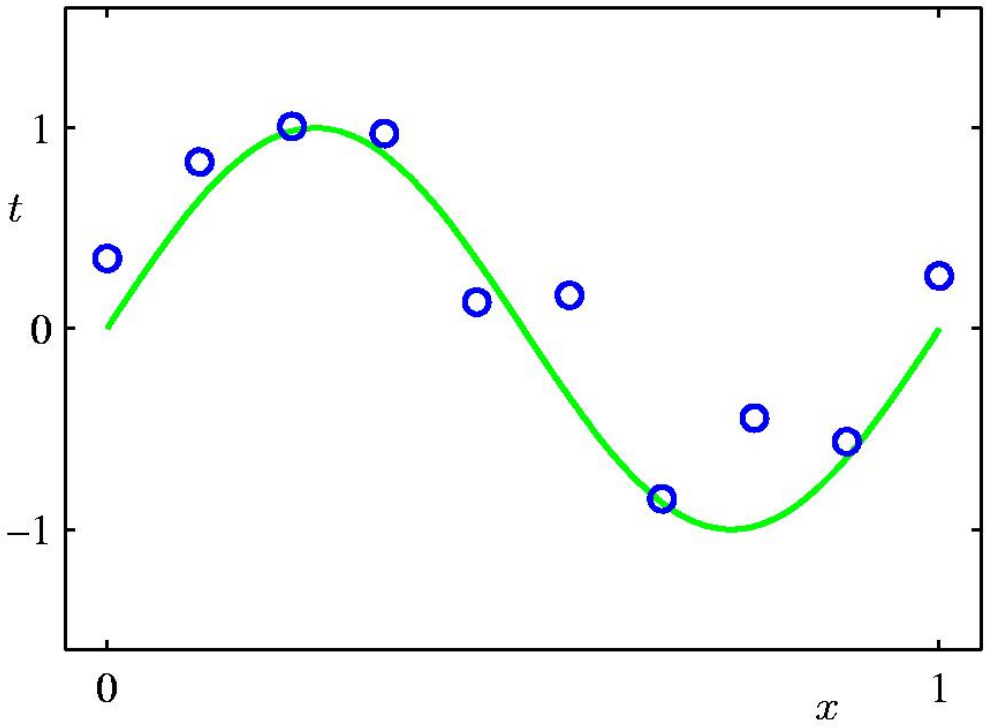
6. Model Selection

Kai Yu

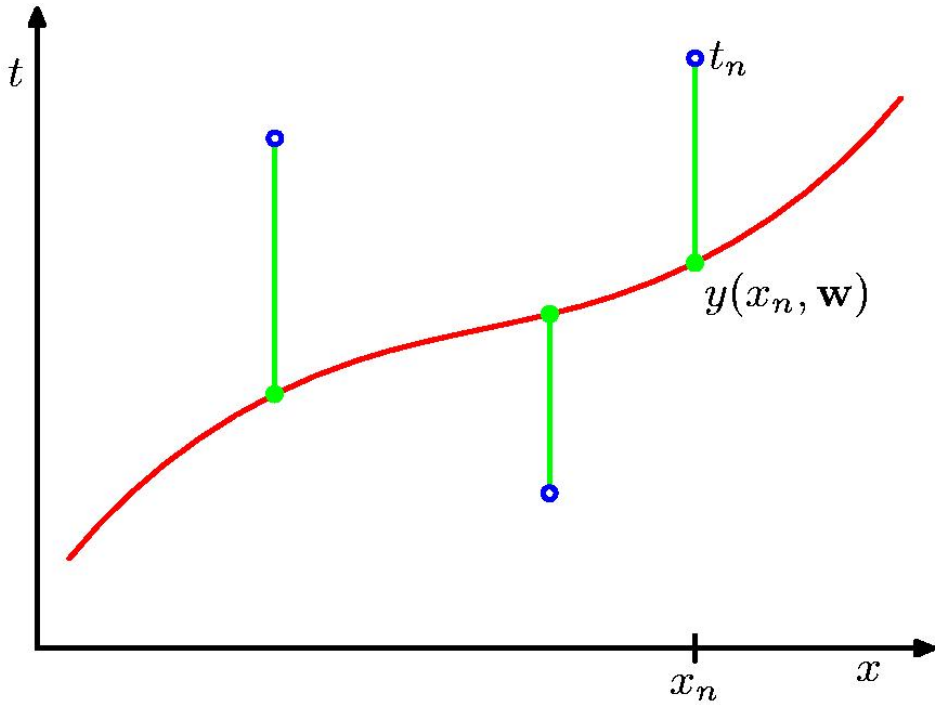
Polynomial Curve Fitting

Blue: Observed data
Green: true distribution



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

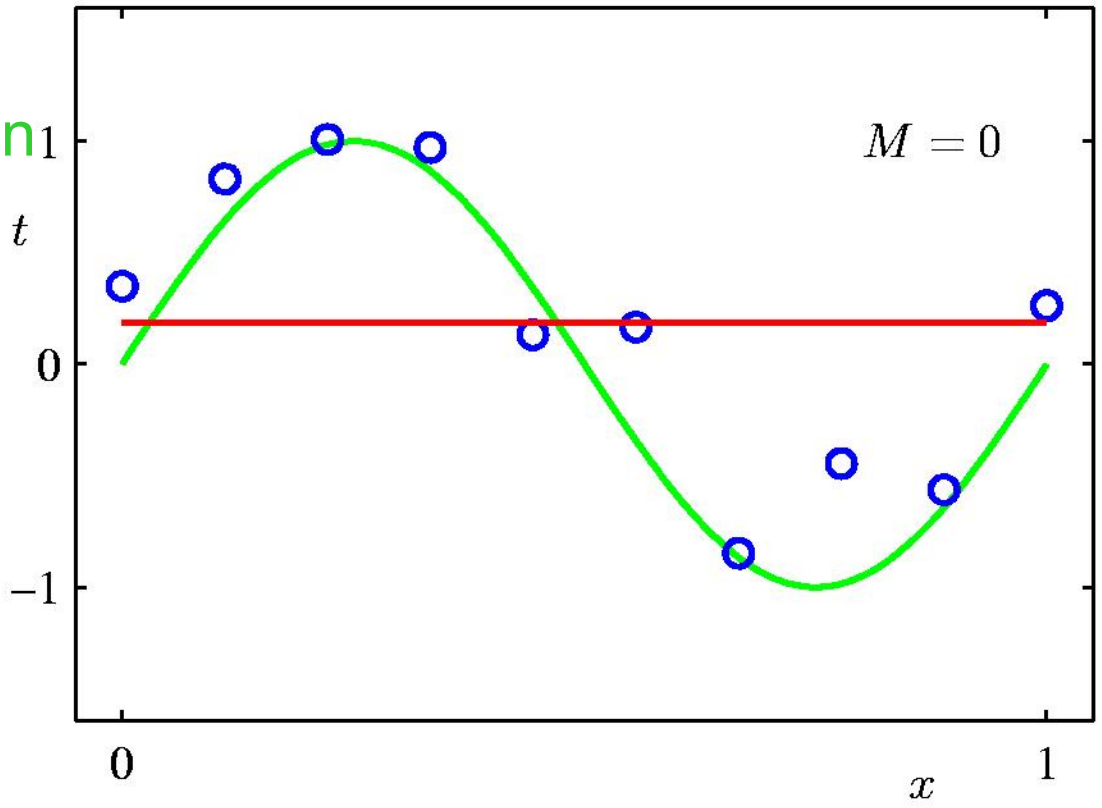
Sum-of-Squares Error Function



$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

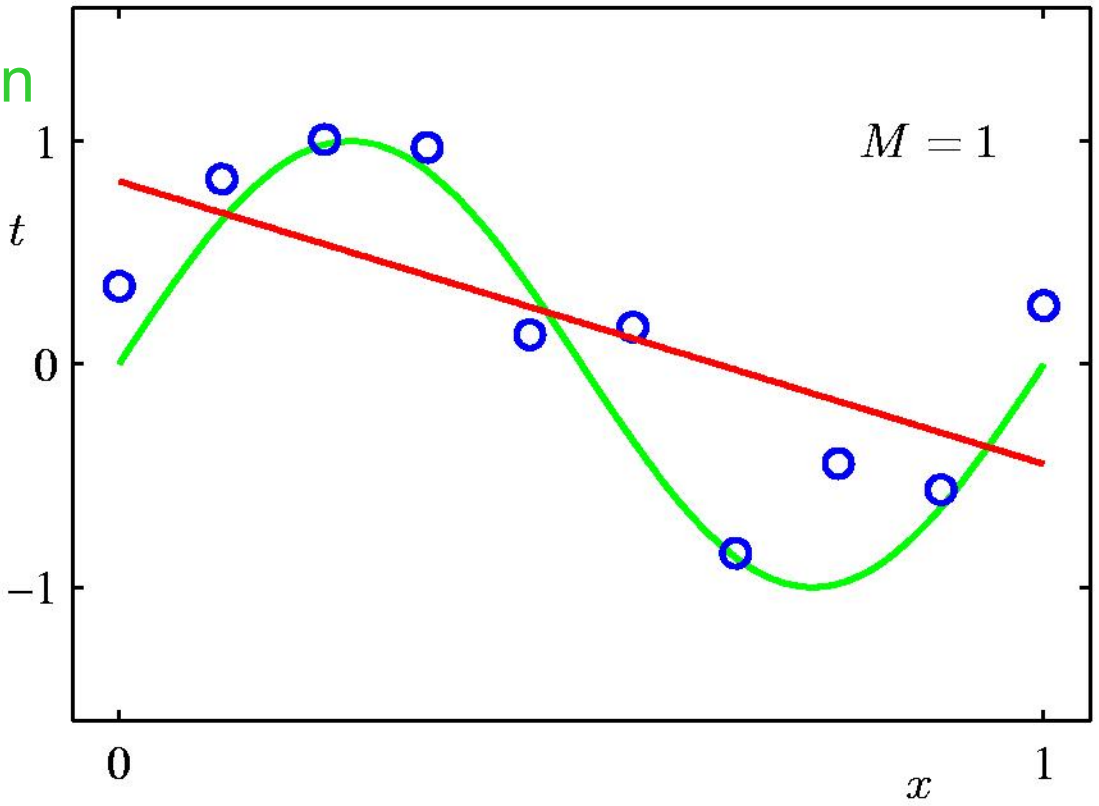
0^{th} Order Polynomial

Blue: Observed data
Red: Predicted curve
Green true distribution



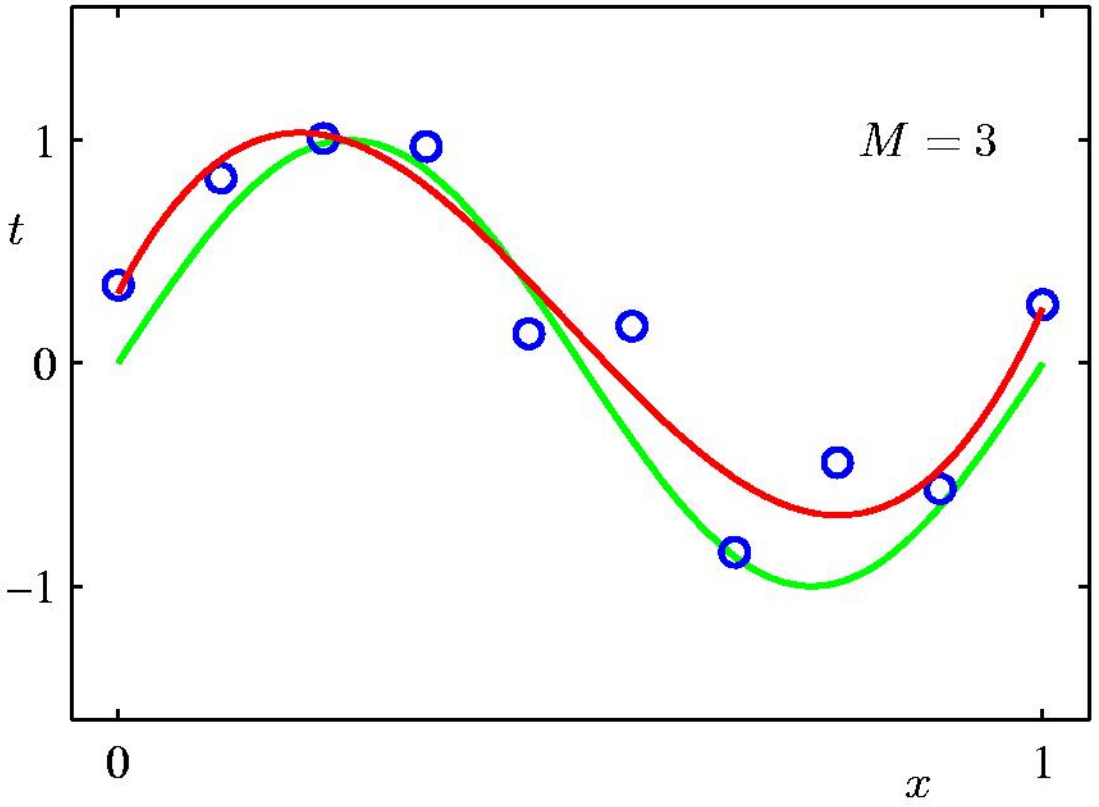
I st Order Polynomial

Blue: Observed data
Red: Predicted curve
Green: true distribution



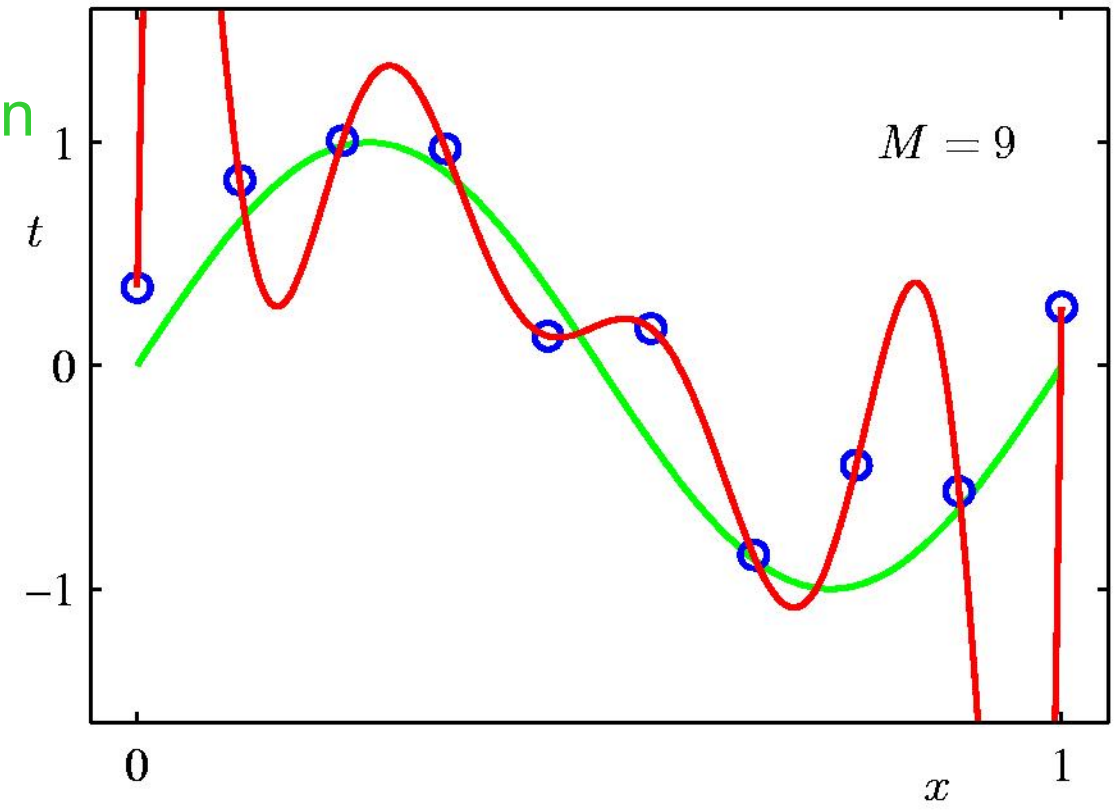
3rd Order Polynomial

Blue: Observed data
Red: Predicted curve
Green true distribution

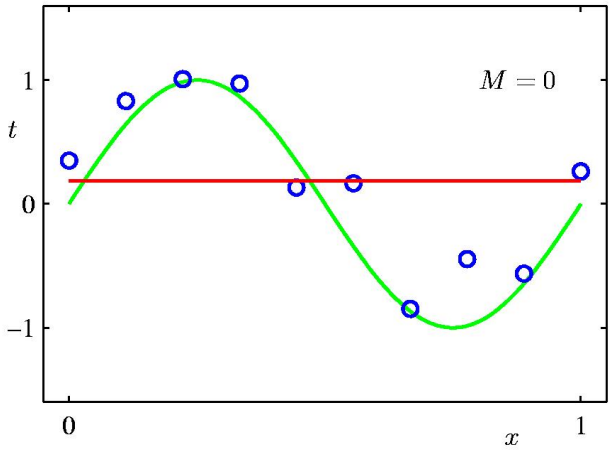


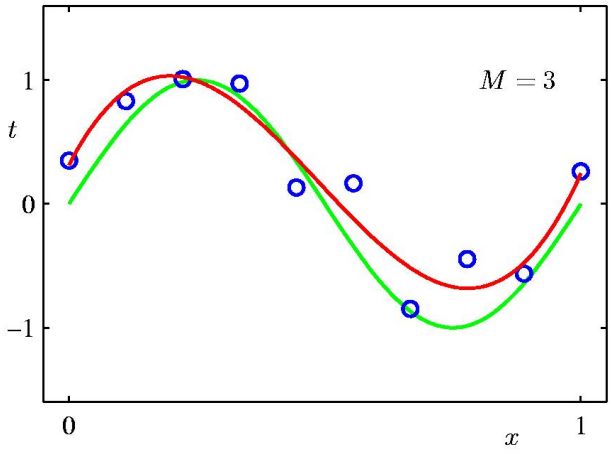
9th Order Polynomial

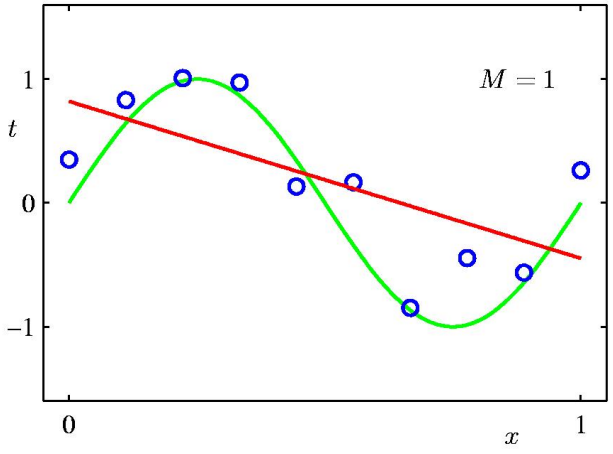
Blue: Observed data
Red: Predicted curve
Green true distribution

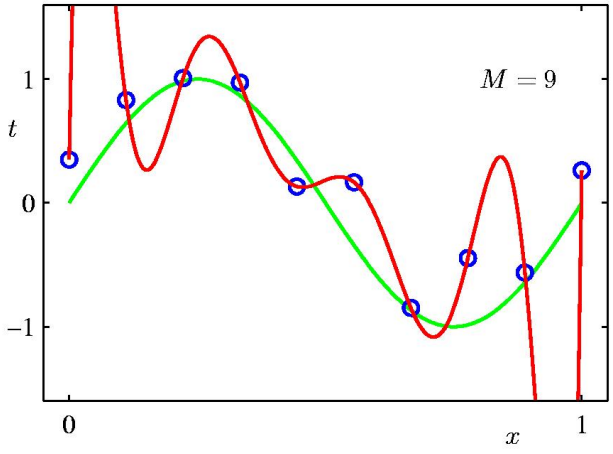


Which of the predicted curve is better?





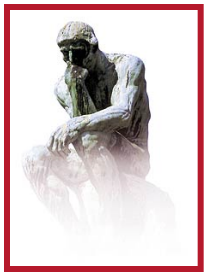




Blue: Observed data
Red: Predicted curve
True: Green true distribution

What do we really want?

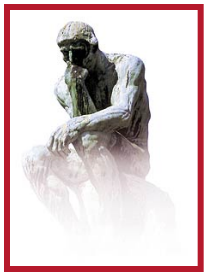
- Why not choose the method with the best fit to the data?



What do we really want?

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

If we were to ask you the homework questions in the midterm, would we have a good estimate of how well you learned the concepts?

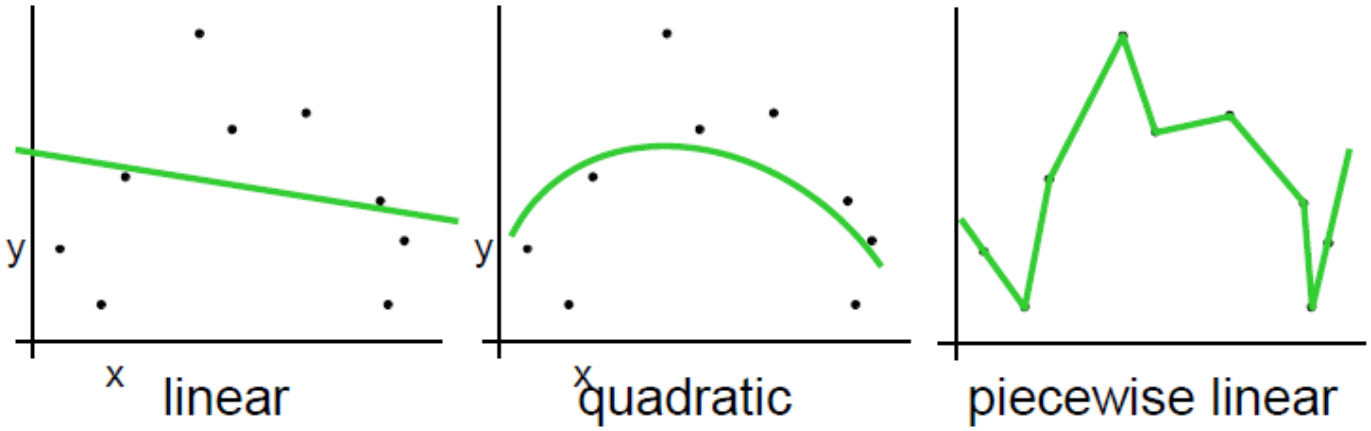


What do we really want?

- 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?

Example



General strategy

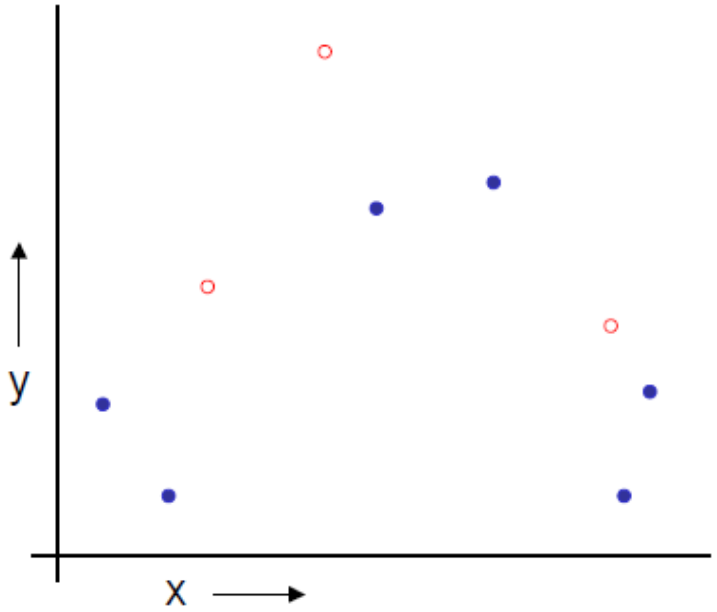
You try to simulate the real word scenario. Test data is your future data. Put it away as far as possible don't look at it.

Validation set is like your test set. You use it to select model. The whole aim is to estimate the models' true error on the sample data you have.

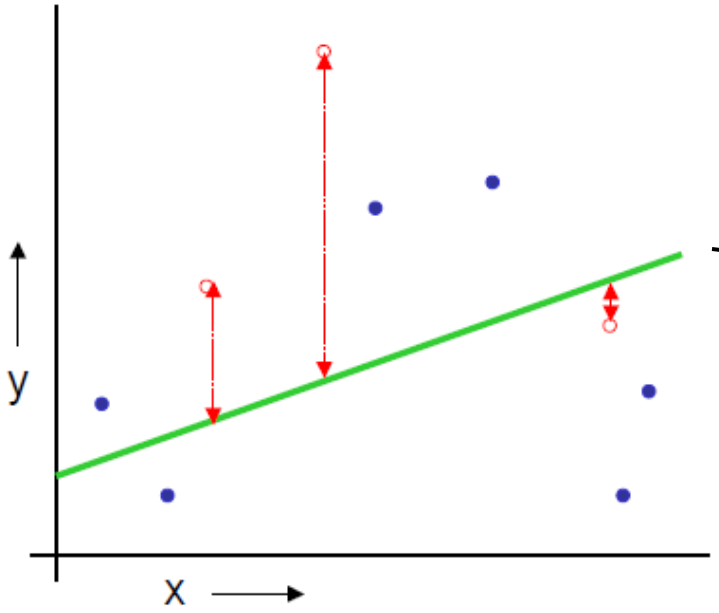
!!! For the rest of the slides ..Assume we put the test data already away. Consider it as the validation data when it says test set.

Test set method

- Randomly split some portion of your data
  Leave it aside as the test set
- The remaining data is the training data



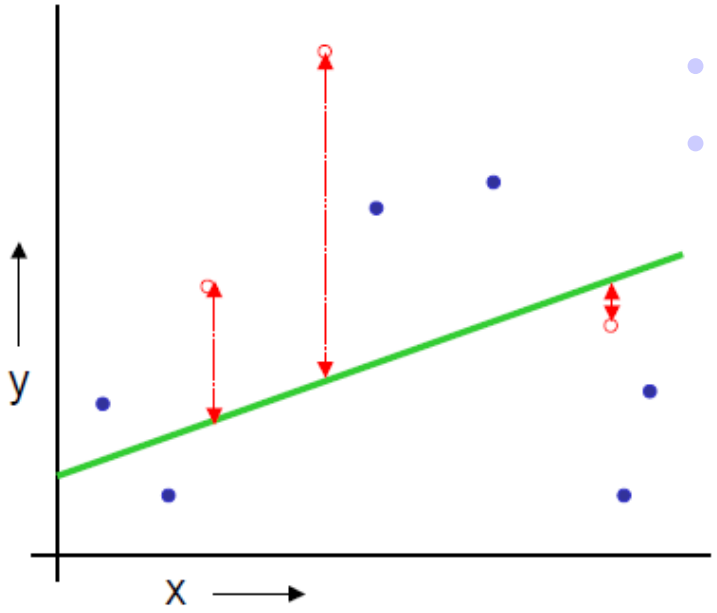
Test set method



- Randomly split some portion of your data
  Leave it aside as the test set
  The remaining data is the training data
  Learn a model from the training set

This is the model you learned.

How good is the prediction?



- Randomly split some portion of your data
  Leave it aside as the test set
- The remaining data is the training data
- Learn a model from the training set
- Estimate your future performance with the test data

Train test set split

- It is simple
- What is the down side ?

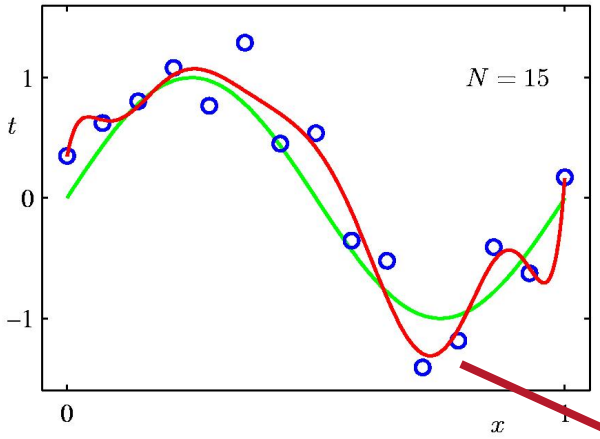
More data is better

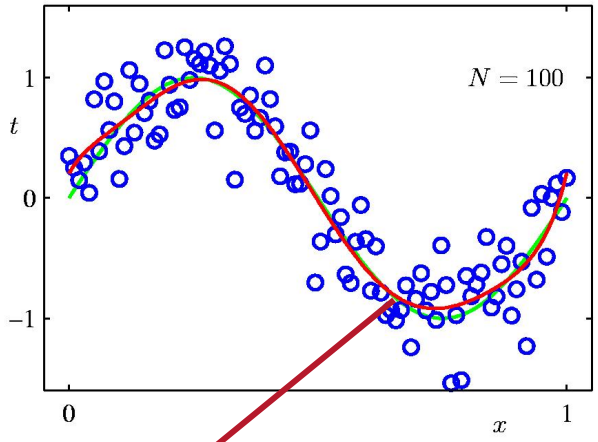
With more data you can learn better

Blue: Observed data

Red: Predicted curve

True: Green true distribution





Compare the predicted curves

Train test set split

- It is simple
- What is the down side ?

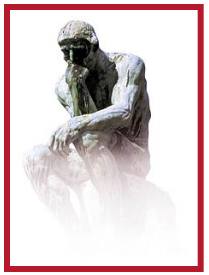
1. You waste some portion of your data.

Train test set split

- It is simple
- What is the down side ?

1. You waste some portion of your data.

What else?



Train test set split

- It is simple
- What is the down side ?

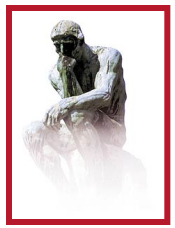
1. You waste some portion of your data.
2. You must be luck or unlucky with your test data

Train test set split

- It is simple
- What is the down side ?

1. You waste some portion of your data.
2. If you don't have much data, you must be luck or unlucky with your test data

How does it translate to statistics?
Your estimator of performance has …?



Train/test set split

- It is simple
- What is the down side ?

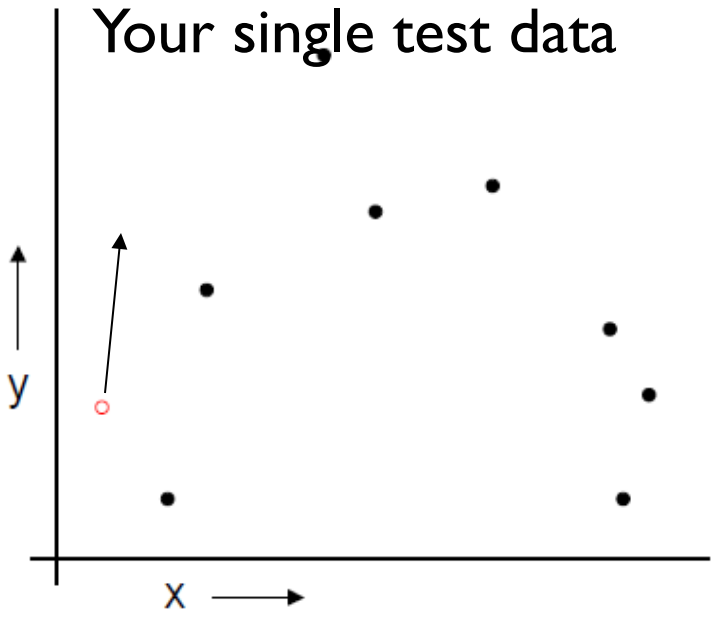
1. You waste some portion of your data.
2. If you don't have much data, you must be luck or unlucky with your test data

How does it translate to statistics?
Your estimator of performance has high variance

Cross Validation

Recycle the data!

LOOCV (Leave-one-out Cross Validation)



Let say we have N data points
k be the index for data points
k=1..N

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

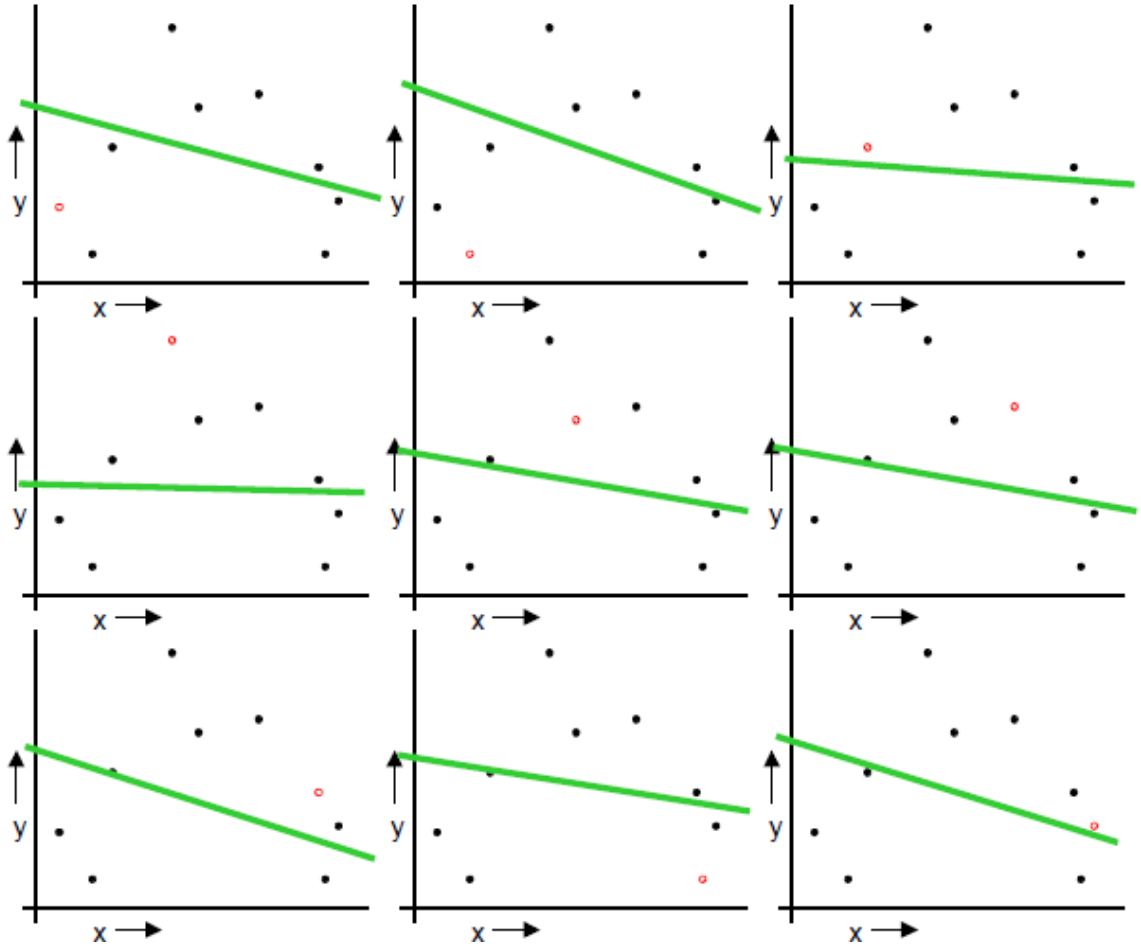
Temporarily remove (x_k,y_k) from the dataset

Train on the remaining N-1 Datapoints

Test your error on (x_k,y_k)

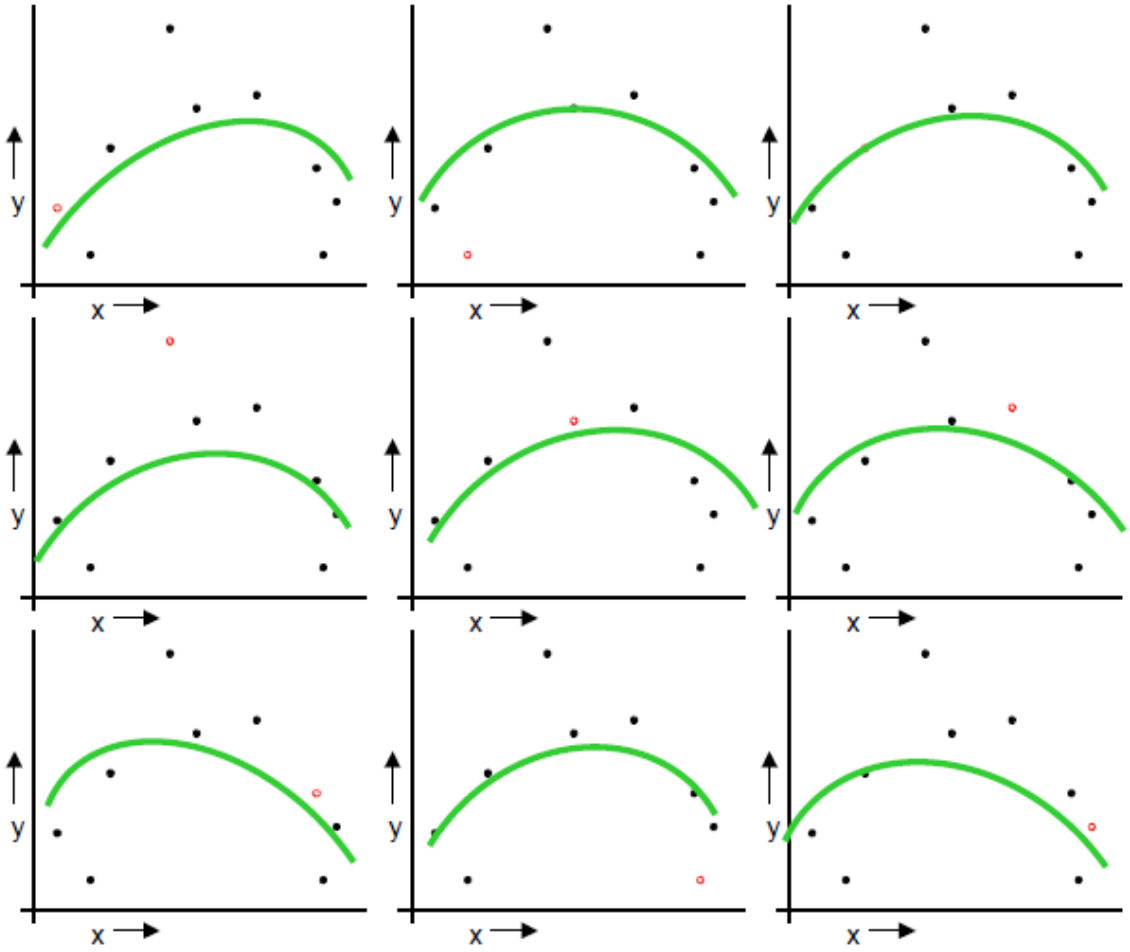
Do this for each k=1..N and report the mean error.

LOOCV (Leave-one-out Cross Validation)



There are N data points..
Do this N times. Notice the test data is changing each time

LOOCV (Leave-one-out Cross Validation)



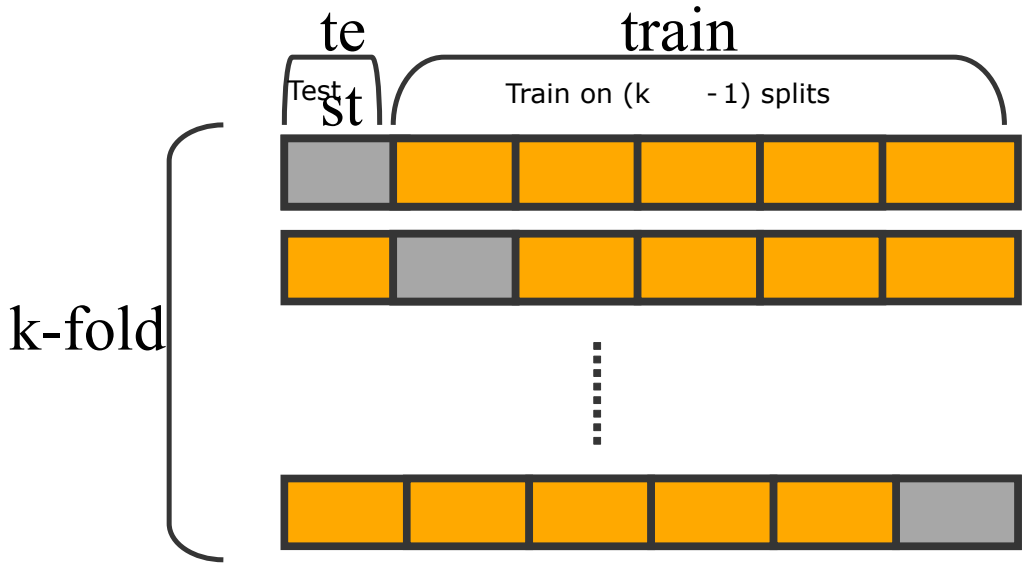
There are N data points..
Do this N times. Notice the test data is changing each time

MSE=3.33

What's the problem of LOOCV?

The computation is expensive!

K-fold cross validation



In 3 fold cross validation, there are 3 runs.
In 5 fold cross validation, there are 5 runs.
In 10 fold cross validation, there are 10 runs.

the error is averaged over all runs

Model Selection

- In-sample error estimates:
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - Minimum Description Length Principle (MDL)
  - Structural Risk Minimization (SRM)

- Extra-sample error estimates:
  - Cross-Validation
  - Bootstrap