Estimating the Performance of Predictive Models with Resampling Methods

Florian Pargent (Florian.Pargent@psy.lmu.de)
Ludwig-Maximilians-Universität München

Why Do We Need Resampling?

How Does Resampling Work?

How To Avoid Common Mistakes?

Why Do We Need Resampling?

Predictive Modeling in Psychology

Breiman and others (2001), Shmueli (2010), Yarkoni and Westfall (2017)

- psychology has a (too) heavy focus on explanation (Yarkoni and Westfall 2017)
- predictive claims (e.g. meta analyses) often not based on realistic estimates of predictive accuracy
- "this has led to irrelevant theory and questionable conclusions…" (Breiman and others 2001)
- increasing amounts of high-dimensional data: complex relationships, hard to hypothesize
- create new measures, reflect on and improve existing theories
- investigate whether theories predict relevant target variables (Shmueli 2010)

Predictive Models

Definition adapted from Kuhn and Johnson (2013)

A predictive model is any (statistical) model that generates (accurate) predictions of some target variable, based on (a series of) predictor variables.

Examples:

- ordinary linear regression
- penalized linear models: lasso, ridge, elastic net
- tree models: decision tree, random forest, gradient boosting
- support vector machines
- neural networks
- ...

Predictive Performance Estimation

The quality of a (fixed) predictive models is evaluated based on its generalization error on new (unseen) data, drawn from the same population:

> *"How well does this predictive model I have already estimated work when I use it to predict observations from my practical application, in which I do not know the target values?"*

First: What is our definition of error (or accuracy)?

Performance Measures for Regression Problems

Quantify a "typical" deviation from the true value!

The statistician's favorite:

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

The social scientist's favorite:

$$R^2 = 1 - \frac{residual\ sum\ of\ squares}{total\ sum\ of\ squares} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

How Does Resampling Work?

Resampling Methods

Plan for Today:

- Holdout
- Cross-Validation
- Repeated Cross-Validation

Further Methods:

- Leave-One-Out Cross-Validation
- Subsampling
- Bootstrap
- ...

Training and Test Set

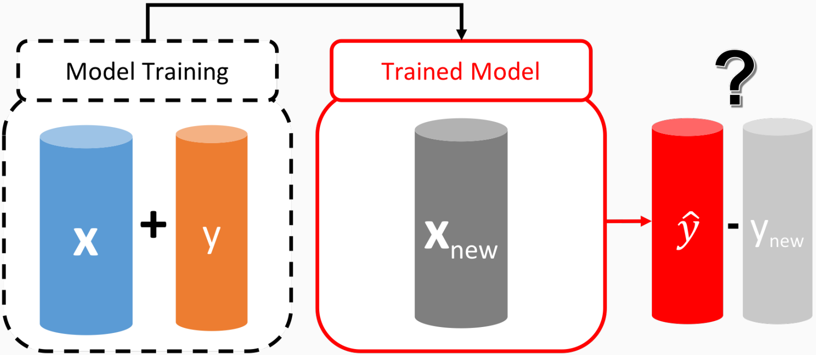
- How well does our model predict **new data** (iid)?
  - Option 1: collect new data **;-)**
  - Option 2: use prediction error in-sample **:-(**
  - Option 3: use available data in a smart way **:-)**

To estimate the performance of our model, split the dataset:

- **Training set**: train the algorithm
- **Test set**: compute performance

-> Holdout – Estimator

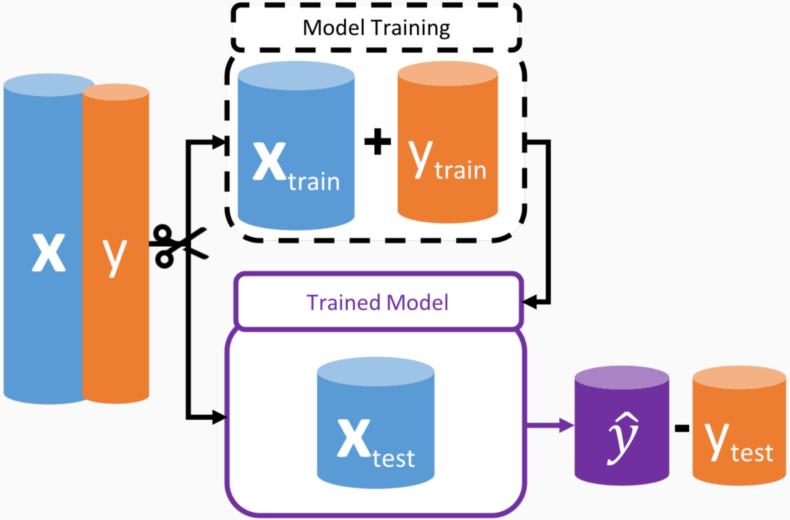
General Idea of Performance Evaluation I



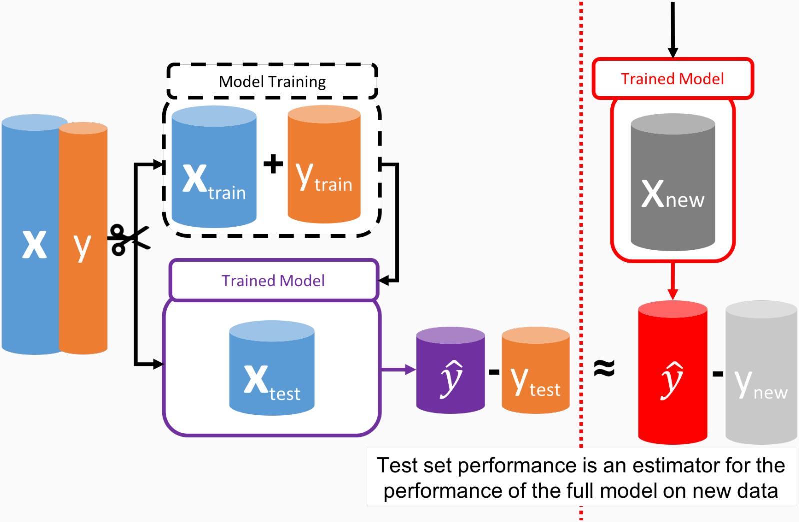
Learn a functional relationship between **X** and y

How does the model perform during application?

General Idea of Performance Evaluation II



General Idea of Performance Evaluation III



IMPORTANT: Do not get confused by the different models!

Full Model:

- trained on the **whole dataset**
- will be used in **practical applications**

Proxy Model:

- trained on a **training set**
- is only a **tool for performance estimation**
- can be **discarded** after test set predictions

Why Do We Have to Separate Training from Test Data?

To avoid getting fooled by **Overfitting**:

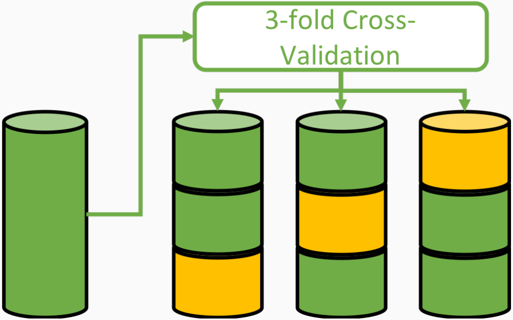
- Model adjusts to a set of given data points too closely
- Sample specific patterns are learned ("fitting the noise")
- Can be compared to "learning something by heart"

Many flexible algorithms predict training data (almost) perfectly:

> *Training ("in-sample") performance is useless to judge the model's performance on new data ("out-of-sample")!*

Improving the Holdout Estimator: Cross-Validation

- **Bias reduction** via big training sets
- **Variance reduction** via aggregation
- Random partitioning in k equally sized parts (often 5 or 10)
- Each part test set once, remaining parts combined training set
- Average the estimated prediction error from all *folds*



Do Not Program Everything Yourself!

Machine learning meta packages in R:

- **mlr** package (Bischl et al. 2016):
  - standardized interface for machine learning
  - detailed tutorial at https://mlr-org.github.io/mlr/
  - mlr-org packages: mlrCPO, mlrMBO, ...

- Alternatives:
  - **caret** package (Kuhn and Johnson 2013)
  - **tidymodels** packages (Max and Wickham 2018)

EXAMPLE: Life Satisfaction

Pargent and Albert-von der Gönna (in press):

- predictive modeling with the GESIS Panel (Bosnjak et al. 2018)
- today's demo: ***Satisfaction Life (Overall)***

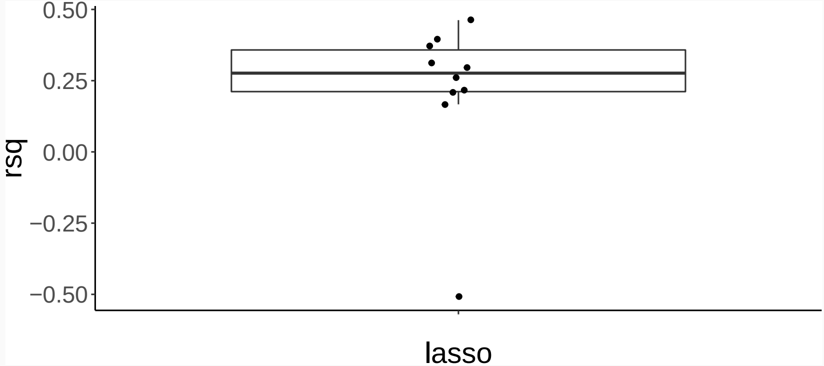
Now we would like to know how satisfied you are with life overall.
Fully unsatisfied | **0 1 2 3 4 5 6 7 8 9 10** | *Fully satisfied*

- 1975 predictor variables
- only use 250 of originally 2389 panelists
- simplified imputation
- **predictive algorithm:** regularized linear model (lasso) by Tibshirani (1996)

EXAMPLE: Insample vs. Out-Of-Sample Performance

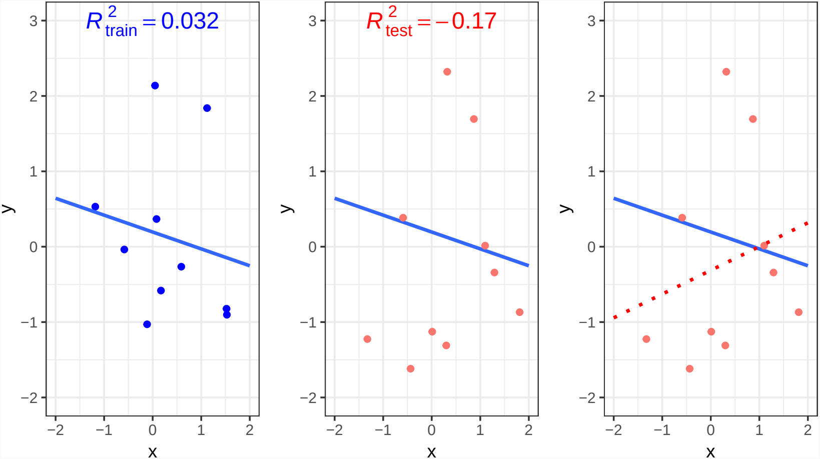
$R^2_{insample} = \mathbf{0.41}$ (insample estimate)

$R^2_{CV} = \mathbf{0.22}$ (estimate from 10-fold CV)



What about that ***NEGATIVE*** R^2 ???

R^2 Can Be Negative Out-Of-Sample



- Train model on training data (positive R^2_{train})
- Predict test data with trained model (negative R^2_{test})

Improving Cross-Validation: Repeated Cross-Validation

Problem:

Cross-validation estimates can be unstable for small datasets...

3 different seeds for our **Life Satisfaction** example:

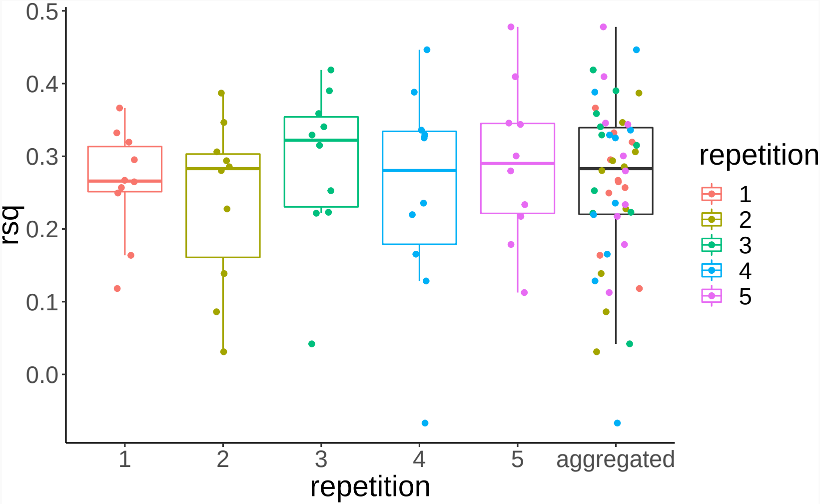
	seed.1	seed.2	seed.3
rsq.test.mean	0.22	0.2	0.3

Solution:

Repeat k-fold cross-validation r times and aggregate the results

EXAMPLE: 5 Times Repeated 10-Fold CV

$R^2_{RepCV} = \mathbf{0.27}$ (estimate from 5 times repeated 10-fold CV)



How To Avoid Common Mistakes?

Variable Selection Done Wrong

Common mistake with many predictor variables:

- correlate all predictors with the target in the complete dataset
- choose the same highly correlated predictors in resampling
- **Problem:** The decision of which variables to select is based on the complete dataset (training set + test set)
  –> **Overfitting**

Don't fool yourself! This shares similarities with...

- *multiple testing*
- *p-hacking*
- *garden of forking paths*

EXAMPLE: Variable Selection Wrong vs. Right

- select the 10 predictors with the highest correlation with the target variable ***Satisfaction life (Overall)***
- ordinary linear model
- 5-fold cross-validation

Variables selected based on the whole dataset:

$R^2_{CV} = 0.38$

Variables selected in each cross-validation fold:

$R^2_{CV} = 0.26$

EXAMPLE: Selected Variables Differ Between Folds!

full model	fold 1	fold 2	fold 3	fold 4	fold 5
dazb025a	dazb025a	dazb021a	dazb021a	dazb025a	dbaw239a
dazb027a	dazb027a	dazb025a	dazb025a	dazb027a	dbaw245a
dbaw239a	dbaw239a	dbaw239a	dazb027a	dbaw239a	debl230a
dbaw245a	dbaw245a	dbaw245a	dbaw239a	dbaw245a	deaw258a
deaw259a	dcaw172a	deaw259a	dbaw245a	deaw259a	deaw259a
deaw265a	deaw259a	deaw267a	deaw259a	deaw265a	deaw265a
deaw267a	dfaw112a	eazb021a	deaw265a	deaw267a	deaw267a
dfaw106a	eazb025a	eazb027a	deaw267a	dfaw106a	dfaw106a
eazb025a	eazb027a	eaaw136a	eazb026a	eaaw135a	eaaw135a
eaaw136a	eaaw136a	eaaw142a	eaaw136a	eaaw136a	eaaw136a

Resampling as a Simulation of Model Application

Which steps are performed until the full model is ready for application?

- imputation of missing values
- transformations of predictors
- variable selection
- hyperparameter tuning
- model estimation
- (model selection)

Repeat all steps for each pair of training and test data!

What if some steps need resampling (e.g. hyperparameter tuning)?

Nested Resampling

- **Inner loop:** tuning, preprocessing, variable selection
- **Outer loop:** evaluation of model performance

Augmented/Fused Algorithms

- some machine learning algorithms are implemented with automatic preprocessing or hyperparameter tuning
- with common machine learning software, simple algorithms can be fused with preprocessing strategies

Treat "augmented" algorithms like "simple" algorithms when estimating predictive performance with resampling!

Life Satisfaction Example:

- *our lasso algorithm (cv.glmnet from the glmnet R package) internally tuned the regularization parameter λ with 10-fold CV*
- *we did not need to specify the inner resampling loop ourself*

Take Home Message

When making predictive claims, social scientists should report realistic estimates of predictive performance!

- With resampling methods, we can estimate the performance on new data for **any** predictive model!
    - To do this, we do not have to know how the algorithm works
    - This allows social scientists to "safely" use machine learning
- However, we have to do the resampling **right**!
    - Repeat all steps from model application during resampling
    - Augmented algorithms can be treated as simple algorithms

Materials and Credits

Slides with code will be uploaded to:
https://osf.io/a8qbt/

Paper “Predictive Modeling with Psychological Panel Data”
with **Johannes Albert-von der Gönna**
https://osf.io/zpse3/

Workshop “An Introduction to Machine Learning in R”
with **Clemens Stachl**
https://osf.io/mnfbd/

Lehrstuhl Psychologische Methodenlehre und Diagnostik
Ludwig-Maximilians-Universität München
of **Prof. Markus Bühner**
http://www.psy.lmu.de/pm/index.html

References I

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "Mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170): 1–5.

Bosnjak, Michael, Tanja Dannwolf, Tobias Enderle, Ines Schaurer, Bella Struminskaya, Angela Tanner, and Kai W. Weyandt. 2018. "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The Gesis Panel." *Social Science Computer Review* 36 (1): 103–15. doi:10.1177/0894439317697949.

Breiman, Leo, and others. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

References II

doi:doi:10.1214/ss/1009213726.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. Springer.

Max, Kuhn, and Hadley Wickham. 2018. *Tidymodels: Easily Install and Load the 'Tidymodels' Packages*. https://CRAN.R-project.org/package=tidymodels.

Pargent, Florian, and Johannes Albert-von der Gönna. in press. "Predictive Modeling with Psychological Panel Data." *Zeitschrift Für Psychologie*.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statist. Sci.* 25 (3). The Institute of Mathematical Statistics: 289–310. doi:10.1214/10-STS330.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1). [Royal

References III

Statistical Society, Wiley]: 267–88. http://www.jstor.org/stable/2346178.

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning." *Perspectives on Psychological Science* 12 (6): 1100–1122. doi:10.1177/1745691617693393.