Supervised Learning Methodology

Christoph Kern

Mannheim Machine Learning Modules

c.kern@uni-mannheim.de

February 6 and 7, 2018

UNIVERSITY OF MANNHEIM

Outline

1. Machine Learning basics
   - Training and test error
   - Validation set, test set, CV
   - Learning curves
   - Performance measures
2. Software Resources
3. References

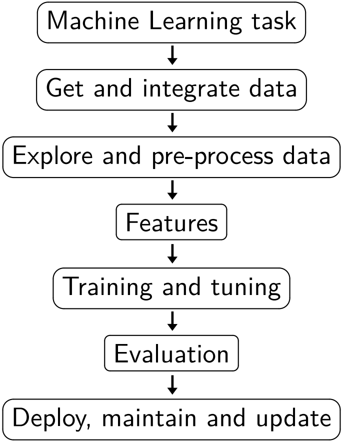
Machine Learning basics

Unsupervised Learning

- Finding patterns in data using a set of input variables X

Supervised Learning

- Predicting an output variable Y based on a set of input variables X
  1. Learn the relationship between input and output using **training data** (with X and Y)
  $$Y = f(X) + \varepsilon$$
  2. Predict the output based on the prediction model (of step 1) for **new test data** ($\sim$only X available)
- continuous Y: regression, categorical Y: classification
- Focus on **prediction** ($\neq$ causation)


Machine Learning task
Get and integrate data
Explore and pre-process data
Features
Training and tuning
Evaluation
Deploy, maintain and update

Training and test error

Training error

$$\overline{\text{err}} = \frac{1}{n} \sum\nolimits_{i=1}^{n} L(y_i, \hat{f}(x_i))$$

- Prediction error based on **training data**

Test error

$$\text{Err}_{\mathcal{T}} = \text{E}(L(Y, \hat{f}(X))|T)$$

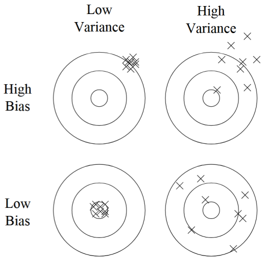
- Prediction error using **test data** (given training data T)

Expected test error decomposition

$$\text{Err}(x_0) = \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0)) + \text{Var}(\varepsilon)$$

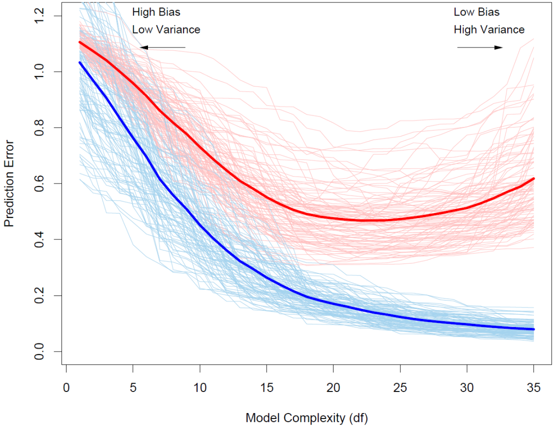
- Minimizing the expected test error
    - Low bias (deviation between $\text{E}(\hat{f}(x_0))$ and $f(x_0)$) **and**
    - Low variance ($\text{Var}(\hat{f}(x_0))$ using different training data)

Figure: Bias and Variance illustration



Domingos (2012)

Figure: Bias-Variance Trade-Off: Training error and test error by model complexity



Hastie et al. (2009)

Validation set, test set, CV

Validation set approach

- Training set & validation set
  1. Fit model using one part of training data
  2. Compute test error for the excluded section

$\rightarrow$ Model assessment

- Training set, validation set & test set
  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

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, ...

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

Standard Errors for CV

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

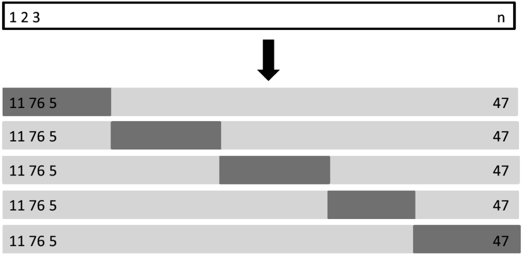
Model selection using k-Fold Cross-Validation

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

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

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



James et al. (2013)

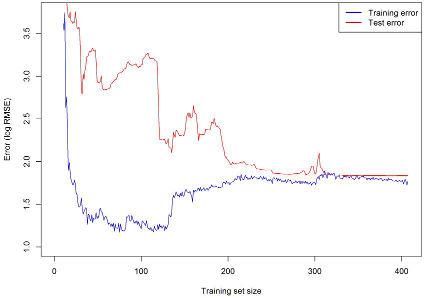
Learning curves

How much data is needed?

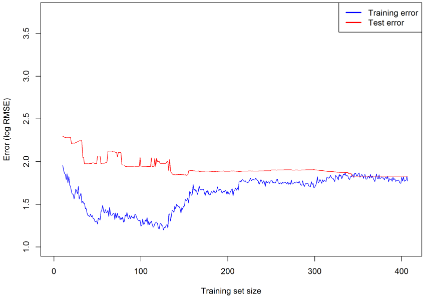
- Idea: Plot training and validation error against training set size
- Allows to study the gain of adding more data
  - Convergence of validation error curve towards training curve
- Can also be used as a diagnosis tool to asses
  - High bias (Underfitting): Curves converge at a high value
  - High variance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

Figure: Learning curves

(a) Linear regression



(b) Regression trees



Performance measures

Performance metrics for regression problems

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

Root mean squared error (RMSE):

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

Mean of absolute errors (MAE):

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

Median of absolute errors (MEDAE):

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

Probabilities, thresholds and prediction for classification

$$y_i = \begin{cases} 1 & if \quad p_i > c \\ 0 & if \quad p_i \leq c \end{cases}$$

Table: Confusion matrix

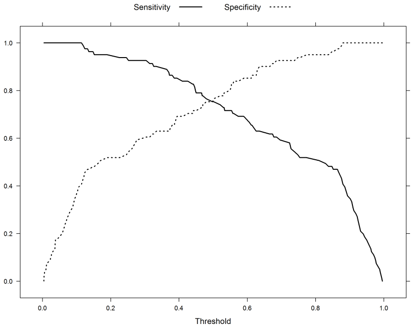
		Prediction 0	Prediction 1	
Reference	0	True Negatives (TN)	False Positives (FP)	N'
	1	False Negatives (FN)	True Positives (TP)	P'
		N	P	

Performance metrics for classification

- Global performance
  - Accuracy: $\frac{TP+TN}{TP+FP+TN+FN}$
  - Misclassification rate: $\frac{FP+FN}{TP+FP+TN+FN}$
  - No Information rate
- Row / column performance
  - Sensitivity (Recall): $\frac{TP}{TP+FN}$
  - Specificity: $\frac{TN}{TN+FP}$
  - Positive predictive value (Precision): $\frac{TP}{TP+FP}$
  - Negative predictive value: $\frac{TN}{TN+FN}$
  - False positive rate: $\frac{FP}{FP+TN}$
  - False negative rate: $\frac{FN}{FN+TP}$
- Combined measures
  - F_1: $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$
  - Cohen's κ: $1 - \frac{1-p_0}{1-p_e}$

Figure: Varying the classification threshold I

(a) Sensitivity and specificity



(b) Precision and recall

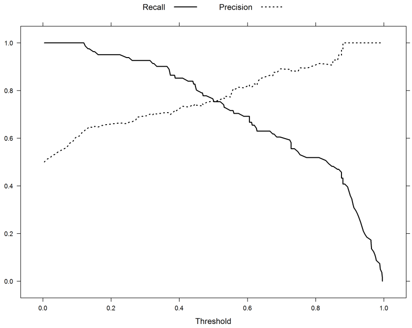
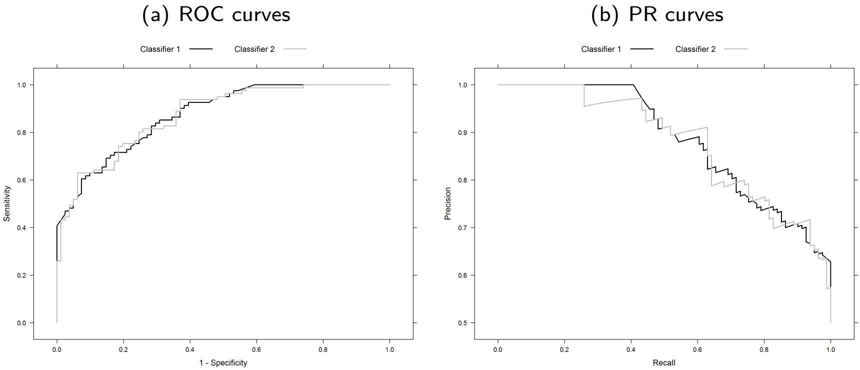


Figure: Varying the classification threshold II



→ AUC-ROC: Area under the receiver operating characteristic curve

→ AUC-PR: Area under the precision–recall curve

Software Resources

Resources for R

- Classification and Regression Training: `caret`
  - `https://topepo.github.io/caret/`
- Machine Learning in R: `mlr`
  - `https://mlr-org.github.io/mlr-tutorial/devel/html/`
- Collection of performance metrics: `MLmetrics`
- ROC and PR curves: e.g. `PRROC`

References

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

Ghani, R., Schierholz, M. (2017). Machine Learning. In: Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., Lane, J. (Eds.). (2017). *Big Data and Social Science: A Practical Guide to Methods and Tools*. Boca Raton, FL: CRC Press Taylor & Francis Group.

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.

Kuhn, M., Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer.

Mullainathan, S., Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives, 31*(2), 87–106.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.