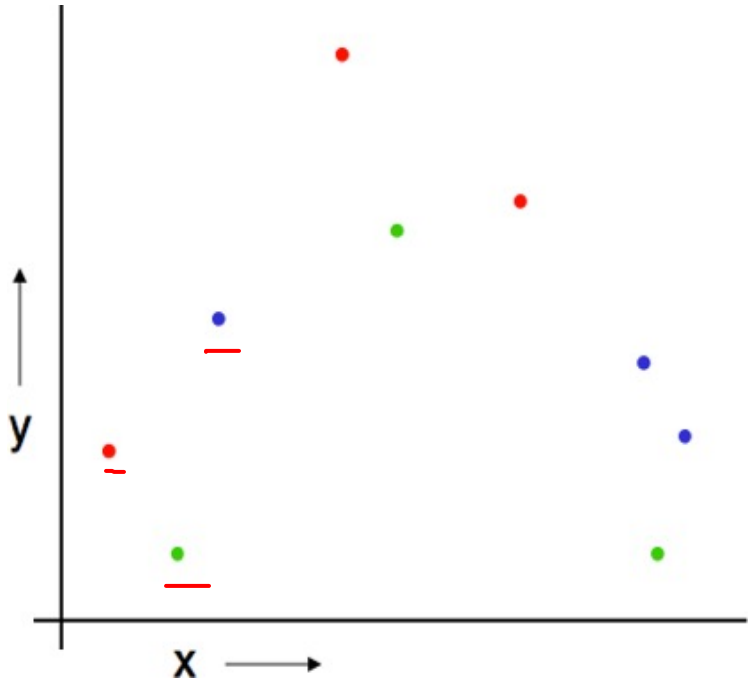
$$MSE_{Loocv} = 3.33$$

AI Academy

NC STATE

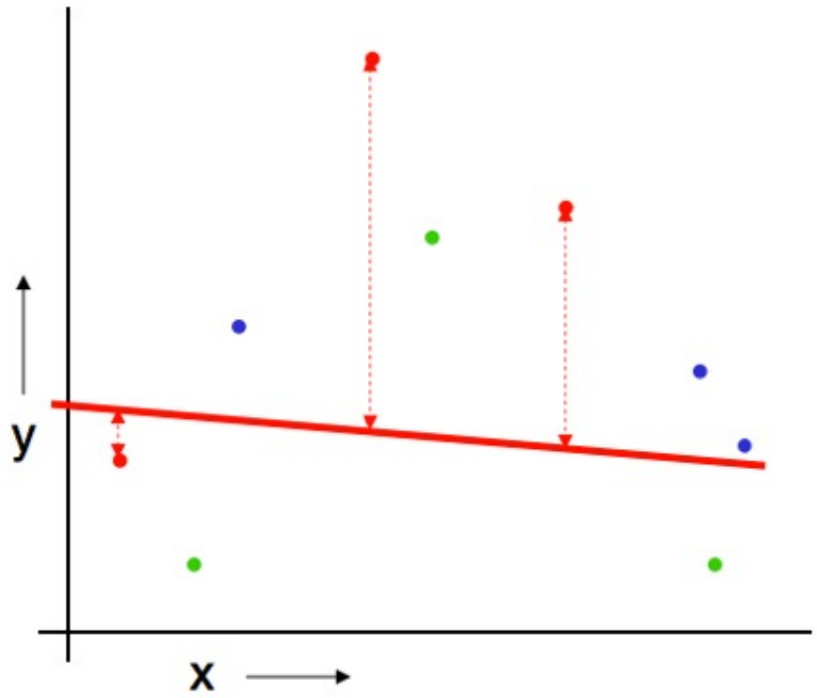
	Cons	Pros
Hold-Out	Variance: Unreliable estimate of future performance.	Quick and easy.
k-fold CV	Good balance between the two.	
LOOCV	Expensive.	No data wasted. More reliable.

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)

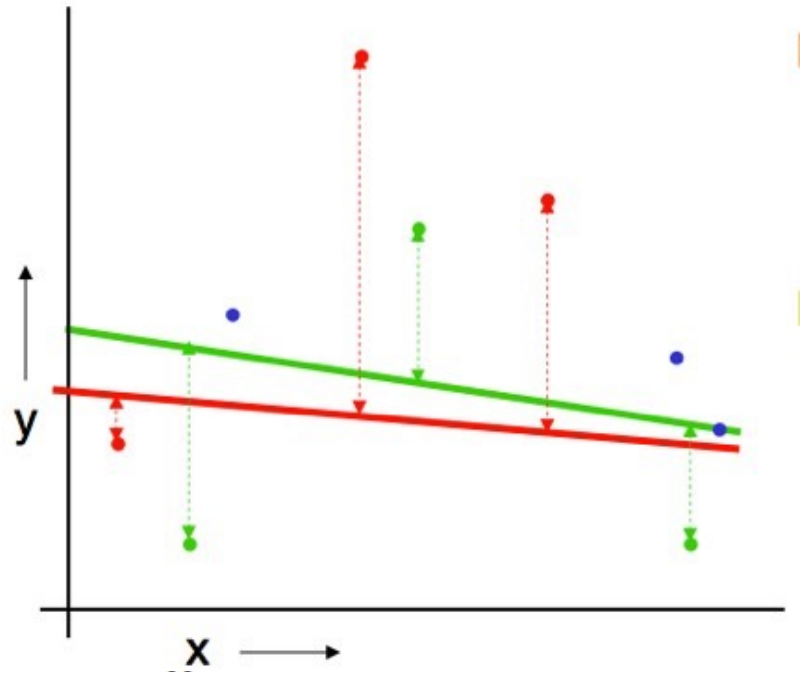
K-fold Cross Validation



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 points *not in the red.* Find the test- set sum of errors on the red points.

K-fold Cross Validation

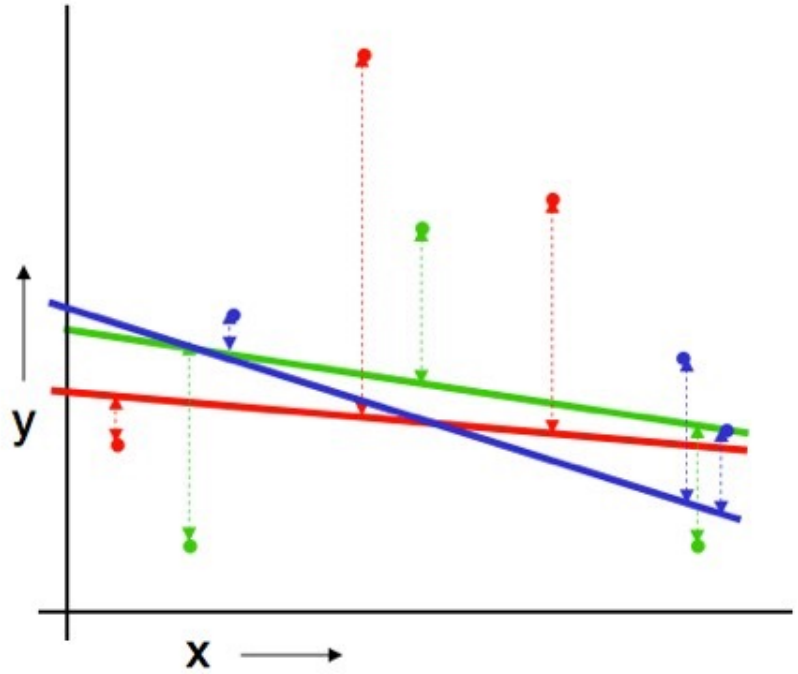


Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red, Green and Blue)

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

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

K-fold Cross Validation



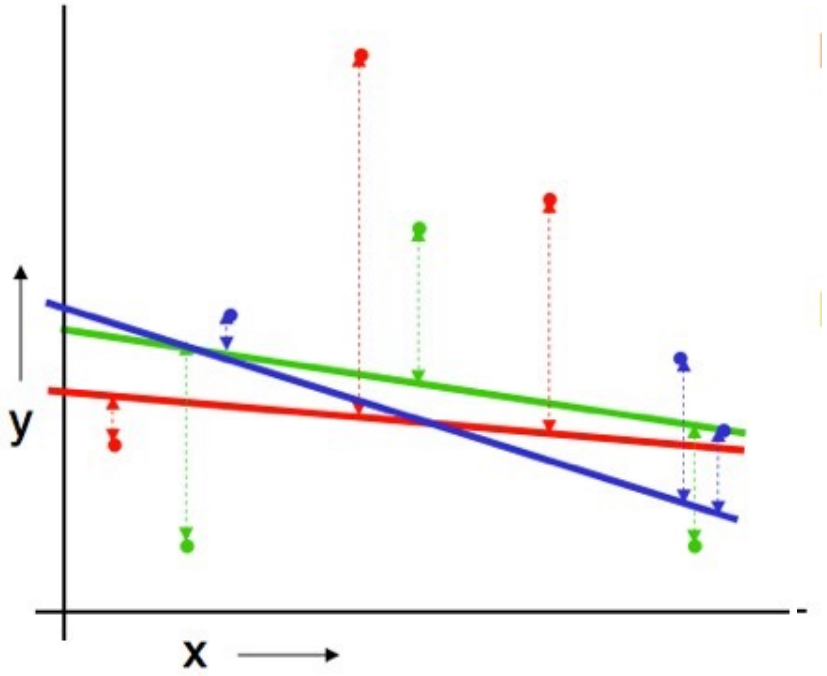
Randomly break the dataset into k partitions (in our example we'll have k=3 partitions colored Red, Green and Blue)

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

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

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

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

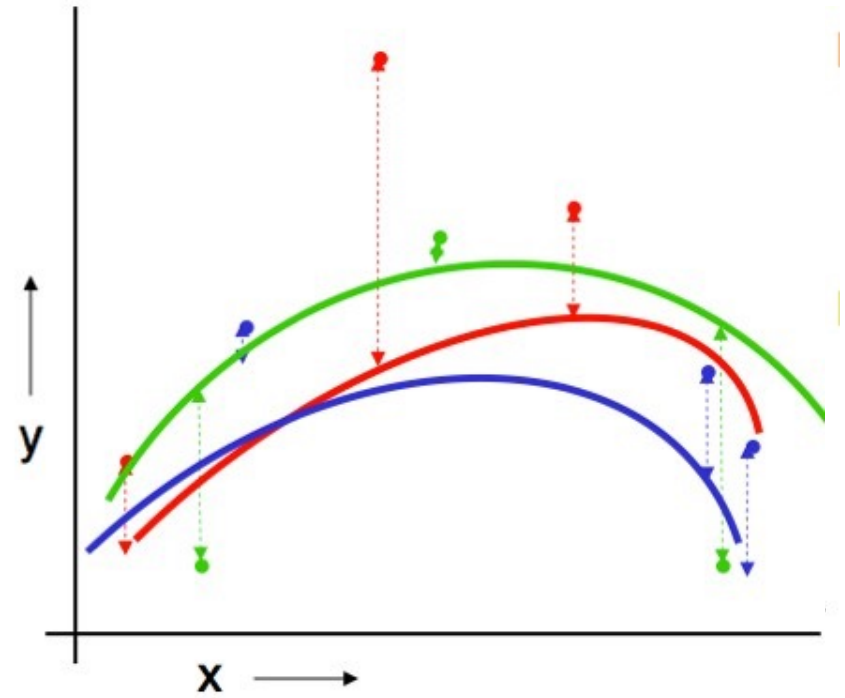
For the blue partition: Train on all points *not in the blue*. Find the test- set sum of errors on the blue points.

Then report the mean error

AI Academy

NC STATE

K-fold Cross Validation



Quadratic Regression $MSE_{3fold} = 1.1$

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 points *not in the red.* Find the test- set sum of errors on the red points.

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

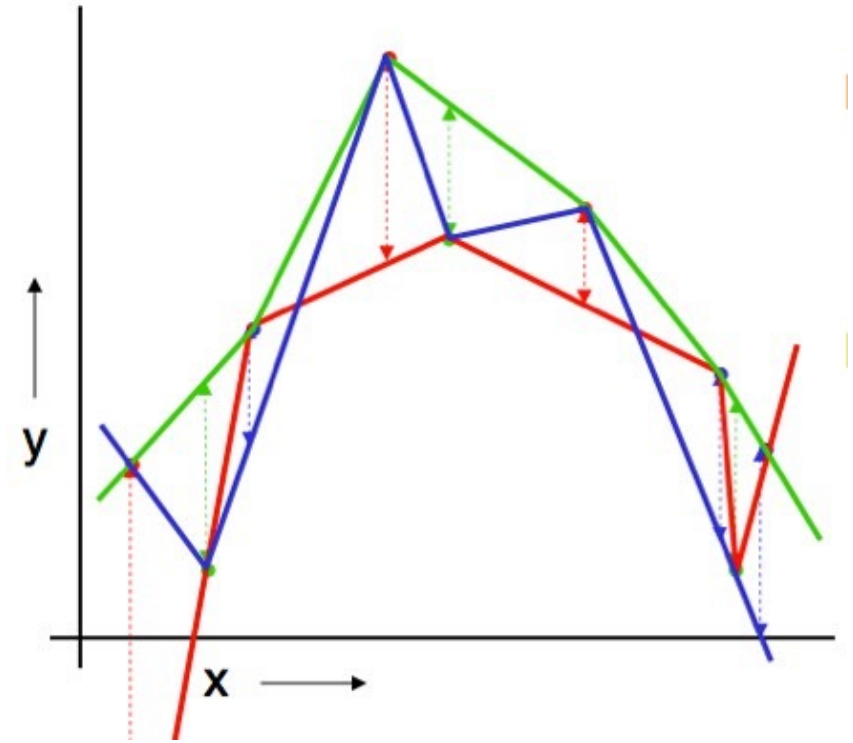
For the blue partition: Train on all points *not in the blue*. Find the test- set sum of errors on the blue points.

Then report the mean error

AI Academy

NC STATE

K-fold Cross Validation



Join the Dots $MSE_{3fold} = 2.93$

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 points *not in the red.* Find the test- set sum of errors on the red points.

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

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

Then report the mean error

How to Choose K?

- **Large K**
  - \+ Many training points (use more data).
  - \- Few testing points so variance of the error estimate would.
  - \- The computational time will be very large.
- **Small K**
  - \- Few training points.
  - \+ Larger testing points so variance of the error estimate will be small.
  - \+ The computational time will be reduced.

Learning Objectives: Cross Validation

You now should be able to:

- Evaluate a model using k-fold cross-validation and leave-one-out cross-validation.

AI Academy
NC STATE

AI Academy

Cross Validation

Exercises

AI Academy

Practice

Use **LOOCV** to evaluate a “decision stump” (1-level decision tree) on this dataset. What is the **testing error**?

Color	Class
Red	**Yes**
Red	*No*
Red	**Yes**
Blue	**Yes**
Blue	**Yes**
Red	*No*

Practice

Use **LOOCV** to evaluate a “decision stump” (1-level decision tree) on this dataset. What is the **testing error**?

Color	Class
Red	**Yes**
Red	*No*
Red	**Yes**
Blue	**Yes**
Blue	**Yes**
Red	*No*

2/6

4/6

.67

Practice

Use **3-fold CV** to evaluate a “decision stump” (1-level decision tree) on this dataset. What is the **testing error**?

Color	Class
Red	**Yes**
Red	*No*
Red	*No*
Blue	**Yes**
Blue	**Yes**
Red	*No*