LINK TAG DER MARKTFORSCHUNG
24. SEPTEMBER 2013
ZWISCHEN MOBILER REVOLUTION UND EVOLUTION BEWÄHRTER METHODEN

INSTITUT

Cross Validation

Dr. Thomas Jensen

Expedia.com

About Me

- PhD from ETH
- Used to be a statistician at Link, now Senior Business Analyst at Expedia
- Manage a database with 720,000 Hotels that are not on contract with Expedia
- Manage two algorithms
  - One to assign a dollar value to each hotel in the database
  - Another to forecast how well/bad we are doing in terms of room availability in a given hotel
- Responsible for metrics

Use Cross Validation
to optimize the algorithms

Overview

- What is R2 and why can it be misleading
- The issue of overfitting
- Definition of cross validation
- How to conduct cross validation
- Best practices when evaluating model fit

How to check if a model fit is good?

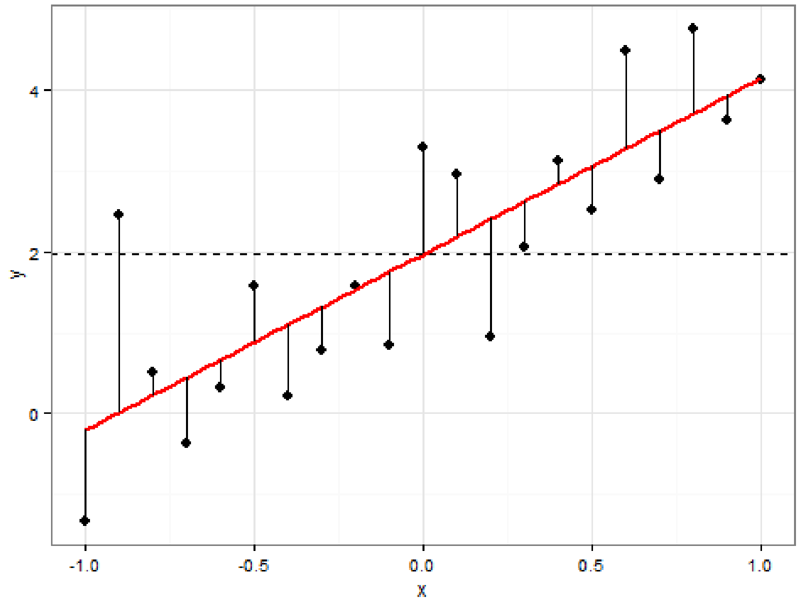
- The **R2 statistic** has become the almost universally standard measure for model fit in linear models
- What is R2?
  - $\boldsymbol{R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2}}$

    Model error

    Variance in the dependent variable
  - It is the ratio of error in a model over the total variance in the dependent variable.
  - Hence the lower the error, the higher the R2 value.

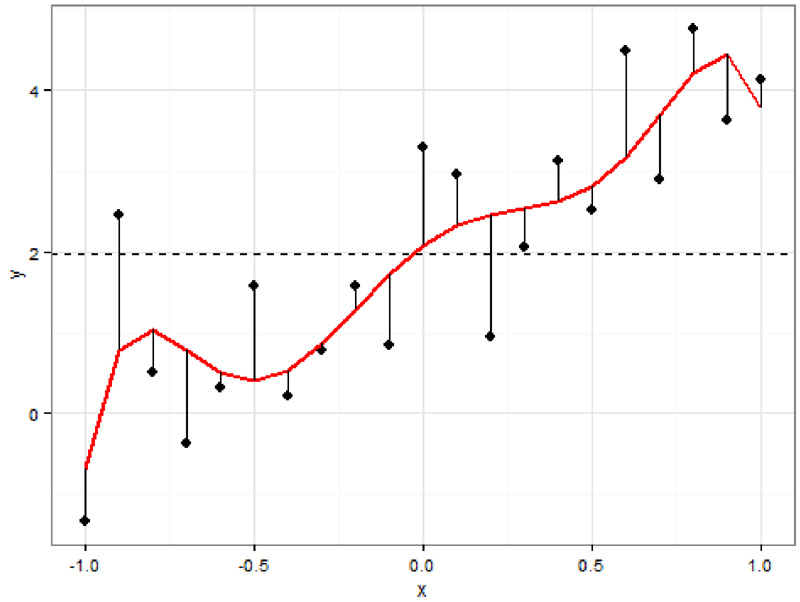
How to check if a model fit is good?

- $\sum(y_i - f_i)^2 = 18.568$
- $\sum(y_i - \bar{y})^2 = 55.001$
- $R^2 = 1 - \frac{18.568}{55.001}$
- $R^2 = 0.6624$
- A decent model fit!



How to check if a model fit is good?

- $\sum(y_i - f_i)^2 = 15.276$
- $\sum(y_i - \bar{y})^2 = 55.001$
- $R^2 = 1 - \dfrac{15.276}{55.001}$
- $R^2 = 0.72$



- Is this a better model?
- No, overfitting!

Overfitting

- Left to their own devices, modeling techniques will overfit the data.
- Classic Example: multiple regression
  - *Every* time you add a variable to the regression, the model's R^2 goes up.
  - Naïve interpretation: *every* additional predictive variable helps explain yet more of the target's variance.
  - But that can't be true!
  - Left to its own devices, Multiple Regression will fit *too many* patterns.
  - A reason why modeling requires subject-matter expertise.

Overfitting

Error on the dataset used to *fit* the model can be misleading

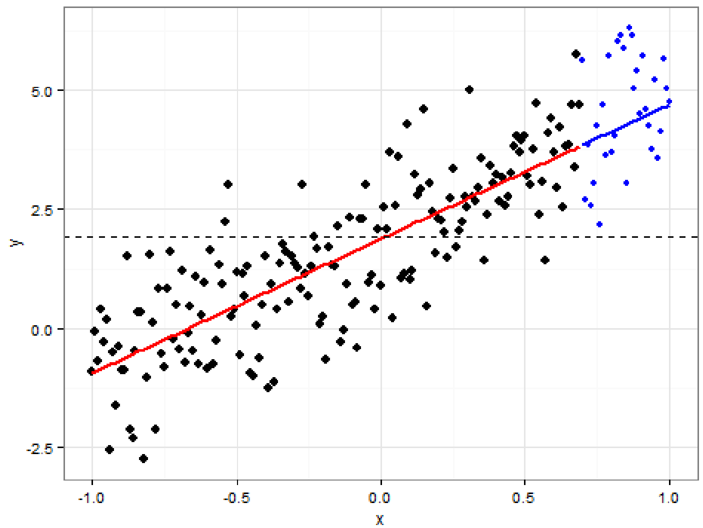
- Doesn't predict future performance.

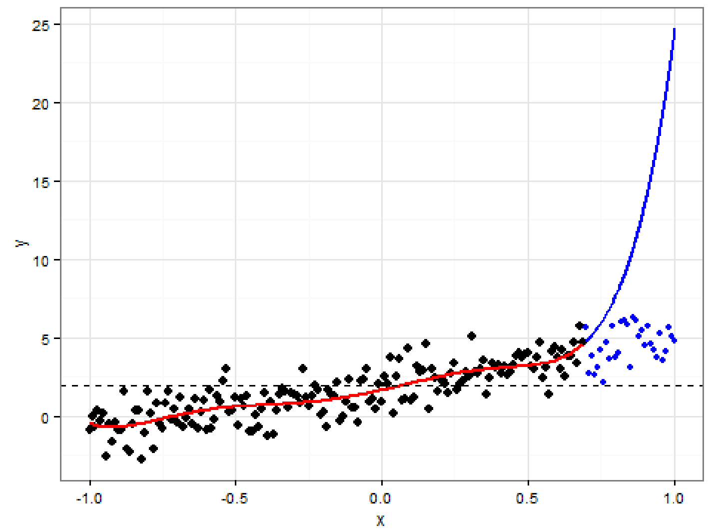
Too much complexity can diminish model's accuracy on future data.

- Sometimes called the Bias-Variance Tradeoff.

Overfitting Cont.

- What are the consequences of overfitting?
  - *"Overfitted models will have high R2 values, but will perform poorly in predicting out-of-sample cases"*



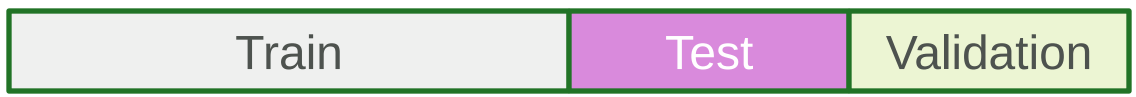


What is Cross Validation

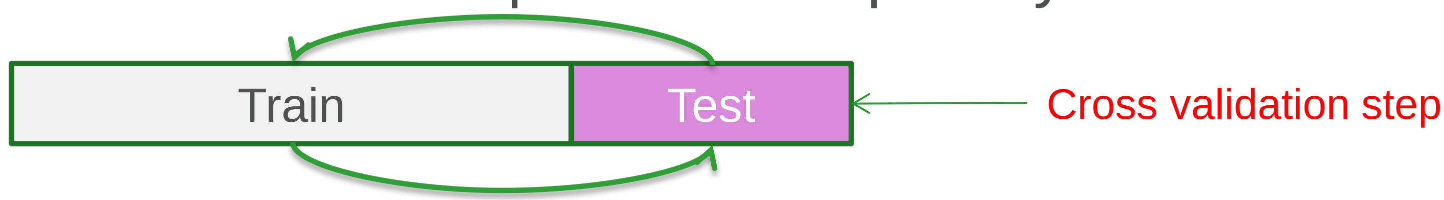
- “A method of assessing the accuracy and validity of a statistical model. The available data set is divided into two parts. Modeling of the data uses one part only. The model selected for this part is then used to predict the values in the other part of the data. A valid model should show good predictive accuracy.”

Cross Validation – the ideal procedure

1. Divide data into three sets, training, validation and test sets



2. Find the optimal model on the training set, and use the test set to check its predictive capability



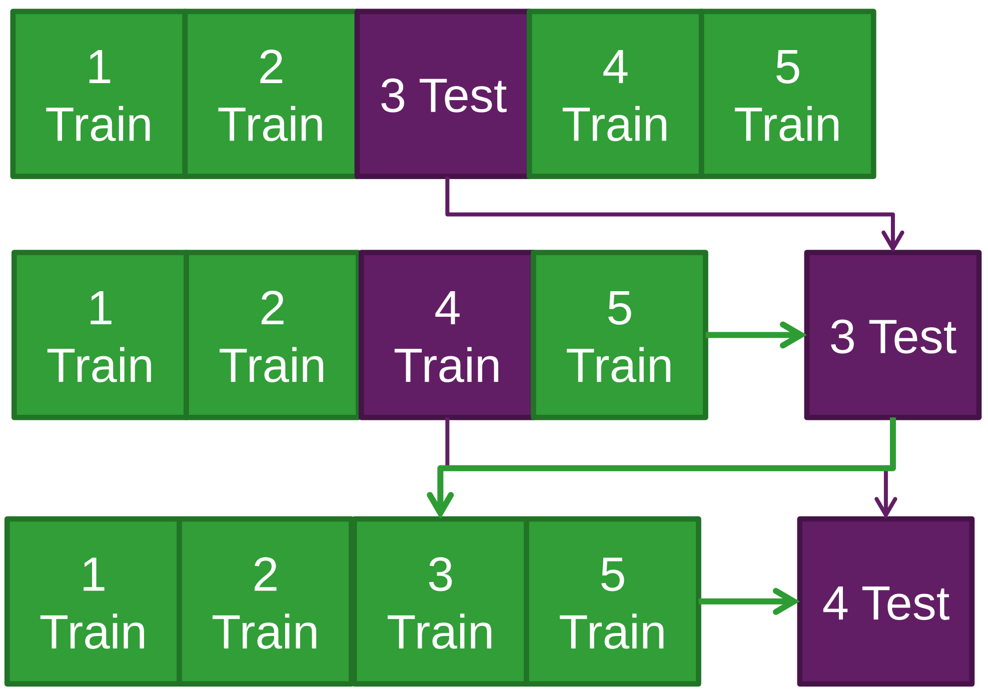
3. See how well the model can predict the test set

4. The validation error gives an unbiased estimate of the predictive power of a model

K-fold Cross Validation

- Since data is often scarce, there might not be enough to set aside for a validation sample
- To work around this issue k-fold CV works as follows:
  1. Split the sample into k subsets of equal size
  2. For each fold estimate a model on all the subsets except one
  3. Use the left out subset to test the model, by calculating a CV metric of choice
  4. Average the CV metric across subsets to get the CV error
- This has the advantage of using all data for estimating the model, however finding a good value for *k* can be tricky

K-fold Cross Validation Example



1. Split the data into 5 samples

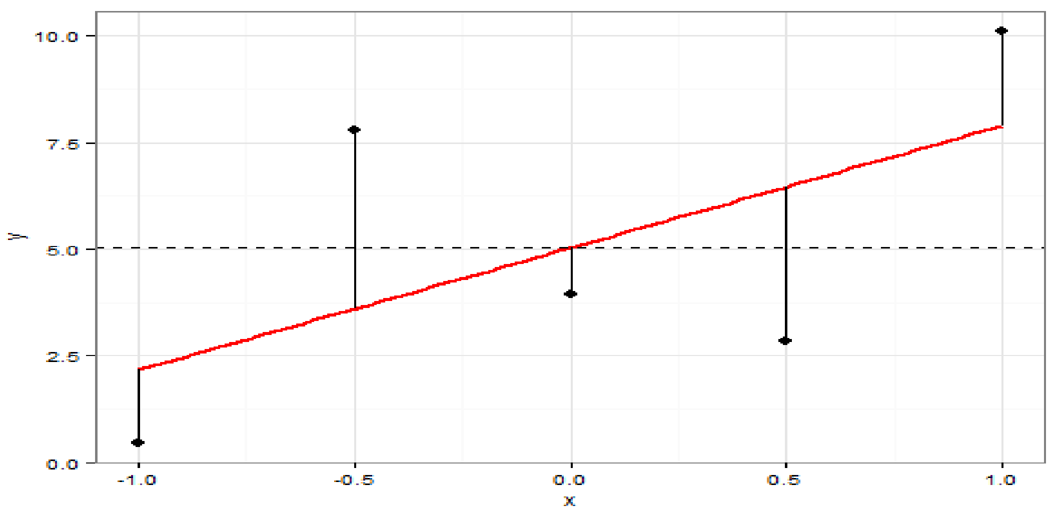
2. Fit a model to the training samples and use the test sample to calculate A CV metric.

3. Repeat the process for the next sample, until all samples have been used to either train or test the model

Cross Validation - Metrics

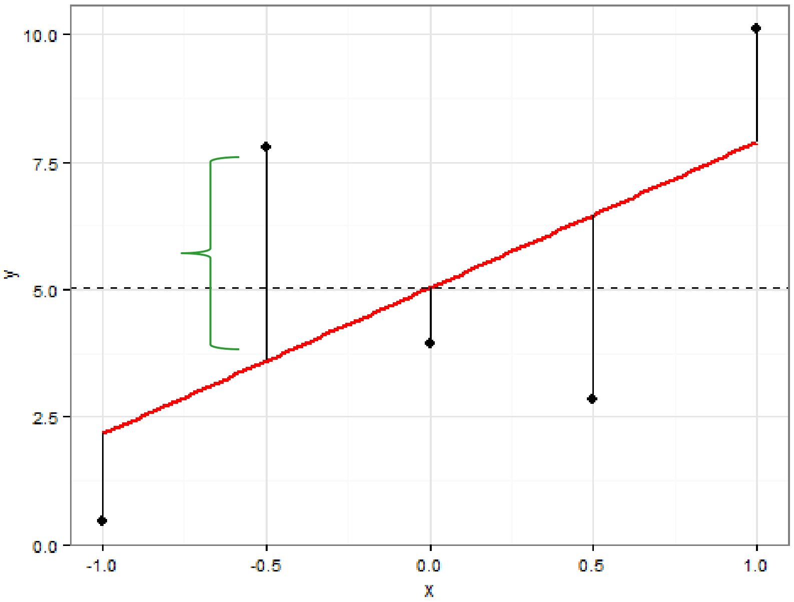
- How do we determine if one model is predicting better than another model?
- The basic relation:
  - $Error_i = y_i - f_i$

The difference between observed (*y*) and predicted value (*f*), when applying the model to unseen data



Cross Validation Metrics

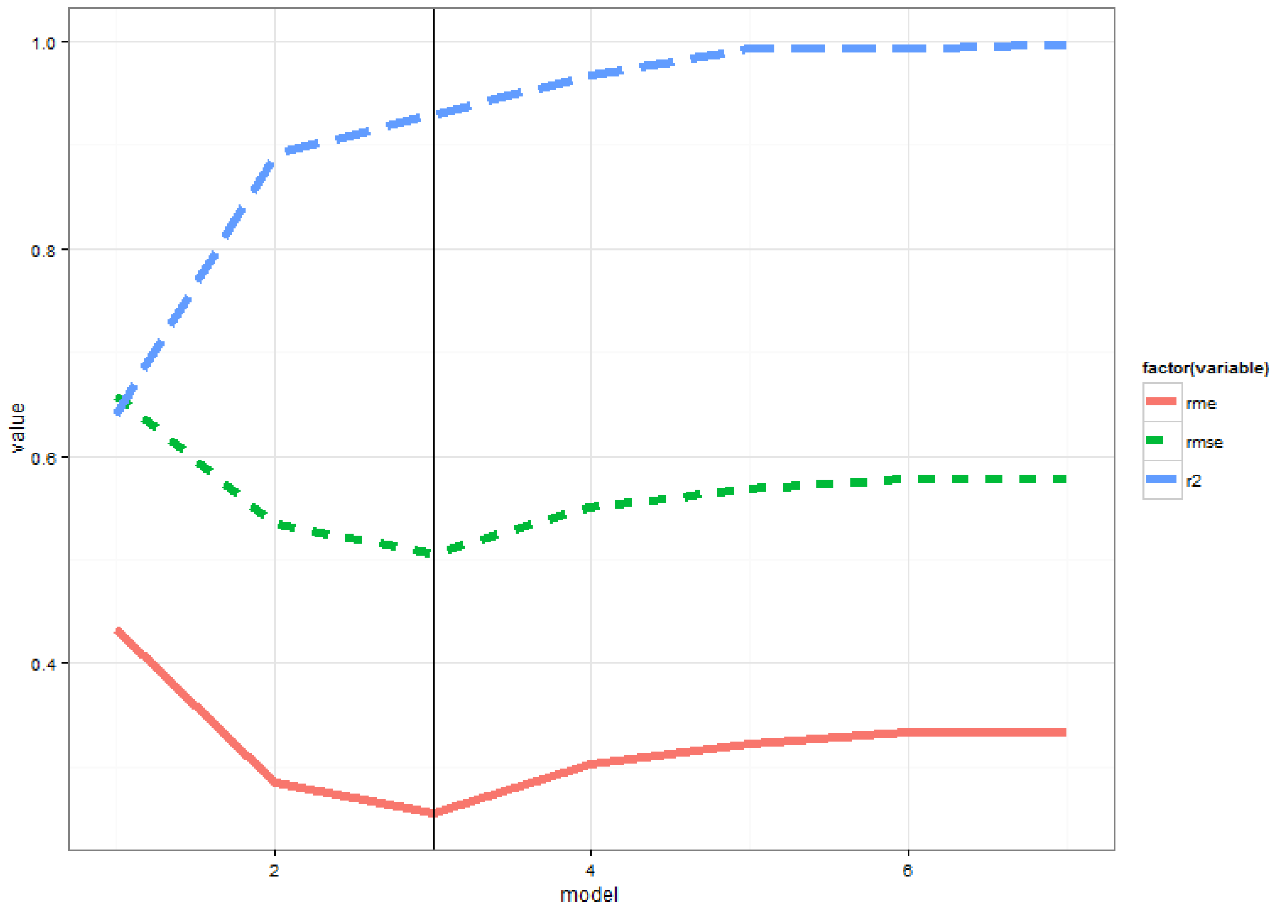
- Mean Squared Error (MSE)
  - $1/n\sum(y_i - f_i)^2$
  - 7.96
- Root Mean Squared Error (RMSE)
  - $\sqrt{1/n\sum(y_i - f_i)^2}$
  - 2.82
- Mean Absolute Percentage Error (MAPE)
  - $(1/n\sum|\frac{y_i - f_i}{y_i}|)$*100
  - 120%



Simulation Example

- Simulated 8 variables from a standard gaussian. The first three variables where related to the dependent variable, the other variables where random noise.
- The true model should have the first three variables, while the other variables should be discarded.
- By using Cross Validation, we should be able to avoid overfitting.

Simulation Example



Best Practice for Reporting Model Fit

1. Use Cross Validation to find the best model
2. Report the RMSE and MAPE statistics from the cross validation procedure
3. Report the R Squared from the model as you normally would.

The added cross-validation information will allow one to evaluate not how much variance can be explained by the model, but also the predictive accuracy of the model. Good models should have a high predictive AND explanatory power!

Conclusion

The take home message

Only looking at $R2$ can lead to selecting models that do not have inferior predictive capability. Instead market researchers should evaluate a model according to a combination of $R2$ and cross validation metrics. This will produce more robust models with better predictive powers.