Model Quality & Metamorphic Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

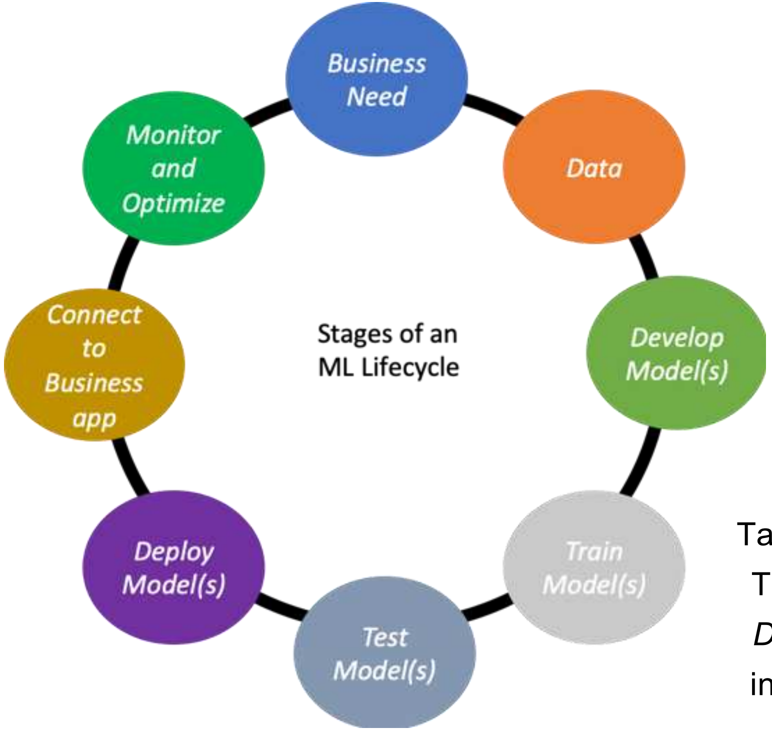
Seminar SE4AI

Outline

1. Evaluating Model Quality (Anjali Tewari)
   - Properties and Factors
   - Metrics and Measures
   - Improving MQ
2. Metamorphic Testing (Johannes Wehrstein)
   - Oracle Problem
   - Deriving Relations
   - Proving Sufficiency of MT
3. Questions & Discussion

MODEL QUALITY

Artificial Intelligence Life Cycle



Talagala, Nisha. "7 Artificial Intelligence Trends and How They Work With Operational Machine Learning." *Oracle Data Science*, blogs.oracle.com/datascience/7-artificial-intelligence-trends-and-how-they-work-with-operational-machine-learning-v2.

ML Testing Properties

TECHNISCHE UNIVERSITÄT DARMSTADT

Correctness	Model relevance	Robustness	Security
Data Privacy	Efficiency	Fairness	Interpretability

Zhang, Jie M., et al. “Machine Learning Testing: Survey, Landscapes and Horizons.” IEEE Transactions on Software Engineering, 2020, pp. 1–1

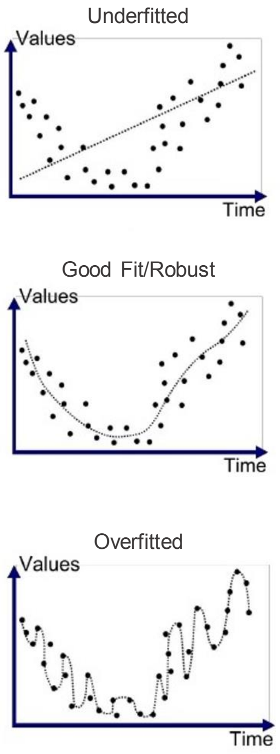
Factors that affect Model Quality

Bias:

- Due to misrepresentation in training sets
- Not enough variance in the testing sets

Outdated models: Model Quality is everchanging because data is everchanging

Overfitting/Underfitting: striking the balance between generalization and optimization



Metrics for Model Quality

Bayes Error Rate: Human Performance Rate

Depending on the type of problem, there can be:

Regression Errors

- Mean Squared Error(MSE)
- Root-Mean-Squared-Error(RMSE).
- Mean-Absolute-Error(MAE).
- R^2 or Coefficient of Determination.
- Adjusted R^2

Classification Errors

$$MSE = \frac{1}{n}\Sigma\left(y - \widehat{y}\right)^2$$

The square of the difference between actual and predicted

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

$$MAE = \frac{1}{n}\Sigma\left|y - \widehat{y}\right|$$

number of data points

Actual output value

Predicted output value

Sum of

The absolute value of the residual

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Classification Error Measures

	Actually A	Actually not A
AI predicts A	True Positive (TP)	False Positive (FP)
AI predicts not A	False Negative (FN)	True Negative (TN)

True positives and true negatives are the correct predictions
False negatives are the wrong predictions or misses
False positives are wrong predictions or false alarms

This matrix represents 2-class problems, matrices for multi-class problems have additional rows and columns for each class.

Measures for Model Quality

Successful Classifications:

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{False negative rate} = \frac{FN}{TP + FN} = 1 - \text{Recall}$$

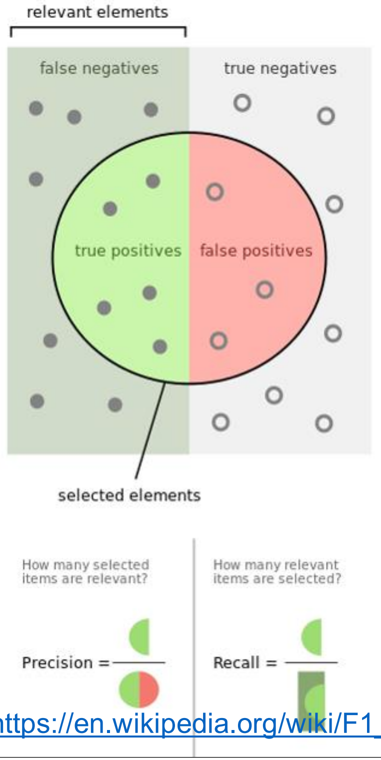
False Classifications (Noise):

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{False positve rate} = \frac{FP}{FP + TN}$$

Combined measure (harmonic mean):

$$\text{F1−Score} = 2 * \frac{\text{recall} * \text{precision}}{\text{recall} + \text{precision}}$$



Source: https://en.wikipedia.org/wiki/F1_score

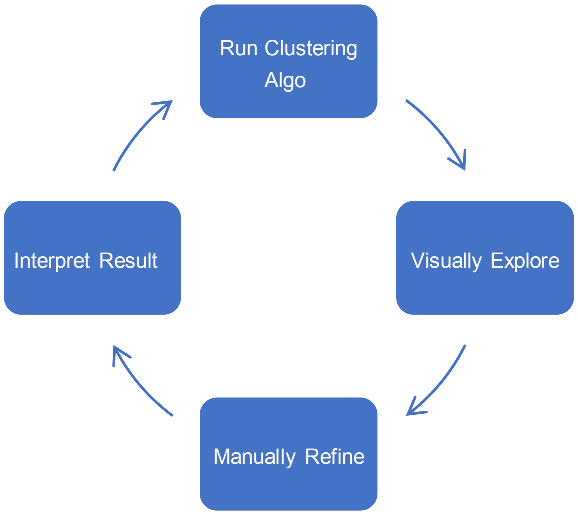
Validation through Experts

Domain expert evaluates the plausibility of a learned model

- Subjective
- Time-intensive
- Costly

But sometimes the only option (e.g. Clustering)

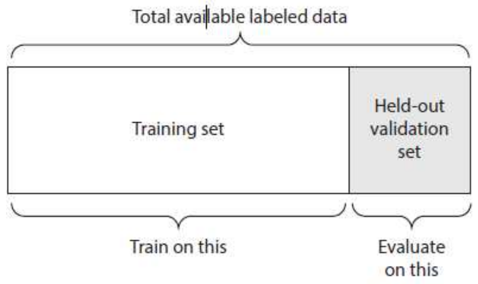
A better solution: Compare generated clusters with manually created clusters

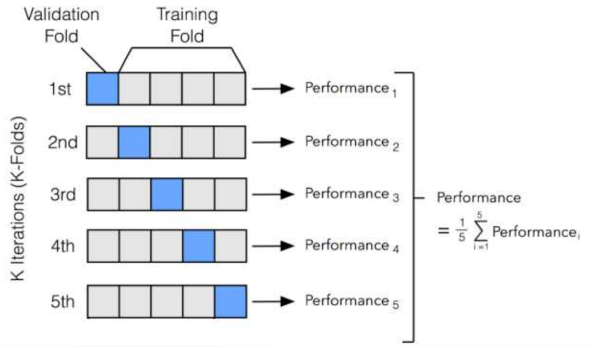


Validation on Data

TECHNISCHE UNIVERSITÄT DARMSTADT

Using	Test Set / Validation Set
Using	K-Fold Validation
Using	Iterative K-Fold Validation with Shuffling





On-line Validation

On-line validation: test learned model in a fielded application

Pro	Cons
Best estimate for overall utility	Bad model may be costly

Methods:

- Telemetry
- A/B Testing

Improving Model Quality

Avoidable bias

- Training a bigger model
- Training longer optimization models

Variance in data

- Getting more data
- Different regularization techniques
- Enlarging hyper-parameter search space

Overfitting to Validation set

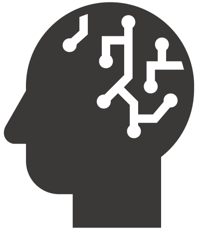
Data Mismatch

METAMORPHIC TESTING

Scenario

Assume we have following scenario:

1. ML based Service
2. Data Scarcity / No Test Oracle



Aim: Make sure that Learning Algorithm works well

Solving The Oracle Problem

ASSERTION CHECKING

N-VERSION PROGRAMMING

METAMORPHIC TESTING

Metamorphic Testing

Approach for both:

- test case generation
- ✓ test result verification

Originally proposed for generating new test cases based on successful ones (Chen et al, 1998)

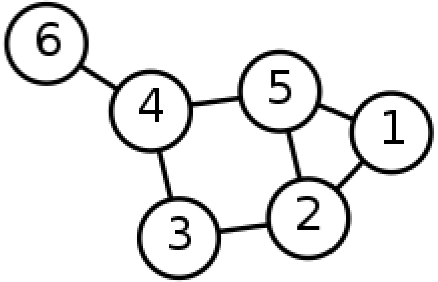
Central element: Metamorphic Relations (MRs)

Metamorphic Testing: A New Approach for Generating Next Test Cases (Chen et al, 1998)

Example Relations for Shortest Path in Graph

Program: $\mathrm{P}(G, a, b)$ (computes shortest path between vertices a and b in undirected graph G)

Proving that result is really the shortest path: difficult



Metamorphic Relations

$$|P(G, b, a)| = |P(G, a, b)|$$

$$|P(G, a, b)| + |P(G, b, c)| \geq |P(G, a, c)|$$

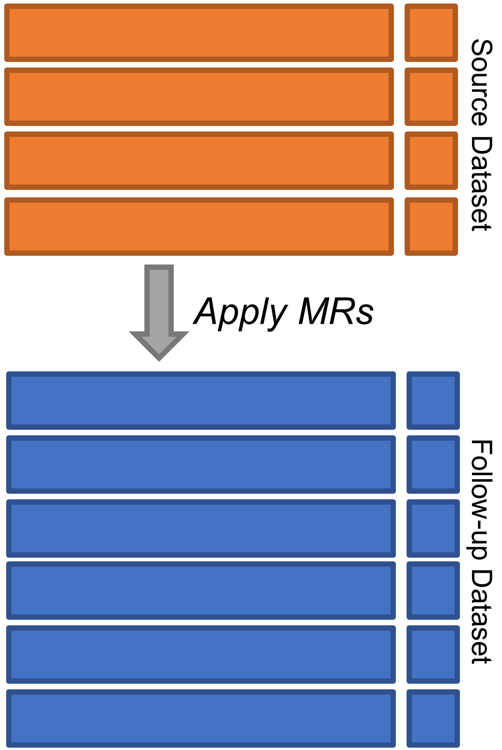
Metamorphic Relations (MRs)

f: function / algorithm
X: Input space
Y: Output space

$$\mathcal{R} \subseteq X^n \times Y^n, n \geq 2$$

$$R(x_1, x_2, \ldots, x_n, f(x_1), f(x_2), \ldots, f(x_n))$$

Caveat:
- MRs = Relations between Testcases ($n \geq 2$), not between Inputs & Outputs (→ Assertion Testing)



Metamorphic Testing Process

Deriving Metamorphic Relations

Derive from problem

Derive from learning algorithm

Deriving MRs from learning algorithm

TECHNISCHE UNIVERSITÄT DARMSTADT

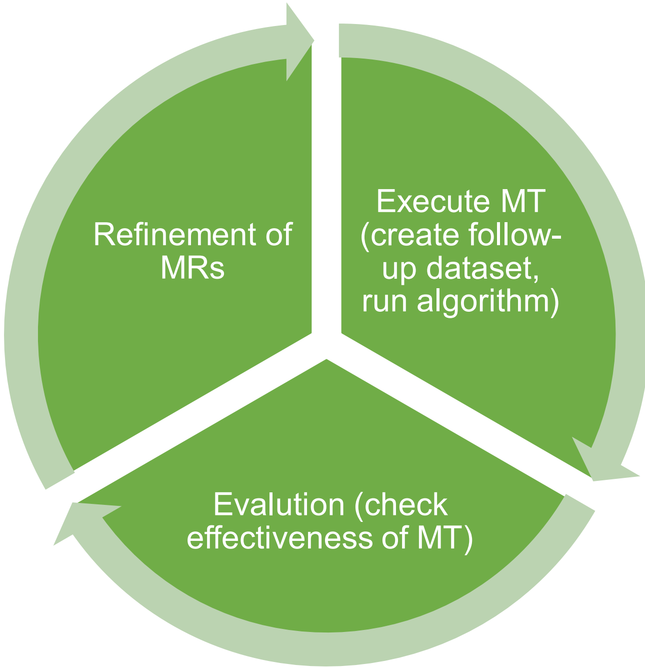
1. Consistence with affine transformation
2. Permutation of class labels / attributes
3. Addition of uninformative attributes
4. Consistence with re-prediction
5. Removal of classes

…

→ MRs are independent from underlying problem

Testing and Validating Machine Learning Classifiers by Metamorphic Testing: Xie et al (2009)

Metamorphic Testing



Proving Sufficiency of MT

- Evaluate testing with test coverage (→ mostly impossible for ML)
- Mutant Testing
- Mutated Tests

MT: Advantages / Disadvantages

Advantages

- Simplicity in concept
- Straightforward implementation
- Ease of automation
- Low costs

Disadvantages

- Difficult generation of MR
- Requires „fast“ learning algorithms
- Difficulty dealing with indeterminism

Sources

Ding, Junhua, et al. “A Framework for Ensuring the Quality of a Big Data Service.” 2016 IEEE International Conference on Services Computing (SCC), 2016, pp. 82–89.

Segura, Sergio, et al. “A Survey on Metamorphic Testing.” IEEE Transactions on Software Engineering, vol. 42, no. 9, 2016, pp. 805–824.

Liu, Huai, et al. “How Effectively Does Metamorphic Testing Alleviate the Oracle Problem.” IEEE Transactions on Software Engineering, vol. 40, no. 1, 2014, pp. 4–22.

Chen, Tsong Yueh, et al. “Metamorphic Testing: A Review of Challenges and Opportunities.” ACM Computing Surveys, vol. 51, no. 1, 2018, pp. 1–27.

Zhou, Zhi Quan, et al. “Metamorphic Testing for Software Quality Assessment: A Study of Search Engines.” IEEE Transactions on Software Engineering, vol. 42, no. 3, 2016, pp. 264–284.

Chen, T. Y., et al. “Metamorphic Testing: A New Approach for Generating Next Test Cases.” ArXiv Preprint ArXiv:2002.12543, 2020.

Zhang, Jie M., et al. “Machine Learning Testing: Survey, Landscapes and Horizons.” IEEE Transactions on Software Engineering, 2020, pp. 1–1.

Chen, Jing, et al. “A Metamorphic Testing Approach for Event Sequences.” PLOS ONE, vol. 14, no. 2, 2019.

Barr, Earl T., et al. “The Oracle Problem in Software Testing: A Survey.” IEEE Transactions on Software Engineering, vol. 41, no. 5, 2015, pp. 507–525.

Khokhar, Muhammad Nadeem, et al. “Metamorphic Testing of AI-Based Applications: A Critical Review.” International Journal of Advanced Computer Science and Applications, vol. 11, no. 4, 2020.

Roman, Victor. *How To Develop a Machine Learning Model From Scratch*. 2 Apr. 2019, towardsdatascience.com/machine-learning-general-process-8f1b510bd8af.

Mello, Arthur. “How Can You Improve Your Machine Learning Model Quality?” *Medium*, Towards Data Science, 2 Apr. 2020, towardsdatascience.com/how-can-you-improve-your-machine-learning-model-quality-b22737d4fe5f.

Fukunaga, Keinosuke Introduction to Statistical Pattern Recognition by ISBN 0122698517, 1990, pp 3 and 97

Kaestner, Christian. “Model Quality.” *17-445: Model Quality*, ckaestne.github.io/seai/F2019/slides/08_model_quality/modelquality.html.

Mishra, Divyanshu. “Regression: An Explanation of Regression Metrics And What Can Go Wrong.” *Medium*, Towards Data Science, 6 Dec. 2019, towardsdatascience.com/regression-an-explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914.

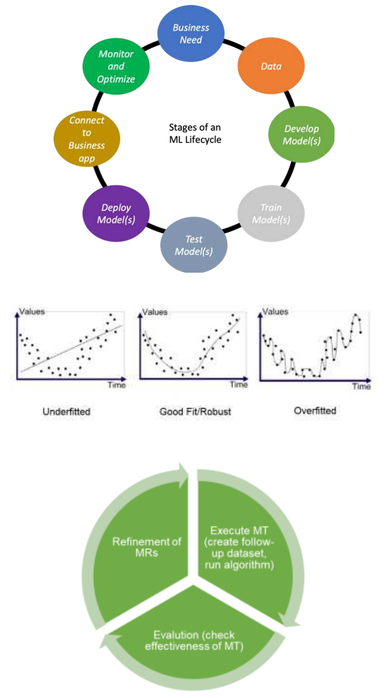
Kohavi, Ron & Longbotham, Roger. (2017). Online Controlled Experiments and A/B Testing. 10.1007/978-1-4899-7687-1_891.

Hand, David, and Peter Christen. “A Note on Using the F-Measure for Evaluating Record Linkage Algorithms.” *Statistics and Computing*, vol. 28, no. 3, 2017, pp. 539–547., doi:10.1007/s11222-017-9746-6.

Perlin, Michael. Quality Assurance for Artificial Intelligence (Part 2). Medium. 09/03/2020.

Your chance to get more…

QUESTIONS



Discussion

- On which kind of ML algorithms Metamorphic Testing is applicable?

Acknowledgements & License

TECHNISCHE UNIVERSITÄT DARMSTADT

- Images are either by the authors of these slides, attributed where they are used, or their source be found under the "Sources" Section.