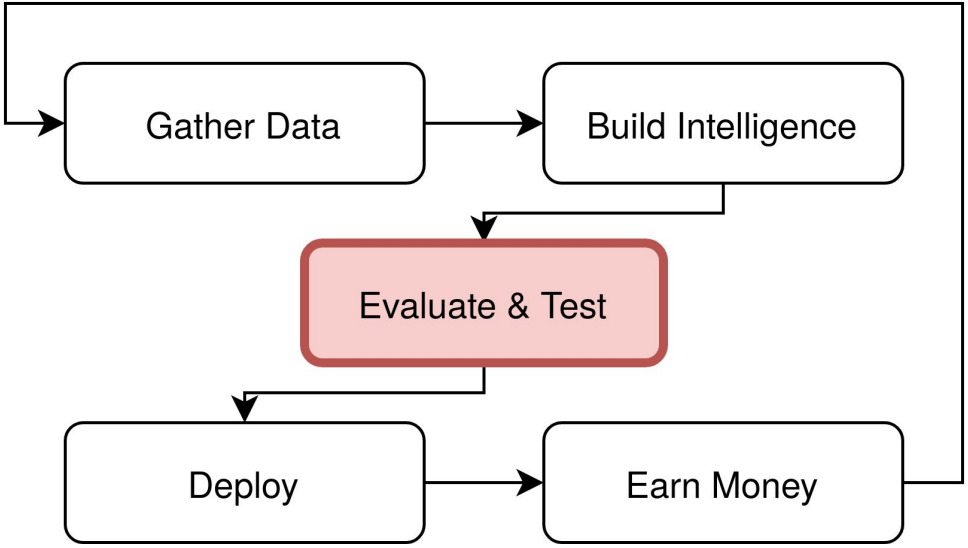
Model Quality and Metamorphic Testing

A talk by Steffen Kuchelmeister and Stephan Wezorke

TECHNISCHE UNIVERSITÄT DARMSTADT



Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

- Model Quality:
  - Ensuring your Model can work
  - Online Testing
  - Offline Testing
  - ROC and AUC
  - F1 Score
- Metamorphic Testing
  - Necessity of Metamorphic Testing
  - Examples
  - Conclusion
- What You Should Take Home
- Literature

Model Quality

https://xkcd.com/1838/

How do you know when the answers start looking right?

Ensuring your Model can work

TECHNISCHE UNIVERSITÄT DARMSTADT

- Standard SE Methods still apply
  - Code review
  - Readable code is especially important
- Test everything you can
  - Input and Output pipeline
  - Use sanity checks
- Always test some examples per Hand

How to evaluate your Model

TECHNISCHE UNIVERSITÄT DARMSTADT

Online Testing:

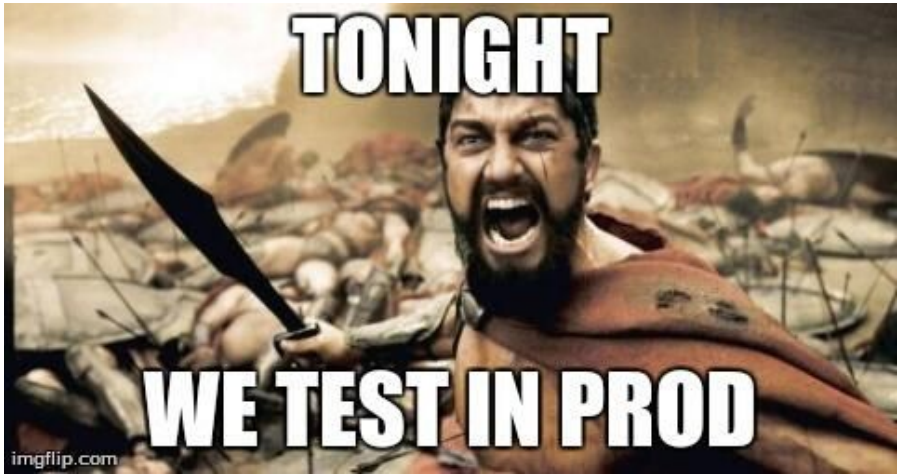
- Deploy your Model
- Measure how the users interact
- Use model with the best user interaction

Offline Testing:

- Have a designated test set
- Evaluate your models on this set
- Use the best performing Models in Production

Online Testing

- Make sure the model isn’t a complete failure
- Only test a small part of your user base
- Make sure you are not excluding some groups of users
- Metrics are dependant on your service



Example Metrics for Online Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

Video Recommendation System:

- How often are users interacting with video recommandations
- How long do they use the application after interacting
- How are recommended Videos rated

Medical Imaging Assistant:

- How often agrees the doctor with the System
- Why did the doctor disagree?
- How often are there misdiagnosis,
  - when the system was right
  - when the system was wrong

For more examples look at [1]

Offline Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Have a designated test set
- Make sure the test data is independent:
  - Separate by Time
- Make sure the set is representative
- Only use the test set after finalizing models
- For example train on data from the 1st-4th of the month and test with data from the 5th

Metrics

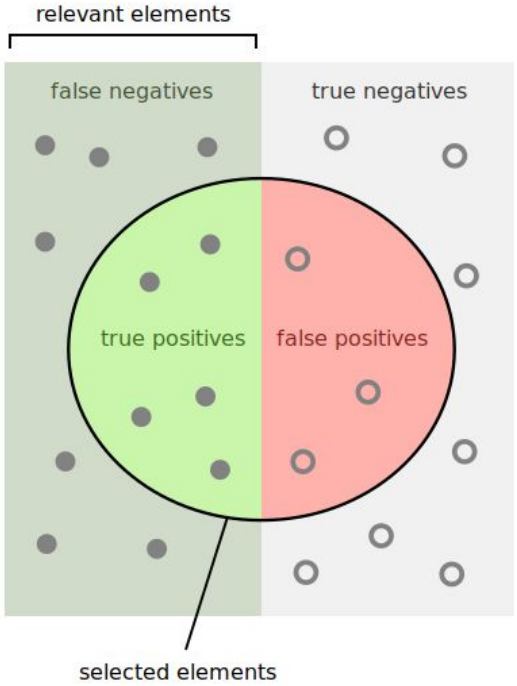
TECHNISCHE UNIVERSITÄT DARMSTADT

Dependant on task:

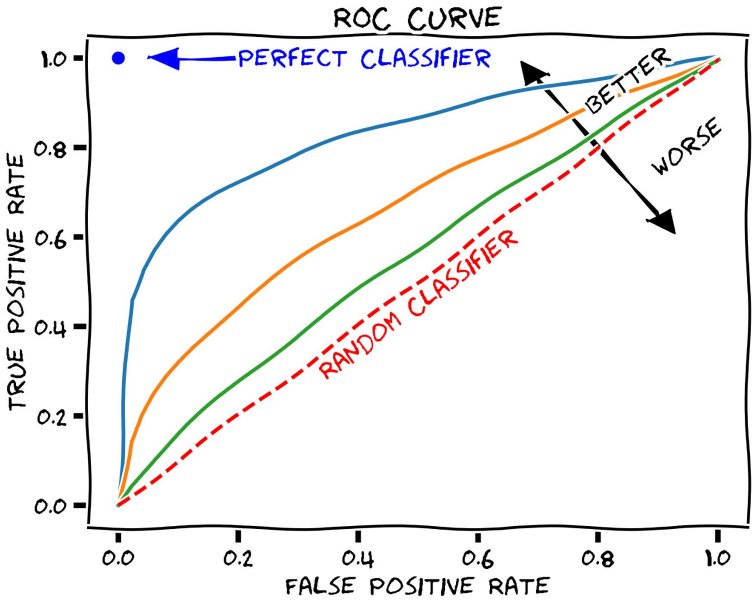
- Classification -> Accuracy, Area under curve, F1
- Regression -> Mean Square Error(be careful with outliers)
- But sometimes more complicated:
    - How to measure the quality of translation
    - How about Image Generation
    - How to compare Graphs or Trees

Always look at the literature for a given task!

Area under Curve(AUC)



https://commons.wikimedia.org/wiki/File:Precisionrecall.svg



https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg

You can change the decision boundary to manipulate the FPR or the TPR

Integrate the ROC to get the AUC

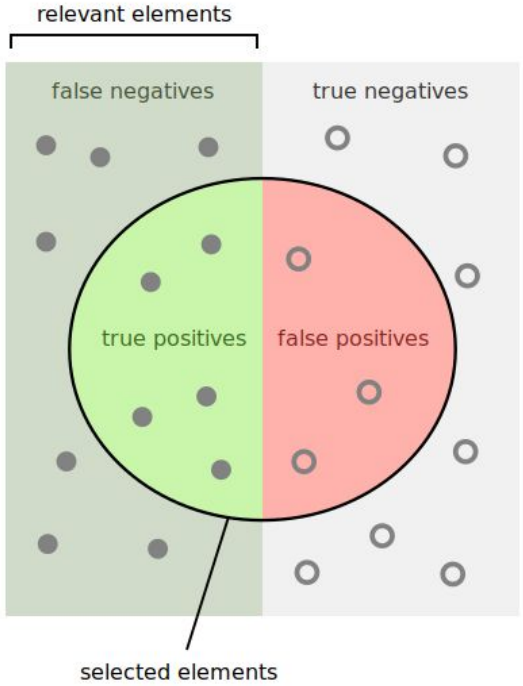
Problems with AUC

TECHNISCHE UNIVERSITÄT DARMSTADT

- Needs shiftable decision boundary
- Some classifiers predict almost every point at 100% or 0%, so it’s not possible to choose a decision boundary
- Looks at all decision boundaries even if only one is relevant
- Fails if the classes are heavily unbalanced
- Looks at region of the ROC, which are often irrelevant

For further discussion look at [7]

F1-Score



https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Improving your Model

TECHNISCHE UNIVERSITÄT DARMSTADT

Use a better Method:

- Compare different types of Models

Use more or better data:

- Collect more data e.g. from production
- Clean your data
- Add data cases where the predictions aren’t up to standard

TECHNISCHE UNIVERSITÄT DARMSTADT

Metamorphic Testing

> *“Program testing can be a very effective way to show the presence of bugs, but it is hopelessly inadequate for showing their absence.*
>
> Edsger W. Dijkstra

Necessity of Metamorphic Testing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Definition:

> A metamorphic relation (MR) is a relation between outputs of a program, when the inputs follow a certain transformation rule.

- Simple examples: $\exp(x) < \exp(x+1); \quad \sin(x+2\pi) = \sin(x)$
- Why do we need MT?
  - Telemetry based on live user experience
  - Oracle problem spoils conventional tests [2]

    → MT one possible solution

MT - Examples

TECHNISCHE UNIVERSITÄT DARMSTADT

- Revisit problem of playing golf
  - [rainy, hot, humid, not windy] →NPG
  - [overcast, cool, dry, windy] → PG
  - [suny, medium, humid, windy] → NPG
- Use numerical predictors, e.g. {rainy, overcast, sunny} → {-1, 0, 1}
  - [-1, 1, 1, 0] → NPG
  - [0, -1, 0, 1] → PG
  - [1, 0, 1, 1] → NPG

MT - Examples (continued)

- k-nearest neighbors (k-NN) predicts based on learned data points that are "closest" to input; distance usually defined by Euclidean metric

$$d(x, x') = \sqrt{\sum_j (x_j - x'_j)^2}$$

- MR: Affine Transformation [3], $x_j \to kx_j + b, k \neq 0$
  - k-NN invariant, because metric only scaled
  - Naive Bayes also not affected, but harder to prove

MT - Examples (continued)

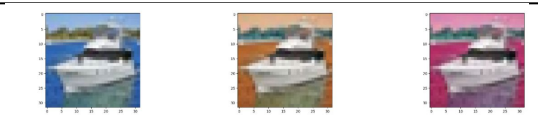
- Re-labelling samples [3]:
  - Relabel subset of training data labeled **NPG** as **NPG***
  - k-NN: test samples previously labeled **PG** not affected
  - NB: not a necessary property

-1	1	1	1	NPG
0	0	1	1	PG
0	1	1	0	PG
-1	1	1	0	NPG
0	1	0	0	PG

-1	1	1	1	NPG
0	0	1	1	PG
0	1	1	0	PG
-1	1	1	0	NPG*
0	1	0	0	PG

MT - Examples (continued)

- Permutation of RGB channels
  in image classification [4]
- Scale mitochondrion by a few percent
  → number of mitochondria persists [5]
- Artificial weather change should not
  change the driving direction
  of autonomous driving software [6]



(a) Original training data (RGB) (b) BGR (c) BRG

(d) GBR (e) GRB (f) RBG

DOI: 10.1145/3213846.3213858

DOI: 10.1145/3238147.3238187

MT - Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- MRs that provably apply may be used to test the algorithms implementation (independent of data)
  - The complexer the algorithm, the harder to find or prove MR
- Otherwise, MRs only validate, that user expectations are met
  - Failure just indicates problem as ML algorithms are not perfect
- Engineering domain specific MRs is difficult task
  - How to properly modify input data and validate modification?
  - E.g.: Is the artificial weather change reasonable?

What You Should Take Home

TECHNISCHE UNIVERSITÄT DARMSTADT

- Use offline and online testing
- Test entire system (not just model)
- Consult the Literature for what metrics to use
- MRs can be used to find errors in implementation or validate user experience
- Finding proper MRs is a difficult (engineering) task

Literature

[1] Chapter 15 of G. Hulten, "*Building Intelligent Systems: A Guide to Machine Learning Engineering*". Apress, 2018. DOI: https://doi.org/10.1007/978-1-4842-3432-7

[2] E. J. Weyuker, "*On Testing Non-Testable Programs*", The Computer Journal 25.4 (1982), 465. DOI: 0.1093/comjnl/25.4.465

[3] X. Xie et al. "*Testing and validating machine learning classifiers by metamorphic testing*", The Journal of Systems and Software 84.4 (2011), 544. DOI: 10.1016/j.jss.2010.11.920

Literature (continued)

[4] A. Dwarakanath et al., "*Identifying implementation bugs in machine learning based image classifiers using metamorphic testing*". In: "*Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis - ISSTA 2018*". ACM Press, 2018, 118. DOI: 10.1145/3213846.3213858

[5] J. Ding, D. Zhang, and X.-H. Hu, "*A Framework for Ensuring the Quality of a Big Data Service*". In: "*2016 IEEE International Conference on Services Computing (SCC)*", 2016, 82. DOI: 10.1109/SCC.2016.18

Literature (continued)

[6] M. Zhang et al. “*DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems*”. In: “*Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*”. ASE, 2018, 132. DOI: 10.1145/3238147.3238187

[7] Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models." *Global ecology and Biogeography* 17.2 (2008): 145-151.