Practical Machine Learning in R

Resampling

Lars Kotthoff\textsuperscript{1,2}
lar sko@uwyo.edu

\textsuperscript{1}with slides from Bernd Bischl and Michel Lang
\textsuperscript{2}slides available at http://www.cs.uwyo.edu/~lar sko/ml-fac
Why do we care?

▷ want to learn general relationships
▷ extreme case: model memorizes data

Why do we care?

▷ want to learn general relationships
▷ extreme case: model memorizes data

My Hobby: Extrapolating

As you can see, by late next month you’ll have over four dozen husbands. Better get a bulk rate on wedding cake.
Example: Polynomial Regression

\[ y = 0.5 + 0.4 \cdot \sin(2\pi x) + \epsilon \]
Model Complexity

- Model complexity ≈ model flexibility
- more complex models can capture more complex relationships
- here: degree of polynomial

![Graph showing model complexity with varying degrees of polynomial (d = 1, d = 3, d = 10).]
Mean squared error for model on training data:

\[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\[ d = 1: 0.04583, \quad d = 3: 0.00182, \quad d = 10: 0.00000 \]

▷ more complex model better?
▷ independent test set
Test Error (Generalization)
Test Error (Generalization)

train: 0.00000, test: 0.00640
Test Error (Generalization)

test error is best for $d = 5$
Bias-Variance Trade-Off

**Bias**: Systematic error of the fitted model

**Variance**: Variance of the fitted models for different samples

Example:

- A polynomial with too few parameters (a too low degree) will make large errors because of a large bias.
- A polynomial with too many parameters (a too high degree) will make large errors because of a large variance.
- Both bias and variance must be small to achieve a good generalization error.
Resampling

- goal: estimate generalization error of model
- (repeatedly) fit models on training sets
- evaluate performance on independent test sets and average performance measure
Subsampling

- randomly sample part of data for training, remainder for testing
- repeat
- holdout = one iteration of subsampling
Stratified Sampling

▷ make sure that sampled data set is representative
▷ e.g. for classification: all classes present with respective percentages

By Dan Kernler - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=36506021
Bootstrap

- randomly sample from data \textit{with replacement}
- training set = unique samples, remainder test set
- repeat
Cross-Validation

- partition data into $k$ sets (folds) of equal size
- use $k - 1$ for training, remainder for testing
- repeat for all possible combinations of train and test sets ($k$ times)
- leave-one-out cross-validation: $k$ equal to total amount of data

By Fabian Flöck - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=51562781
Exercises