Data Splitting

ML Basics

In-sample prediction error

Estimating the test error with training data

- Setup: Add training optimism $\hat{\omega}$ to training error

$$\widehat{\text{Err}}_{in} = \overline{\text{err}} + \hat{\omega}$$

- Corrected fit measure for OLS regression

$$C_p = \overline{\text{err}} + 2\frac{d}{n}\hat{\sigma}_{\varepsilon}^2$$

- Corrected fit measures for ML-based methods

$$AIC = -\frac{2}{n}LL + 2\frac{d}{n}$$
$$BIC = -2LL + \log(n)d$$

Validation set, test set, CV

Training set & test set

- Estimate prediction error on new data
  1. Fit model using one part of training data
  2. Compute test error for the excluded section

$\rightarrow$ Model assessment

Figure: 80/20 train-test split

Train	*Test*

Training set, validation set & test set

- Compare models and estimate prediction error
  1. Fit models using training part of training data
  2. Choose best model using validation set
  3. Evaluate final model using test set

$\rightarrow$ Model tuning & assessment

Figure: 50/25/25 Train-validation-test split

Train	*Validate*	*Test*

Validation set, test set, CV

Training set & test set

- Estimate prediction error on new data
  1. Fit model using one part of training data
  2. Compute test error for the excluded section

$\rightarrow$ Model assessment

Figure: 80/20 train-test split

Train	*Test*

Training set, validation set & test set

- Compare models and estimate prediction error
  1. Fit models using training part of training data
  2. Choose best model using validation set
  3. Evaluate final model using test set

$\rightarrow$ Model tuning & assessment

Figure: 50/25/25 Train-validation-test split

Train	*Validate*	*Test*

Leave test data untouched until the end of analysis!

Validation set, test set, CV

Cross-Validation

- LOOCV (Leave-One-Out Cross-Validation)
  1. Fit model on training data while excluding one case
  2. Compute test error for the excluded case
  3. Repeat step 1 & 2 n times
- k-Fold Cross-Validation
  1. Fit model on training data while excluding one group
  2. Compute test error for the excluded group
  3. Repeat step 1 & 2 k times (e.g. $k = 5$, $k = 10$)
- Outlook: nested CV, repeated CV, ...

$$CV(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

Validation set, test set, CV

Standard Errors for CV

$$\frac{1}{\sqrt{K}}\text{sd}\{CV_1(\hat{f}^{-(1)}), \ldots, CV_K(\hat{f}^{-(K)})\}$$

Model selection using *k*-Fold Cross-Validation

- Choose model with smallest cross-validated error
- Choose smallest model within one standard error of the smallest cross-validated error (1-SE Rule)

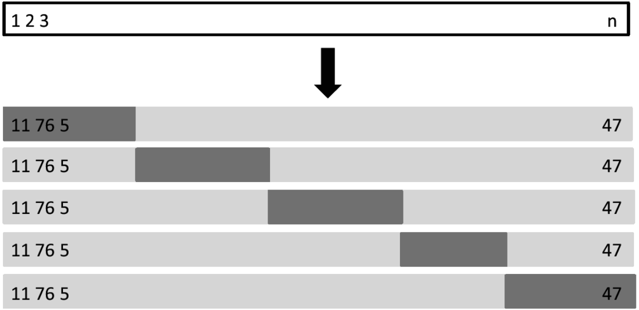
Validation set, test set, CV

More on data splitting

- Simple random splits
  - General approach for “unstructured” data
  - Typically 75% or 80% go into training set
- Stratified splits
  - For classification problems with class imbalance
  - Sampling within each class of Y to preserve class distribution
- Splitting by groups
  - For (temporal) structured data
  - Use specific groups (temporal holdouts) for validation

Validation set, test set, CV

Figure: 5-Fold Cross-Validation with training set and validation set (example)



James et al. (2013)

Learning curves

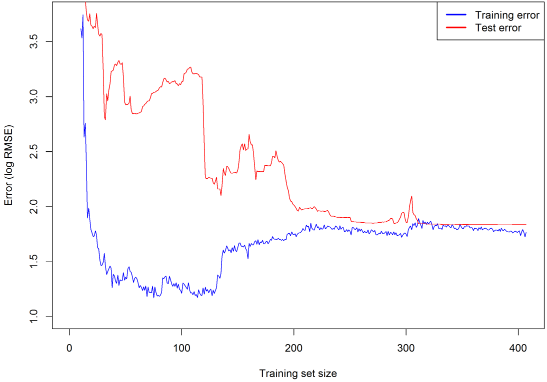
How much data is needed?

- Idea: Plot training and validation error against training set size
- Allows to study the gain of adding more data
  - Convergence of validation error curve towards training curve
- Can also be used as a diagnosis tool to asses
  - High bias (Underfitting): Curves converge at a high value
  - High variance (Overfitting): Large gap between curves

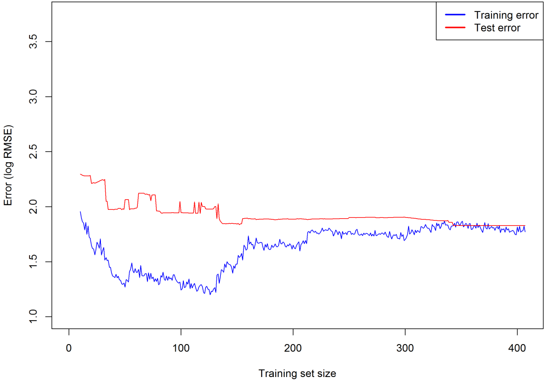
Learning curves

Figure: Learning curves

(a) Linear regression



(b) Regression trees



Performance measures for regression

Performance measures for regression

r^2 score:

$$r^2 = \text{corr}(y_i, \hat{f}(x_i))^2$$

Residual Sum of Squares (RSS):

$$\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Mean of squared errors (MSE):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Root mean squared error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}$$

Performance measures for regression

Mean of absolute errors (MAE):

$$\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{f}(x_i))|$$

Median of absolute errors (MEDAE):

$$\text{median}(|y_1 - \hat{f}(x_1)|, \ldots, |y_n - \hat{f}(x_n)|)$$

Median of squared errors (MEDSE):

$$\text{median}((y_1 - \hat{f}(x_1))^2, \ldots, (y_n - \hat{f}(x_n))^2)$$

Software Resources

Resources for R

- Overview
  - https://cran.r-project.org/web/views/MachineLearning.html
- caret
  - http://topepo.github.io/caret/index.html
- mlr
  - https://mlr-org.github.io/mlr-tutorial/devel/html/
- H2O
  - http://docs.h2o.ai/

References

Buskirk, T. D., Kirchner, A., Eck, A., Signorino, C. S. (2018). An Introduction to Machine Learning Methods for Survey Researchers. *Survey Practice 11*(1).

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM 55*(10), 78–87.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.