

Cross-validation for detecting and preventing overfitting

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials>. Comments and corrections gratefully received.

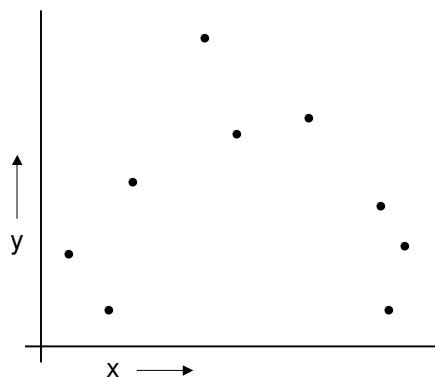
Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Copyright © Andrew W. Moore

Slide 1

A Regression Problem



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

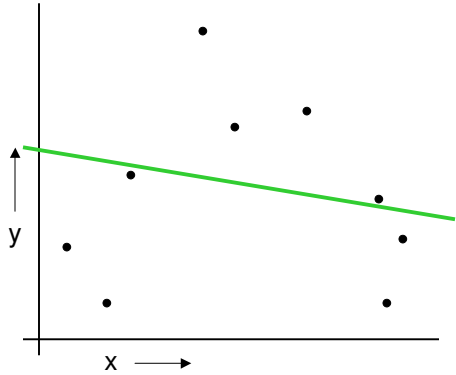
Can we learn f from this data?

Let's consider three methods...

Copyright © Andrew W. Moore

Slide 2

Linear Regression



Copyright © Andrew W. Moore

Slide 3

Linear Regression

Univariate Linear regression with a constant term:

X	Y
3	7
1	3
⋮	⋮



$$\mathbf{x} = \begin{bmatrix} 3 \\ 1 \\ \vdots \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 7 \\ 3 \\ \vdots \end{bmatrix}$$

$x_1 = (3) \dots$ $y_1 = 7 \dots$

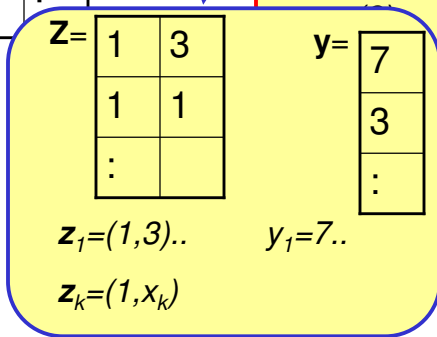
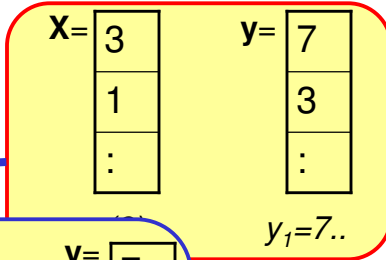
Copyright © Andrew W. Moore

Slide 4

Linear Regression

Univariate Linear regression with a constant term:

X	Y
3	7
1	3
⋮	⋮



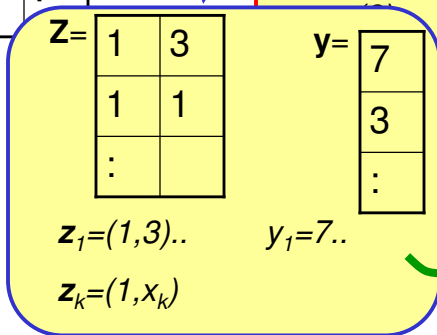
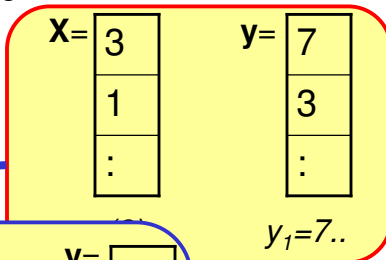
Copyright © Andrew W. Moore

Slide 5

Linear Regression

Univariate Linear regression with a constant term:

X	Y
3	7
1	3
⋮	⋮



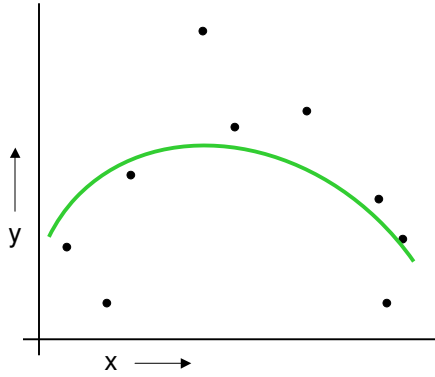
$$\beta = (Z^T Z)^{-1} (Z^T y)$$

$$y^{est} = \beta_0 + \beta_1 x$$

Copyright © Andrew W. Moore

Slide 6

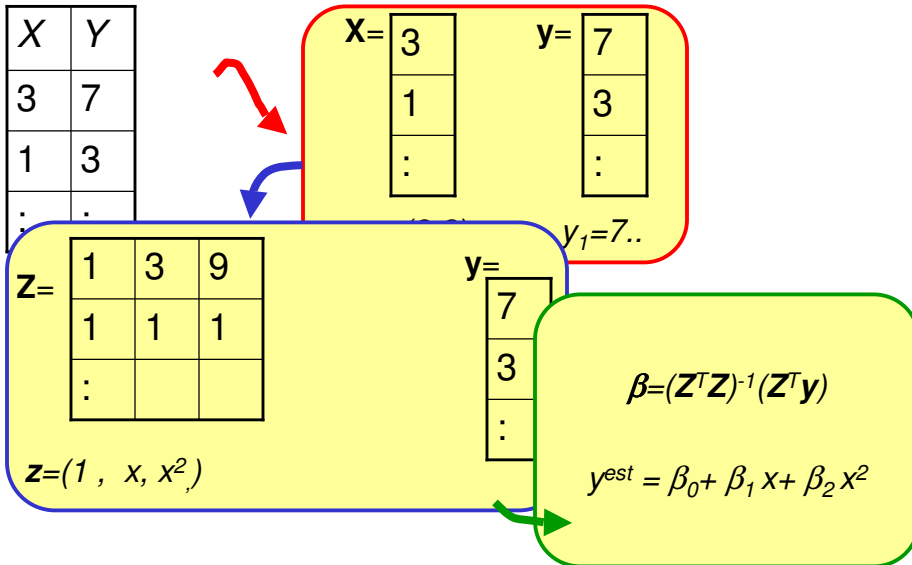
Quadratic Regression



Copyright © Andrew W. Moore

Slide 7

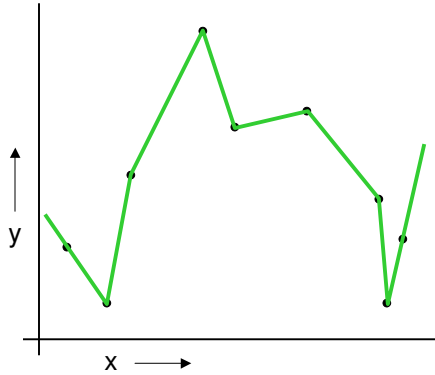
Quadratic Regression



Copyright © Andrew W. Moore

Slide 8

Join-the-dots

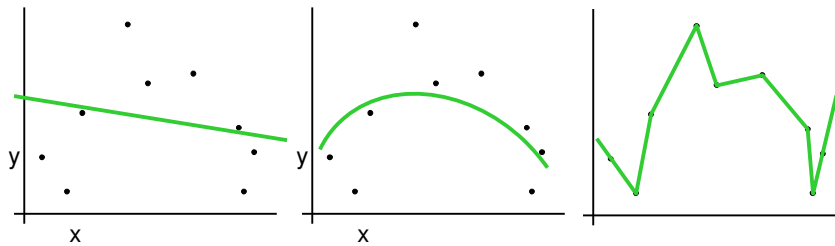


Also known as **piecewise linear nonparametric regression** if that makes you feel better

Copyright © Andrew W. Moore

Slide 9

Which is best?

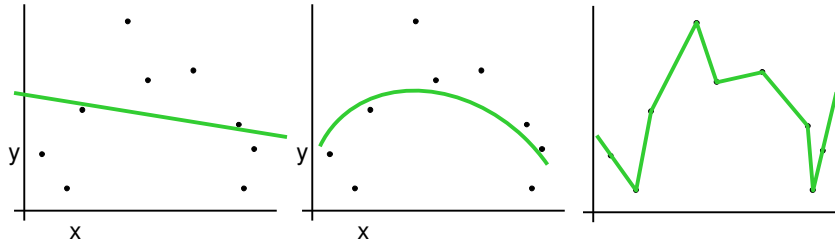


Why not choose the method with the best fit to the data?

Copyright © Andrew W. Moore

Slide 10

What do we really want?



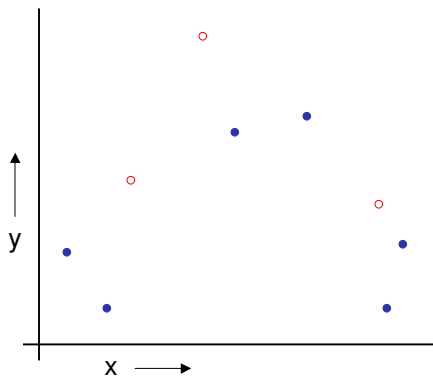
Why not choose the method with the best fit to the data?

“How well are you going to predict future data drawn from the same distribution?”

Copyright © Andrew W. Moore

Slide 11

The test set method



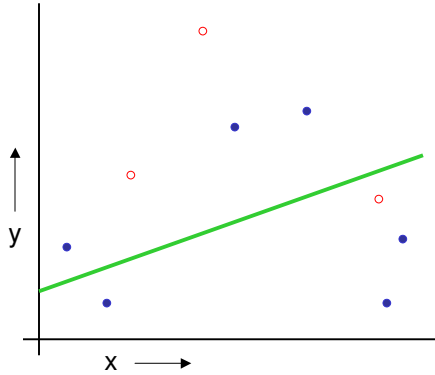
1. Randomly choose 30% of the data to be in a **test set**

2. The remainder is a **training set**

Copyright © Andrew W. Moore

Slide 12

The test set method



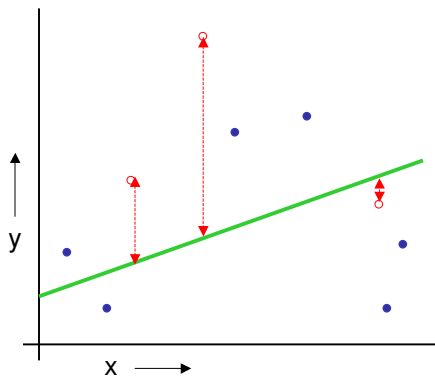
1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set

(Linear regression example)

Copyright © Andrew W. Moore

Slide 13

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)

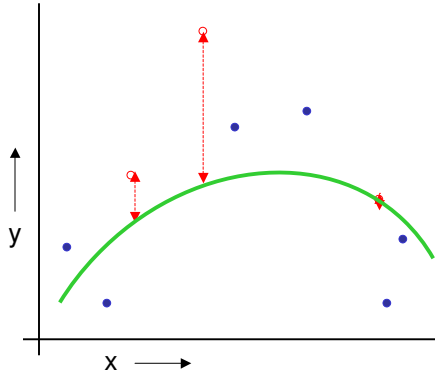
Mean Squared Error = 2.4

4. Estimate your future performance with the **test set**

Copyright © Andrew W. Moore

Slide 14

The test set method



(Quadratic regression example)

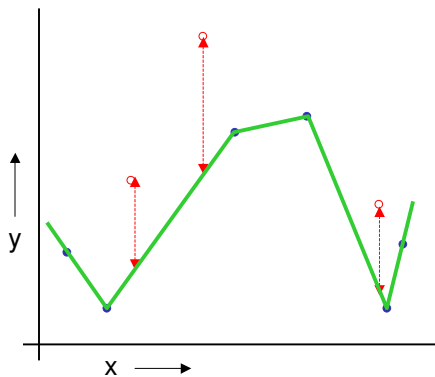
Mean Squared Error = 0.9

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

Copyright © Andrew W. Moore

Slide 15

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

Copyright © Andrew W. Moore

Slide 16

The test set method

Good news:

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

Bad news:

- What's the downside?

Copyright © Andrew W. Moore

Slide 17

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

We say the "test-set estimator of performance has high variance"

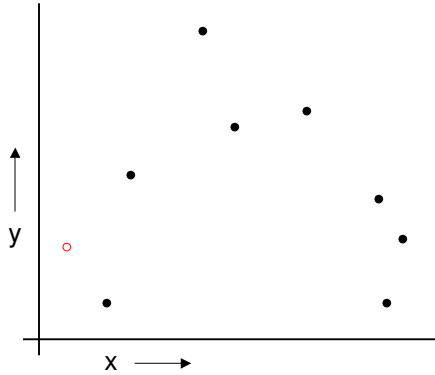
Copyright © Andrew W. Moore

Slide 18

LOOCV (Leave-one-out Cross Validation)

For $k=1$ to R

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



Copyright © Andrew W. Moore

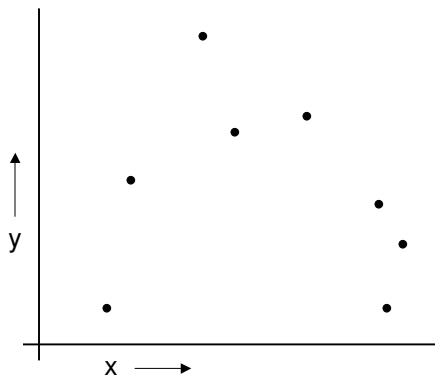
Slide 19

LOOCV (Leave-one-out Cross Validation)

For $k=1$ to R

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

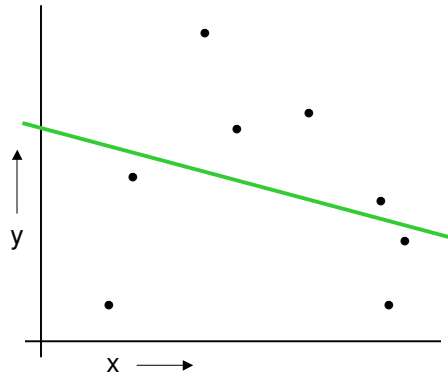
2. Temporarily remove (x_k, y_k) from the dataset



Copyright © Andrew W. Moore

Slide 20

LOOCV (Leave-one-out Cross Validation)



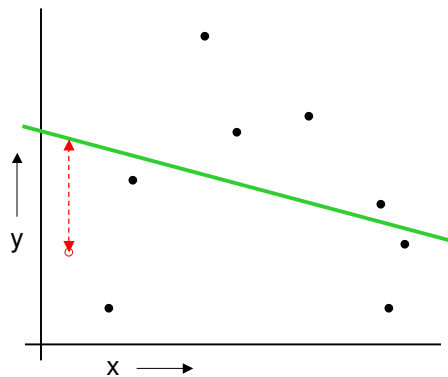
For $k=1$ to R

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

Copyright © Andrew W. Moore

Slide 21

LOOCV (Leave-one-out Cross Validation)



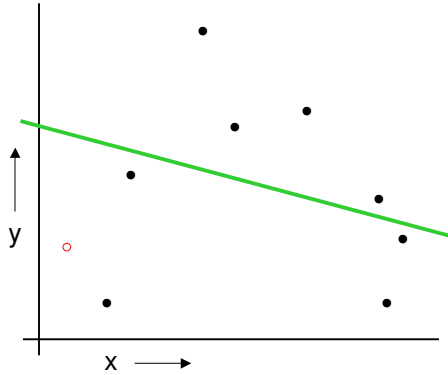
For $k=1$ to R

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)

Copyright © Andrew W. Moore

Slide 22

LOOCV (Leave-one-out Cross Validation)

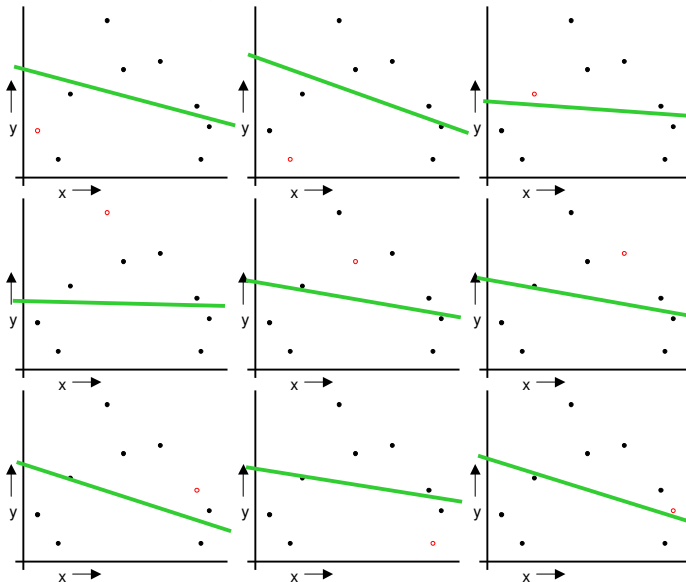


For $k=1$ to R

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.

LOOCV (Leave-one-out Cross Validation)



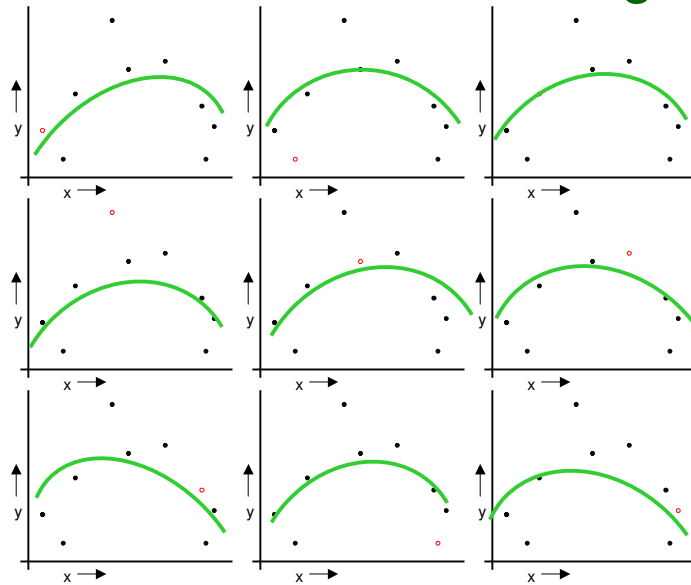
For $k=1$ to R

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)

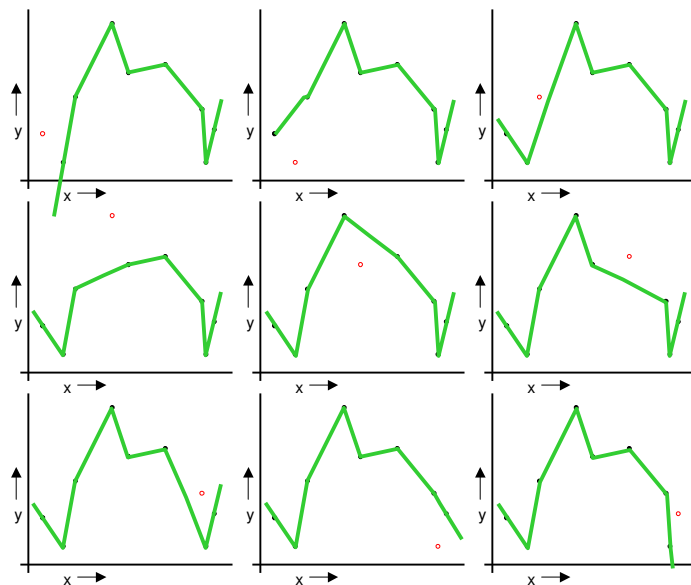
When you've done all points, report the mean error.

$$MSE_{LOOCV} = 0.962$$

Copyright © Andrew W. Moore

Slide 25

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.

$$MSE_{LOOCV} = 3.33$$

Copyright © Andrew W. Moore

Slide 26

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

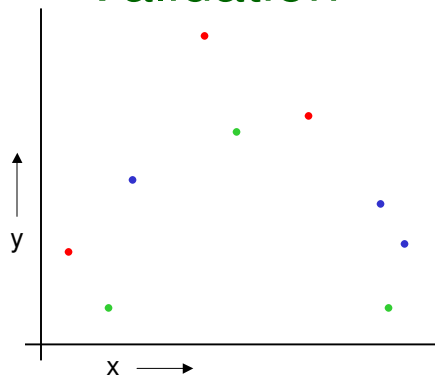
..can we get the best of both worlds?

Copyright © Andrew W. Moore

Slide 27

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)

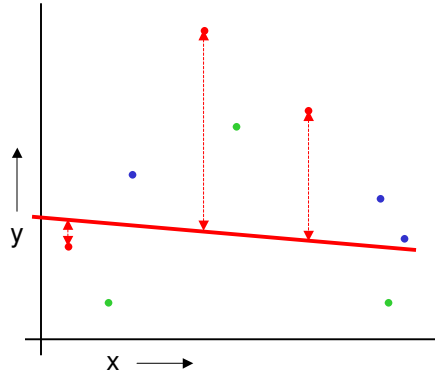


Copyright © Andrew W. Moore

Slide 28

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)



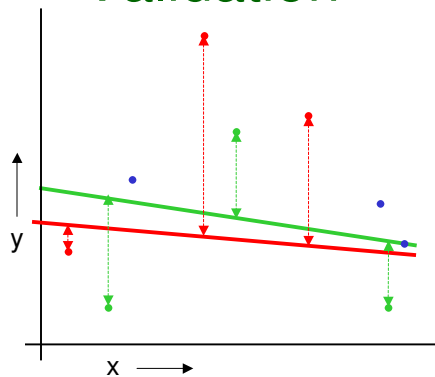
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

Copyright © Andrew W. Moore

Slide 29

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.

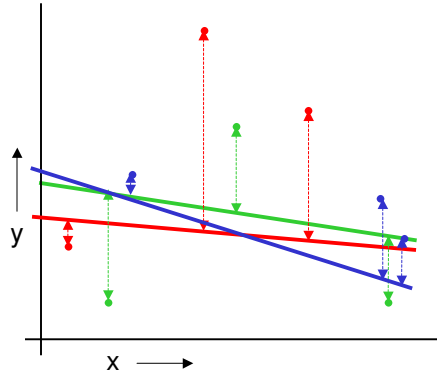
For the green partition: Train on all the points not in the green partition. Find the test-set sum of errors on the green points.

Copyright © Andrew W. Moore

Slide 30

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)



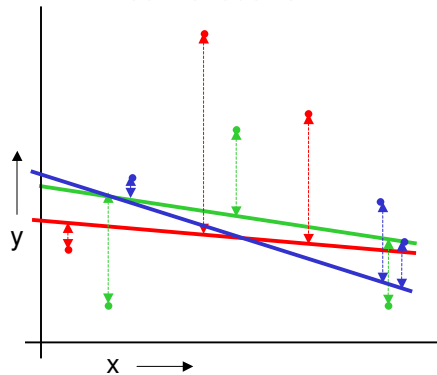
For the red partition: Train on all the points not in the red partition. Find the test-set sum of errors on the red points.

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

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.

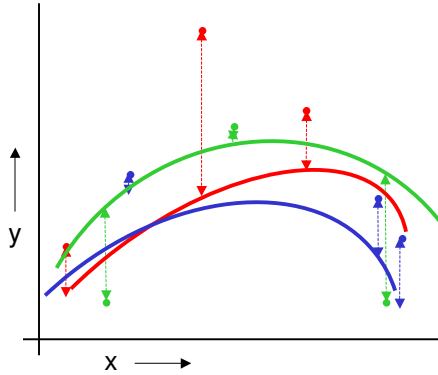
For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Linear Regression
 $MSE_{3FOLD}=2.05$

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.

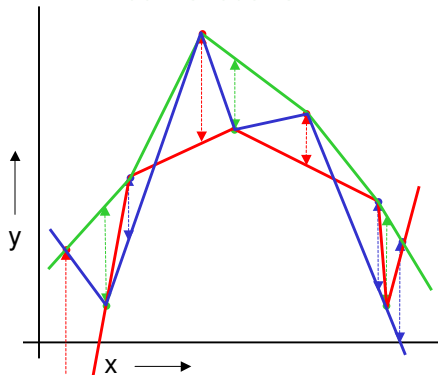
For the blue partition: Train on all the points not in the blue partition. Find the test-set sum of errors on the blue points.

Then report the mean error

Quadratic Regression
 $MSE_{3FOLD}=1.11$

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

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
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
N-fold	Identical to Leave-one-out	

Copyright © Andrew W. Moore

Slide 35

CV-based Model Selection

- We're trying to decide which algorithm to use.
- We train each machine and make a table...

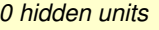



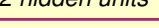







i	f_i	TRAINERR	10-FOLD-CV-ERR	Choice
1	f_1			
2	f_2			
3	f_3			⊗
4	f_4			
5	f_5			
6	f_6			

Copyright © Andrew W. Moore

Slide 36

CV-based Model Selection

- Example: Choosing number of hidden units in a one-hidden-layer neural net.
- Step 1: Compute 10-fold CV error for six different model classes:

Algorithm	TRAINERR	10-FOLD-CV-ERR	Choice
0 hidden units			
1 hidden units			
2 hidden units			<input checked="" type="checkbox"/>
3 hidden units			
4 hidden units			
5 hidden units			












- Step 2: Whichever model class gave best CV score: train it with all the data, and that's the predictive model you'll use.

Copyright © Andrew W. Moore

Slide 37

CV-based Model Selection

- Example: Choosing "k" for a k-nearest-neighbor regression.
- Step 1: Compute LOOCV error for six different model classes:

Algorithm	TRAINERR	10-fold-CV-ERR	Choice
K=1			
K=2			
K=3			
K=4			<input checked="" type="checkbox"/>
K=5			
K=6			

- Step 2: Whichever model class gave best CV score: train it with all the data, and that's the predictive model you'll use.

Copyright © Andrew W. Moore

Slide 38

CV-based Model Selection

- Example: Choosing “k” for a k-nearest-neighbor regression.
- Step 1: Compute LOOCV error for six different classes:

Algorithm	Training Error	Test Error
K=1	██████████	██████████
K=2	██████████	██████████
K=3	██████████	██████████
K=4	██████████	██████████
K=5	██████████	██████████
K=6	██████████	██████████

Why did we use 10-fold-CV for neural nets and LOOCV for k-nearest neighbor?

And why stop at K=6

Are we guaranteed that a local optimum of K vs LOOCV will be the global optimum?

What should we do if we are depressed at the expense of doing LOOCV for K= 1 through 1000?

The reason is Computational. For k-NN (and all other nonparametric methods) LOOCV happens to be as cheap as regular predictions.

No good reason, except it looked like things were getting worse as K was increasing

Sadly, no. And in fact, the relationship can be very bumpy.

Idea One: K=1, K=2, K=4, K=8, K=16, K=32, K=64 ... K=1024

Idea Two: Hillclimbing from an initial guess at K

- Step 2: Whichever model class gave best CV score: train it with all the data, and that’s the predictive model you’ll use.

CV-based Model Selection

- Can you think of other decisions we can ask Cross Validation to make for us, based on other machine learning algorithms in the class so far?

CV-based Model Selection

- Can you think of other decisions we can ask Cross Validation to make for us, based on other machine learning algorithms in the class so far?
 - Degree of polynomial in polynomial regression
 - Whether to use full, diagonal or spherical Gaussians in a Gaussian Bayes Classifier.
 - The Kernel Width in Kernel Regression
 - The Kernel Width in Locally Weighted Regression
 - The Bayesian Prior in Bayesian Regression

These involve choosing the value of a real-valued parameter. What should we do?

Copyright © Andrew W. Moore

Slide 41

CV-based Model Selection

- Can you think of other decisions we can ask Cross Validation to make for us, based on other machine learning algorithms in the class so far?
 - Degree of polynomial in polynomial regression
 - Whether to use full, diagonal or spherical Gaussians in a Gaussian Bayes Classifier.
 - The Kernel Width in Kernel Regression
 - The Kernel Width in Locally Weighted Regression
 - The Bayesian Prior in Bayesian Regression

These involve choosing the value of a real-valued parameter. What should we do?

Idea One: Consider a discrete set of values (often best to consider a set of values with exponentially increasing gaps, as in the K-NN example).

Idea Two: Compute $\frac{\partial \text{LOOCV}}{\partial \text{Parameter}}$ and then do gradient descent.

Copyright © Andrew W. Moore

Slide 42

CV-based Model Selection

- Can you think of other decisions we can ask Cross Validation to make for us, based on other machine learning algorithms in the class so far?
 - Degree of polynomial in polynomial regression
 - Whether to use full, diagonal or spherical Gaussians in a Gaussian Bayes Classifier.
 - The Kernel Width in Kernel Regression
 - The Kernel Width in Locally Weighted Regression
 - The Bayesian Prior in Bayesian Regression

These involve choosing the value of a real-valued parameter. What should we do?

Idea One: Consider a discrete set of values (often best to consider a set of values with exponentially increasing gaps, as in the K-NN example).

Idea Two: Compute $\frac{\partial \text{LOOCV}}{\partial \text{Parameter}}$ and then do gradient descent.

Copyright © Andrew W. Moore

Slide 43

CV-based Algorithm Choice

- Example: Choosing which regression algorithm to use
- Step 1: Compute 10-fold-CV error for six different model classes:

Algorithm	TRAINERR	10-fold-CV-ERR	Choice
1-NN			
10-NN			
Linear Reg'n			
Quad reg'n			☒
LWR, KW=0.1			
LWR, KW=0.5			

- Step 2: Whichever algorithm gave best CV score: train it with all the data, and that's the predictive model you'll use.

Copyright © Andrew W. Moore

Slide 44

Cross-validation for classification

- Instead of computing the sum squared errors on a test set, you should compute...

Copyright © Andrew W. Moore

Slide 45

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.

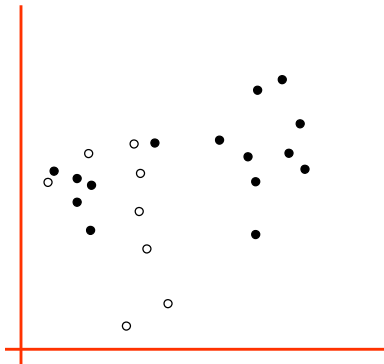
Copyright © Andrew W. Moore

Slide 46

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.



- What's LOOCV of 1-NN?
- What's LOOCV of 3-NN?
- What's LOOCV of 22-NN?

Copyright © Andrew W. Moore

Slide 47

Cross-Validation for classification

- Choosing k for k -nearest neighbors
- Choosing h for the Parzen windows
- Any other “free” parameter of a classifier
- Choosing which classifier to use
- Choosing Features to use

Copyright © Andrew W. Moore

Slide 48

Feature Selection

- Suppose you have a learning algorithm LA and a set of input attributes $\{X_1, X_2 \dots X_m\}$
- You expect that LA will only find some subset of the attributes useful.
- Question: How can we use cross-validation to find a useful subset?
- Four ideas:
 - Forward selection
 - Backward elimination
 - Hill Climbing
 - Stochastic search (Simulated Annealing or GAs)

Another fun area in which Andrew has spent a lot of his wild youth