What is Cross Validation?

Talks n' Beer III

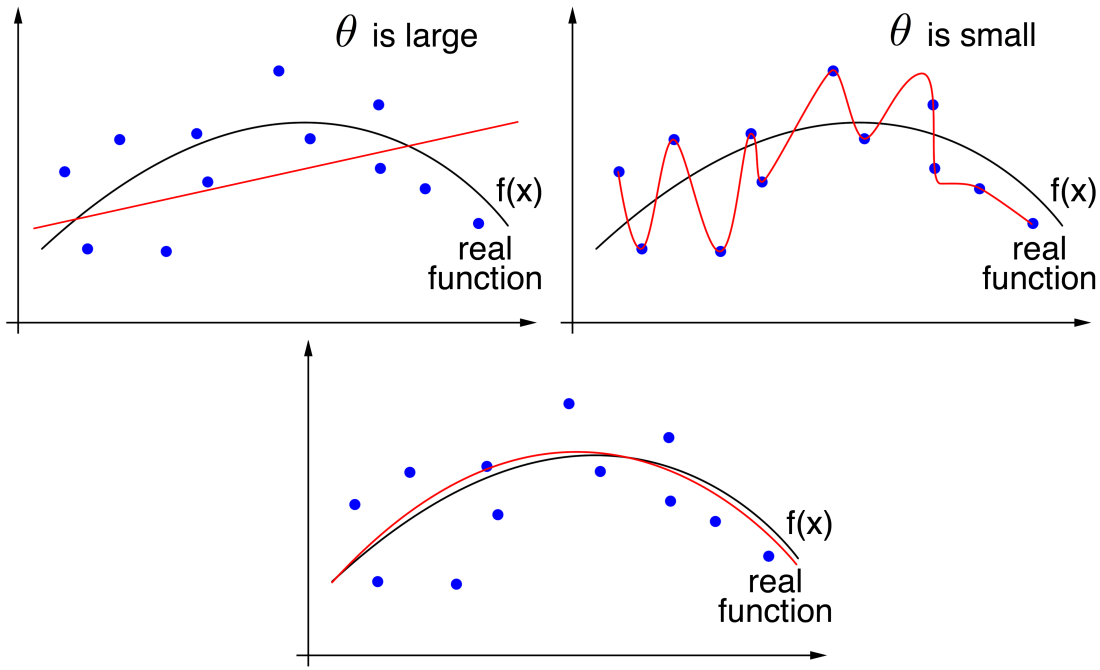
27. Juni 2013

Given:
Labeled data $(x_i, y_i)_{i=1}^{N}$, with $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$ and the data that we want to predict $x^* \in \mathbb{R}^D$.
Goal:
Do a correct prediction of the labels $\hat{y}^*$ for each point x^*.
Predictions are made with a so called model:

$$\hat{f}(x, \theta) = \hat{y} \text{ and } \theta \in \mathbb{R} \tag{1}$$

Example for a model with different θ parameters

The blue points are our given measured data $(x_i, y_i)_{i=1}^{N}$. We like to fit a model $\hat{f}(x, \theta)$. θ determines how flexibel our model can be.



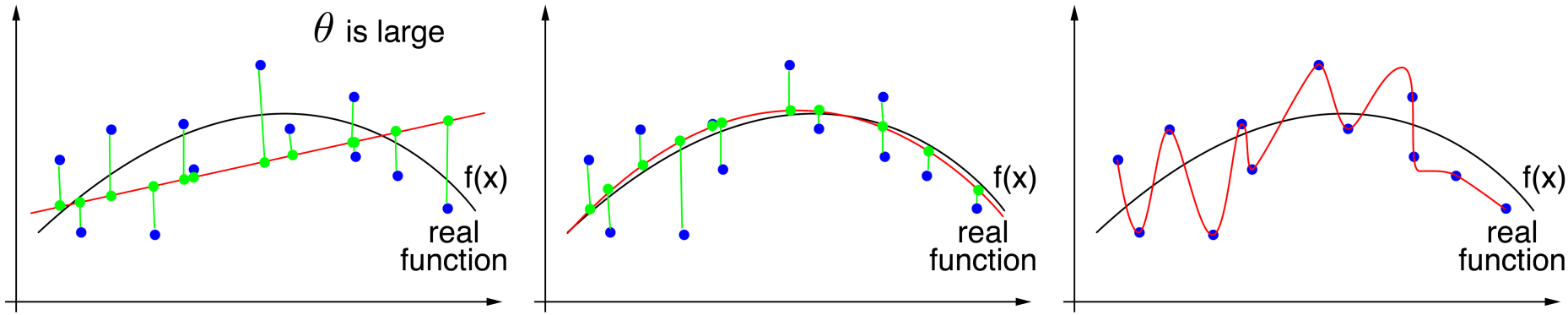
Depending on the choice of our parameter θ we get different prediction accuracies for our observed data.

How do we determine the goodness of fit?

We simply apply an error measure, i.e.

$$\text{err} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2}$$

And then we try to minimize this on our data.



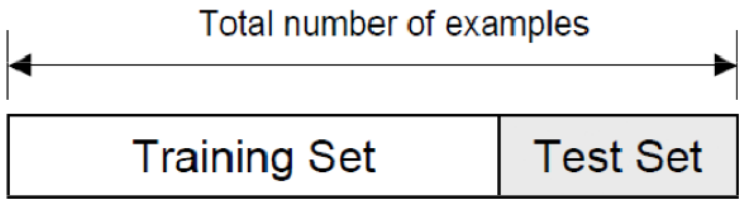
Result: We would get the best fit for the small θ according to our error measure.

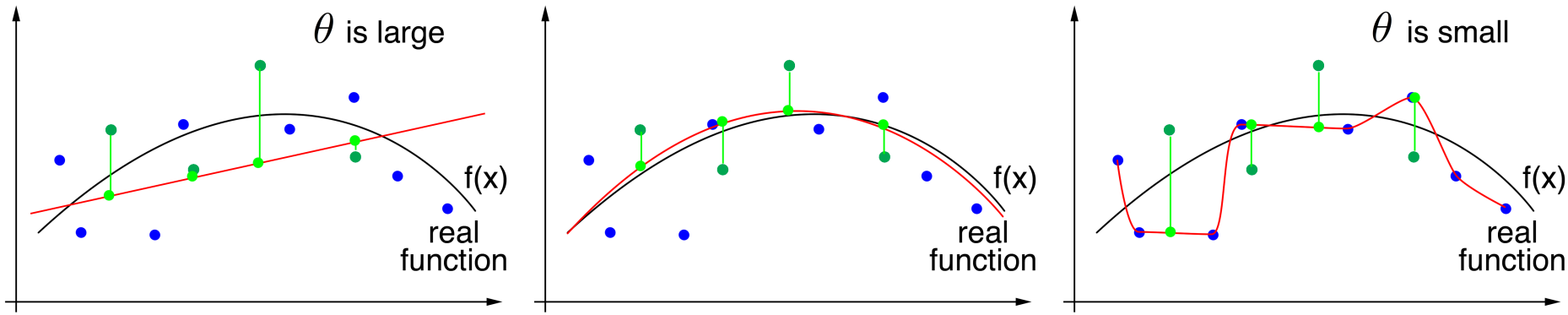
What is the problem?

The problem is, that we will overfit the model such that we perfectly predict the data we learned on.
But we would perform very very bad on new unseen data.
Even worse: This way we wouldn't even notice it! The solution?

The solution is Cross Validation

1. Ignore a part of the training data and adjust the parameter on the rest.
2. Then evaluate your model on the ignored part.

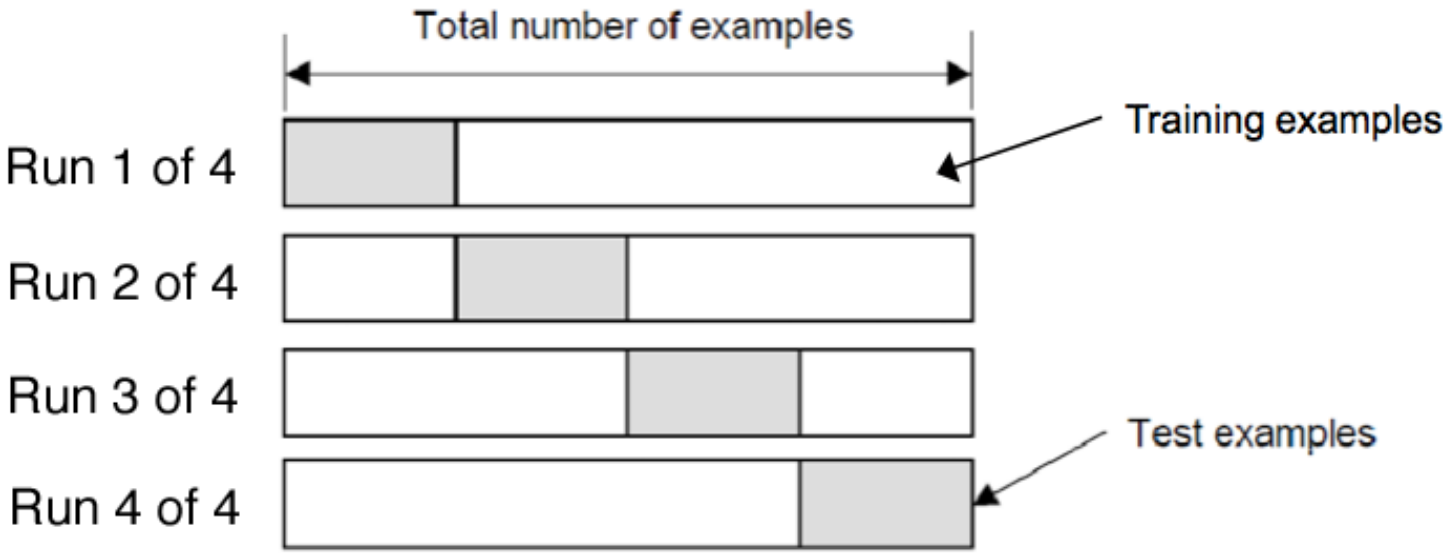




Plot:	Left	Middle	Right
Error:	1.98	1.45	2.14

Cross Validation

But we can become even better! Instead of just one split we will split up the data into several parts, e.g. 4 folds



Calculate the average of the error we made on each testing data

$$\text{err}_{\text{avg}} = \frac{1}{4}\sum_{i=1}^{4}\text{err}_i \tag{3}$$

Select the model parameters θ that yield the lowest average error.

What is the benefit of Cross Validation?

1. Simple yet powerful method for estimating parameters in a supervised setting
2. Error is statistically estimated (implies further statistical analysis of the method itself)
3. Can be applied onto nearly any model

Therefore cross validation is a very important method that everybody who works with data should know!
No excuses!

Thx

Thank you!