MODEL QUALITY

Christian Kaestner

Required reading:

LEARNING GOALS

- Select a suitable metric to evaluate prediction accuracy of a model and to compare multiple models
- Select a suitable baseline when evaluating model accuracy
- Explain how software testing differs from measuring prediction accuracy of a model
- Curate validation datasets for assessing model quality, covering subpopulations as needed
- Use invariants to check partial model properties with automated testing
- Develop automated infrastructure to evaluate and monitor model quality
THIS LECTURE

FIRST PART: MEASURING PREDICTION ACCURACY

the data scientist's perspective

SECOND PART: LEARNING FROM SOFTWARE TESTING

how software engineering tools may apply to ML
"Programs which were written in order to determine the answer in the first place. There would be no need to write such programs, if the correct answer were known" (Weyuker, 1982).
PREDICTION ACCURACY OF A MODEL

\[ \text{model}: X \rightarrow Y \]

validation data (tests?): sets of \((X, Y)\) pairs indicating desired outcomes for select inputs

For our discussion: any form of model, including machine learning models, symbolic AI components, hardcoded heuristics, composed models, ...
Today's focus is on the quality of the produced *model*, not the algorithm used to learn the model or the data used to train the model:

i.e. assuming *Decision Tree Algorithm* and feature extraction are correctly implemented (according to specification), is the model learned from data any good?

The model is just one component of the entire system.

Focus on measuring quality, not debugging the source of quality problems (e.g., in data, in feature extraction, in learning, in infrastructure)
CASE STUDY: CANCER DETECTION
Application to be used in hospitals to screen for cancer, both as routine preventative measure and in cases of specific suspicions. Supposed to work together with physicians, not replace.
THE SYSTEMS PERSPECTIVE

System is more than the model

Includes deployment, infrastructure, user interface, data infrastructure, payment services, and often much more

Systems have a goal:

- maximize sales
- save lives
- entertainment
- connect people

Models can help or may be essential in those goals, but are only one part

Today: Narrow focus on prediction accuracy of the model
CANCER PREDICTION WITHIN A HEALTHCARE APPLICATION

(CC BY-SA 4.0, Martin Sauter)
MANY QUALITIES

Prediction accuracy of a model is important

But many other quality matters when building a system:

- Model size
- Inference time
- User interaction model
- Kinds of mistakes made
- How the system deals with mistakes
- Ability to incrementally learn
- Safety, security, fairness, privacy
- Explainability

Today: Narrow focus on prediction accuracy of the model
COMPARING MODELS

Compare two models (same or different implementation/learning technology) for the same task:

- Which one supports the system goals better?
- Which one makes fewer important mistakes?
- Which one is easier to operate?
- Which one is better overall?
- Is either one good enough?
ON TERMINOLOGY: PERFORMANCE

In machine learning, "performance" typically refers to accuracy

"this model performs better" = it produces more accurate results

Be aware of ambiguity across communities.

When speaking of "time", be explicit: "learning time", "inference time", "latency",
...

(see also: performance in arts, job performance, company performance, performance test (bar exam) in law, software/hardware/network performance)
MEASURING PREDICTION ACCURACY FOR CLASSIFICATION TASKS

(The Data Scientists Toolbox)
**CONFUSION/ERROR MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>Actually A</th>
<th>Actually B</th>
<th>Actually C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI predicts A</td>
<td>10</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>AI predicts B</td>
<td>3</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>AI predicts C</td>
<td>5</td>
<td>22</td>
<td>82</td>
</tr>
</tbody>
</table>

Accuracy = correct predictions (diagonal) out of all predictions

Example's accuracy = \[\frac{10 + 24 + 82}{10 + 6 + 2 + 3 + 24 + 10 + 5 + 22 + 82} = .707\]
IS 99% ACCURACY GOOD?

-> depends on problem; can be excellent, good, mediocre, terrible

10% accuracy can be good on some tasks (information retrieval)

Always compare to a base rate!

Reduction in error = \[
\frac{(1 - accuracy_{baseline}) - (1 - accuracy_f)}{1 - accuracy_{baseline}}
\]

- from 99.9% to 99.99% accuracy = 90% reduction in error
- from 50% to 75% accuracy = 50% reduction in error
BASELINES?

Suitable baselines for cancer prediction? For recidivism?
Many forms of baseline possible, many obvious: Random, all true, all false, repeat last observation, simple heuristics, simpler model
TYPES OF MISTAKES

Two-class problem of predicting event A:

<table>
<thead>
<tr>
<th></th>
<th>Actually A</th>
<th>Actually not A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI predicts A</strong></td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td><strong>AI predicts not A</strong></td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

True positives and true negatives: correct prediction

False negatives: wrong prediction, miss, Type II error

False positives: wrong prediction, false alarm, Type I error
### MULTI-CLASS PROBLEMS VS TWO-CLASS PROBLEM

<table>
<thead>
<tr>
<th></th>
<th>Actually A</th>
<th>Actually B</th>
<th>Actually C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI predicts A</td>
<td>10</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>AI predicts B</td>
<td>3</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>AI predicts C</td>
<td>5</td>
<td>22</td>
<td>82</td>
</tr>
</tbody>
</table>
### MULTI-CLASS PROBLEMS VS TWO-CLASS PROBLEM

<table>
<thead>
<tr>
<th></th>
<th>Actually A</th>
<th>Actually B</th>
<th>Actually C</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI predicts A</td>
<td>10</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>AI predicts B</td>
<td>3</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>AI predicts C</td>
<td>5</td>
<td>22</td>
<td>82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AI predicts A</td>
<td>10</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI predicts not A</td>
<td>8</td>
<td>138</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI predicts B</td>
<td></td>
<td></td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>AI predicts not B</td>
<td></td>
<td></td>
<td>28</td>
<td>99</td>
</tr>
</tbody>
</table>
Individual false positive/negative classifications can be derived by focusing on a single value in a confusion matrix. False positives/recall/etc are always considered with regard to a single specific outcome.
TYPES OF MISTAKES IN IDENTIFYING CANCER?
**MEASURES**

Measuring success of correct classifications (or missing results):

- Recall = $\frac{TP}{TP+FN}$
  - aka true positive rate, hit rate, sensitivity; *higher is better*
- False negative rate = $\frac{FN}{TP+FN} = 1 - \text{recall}$
  - aka miss rate; *lower is better*

Measuring rate of false classifications (or noise):

- Precision = $\frac{TP}{TP+FP}$
  - aka positive predictive value; *higher is better*
- False positive rate = $\frac{FP}{FP+TN}$
  - aka fall-out; *lower is better*

Combined measure (harmonic mean):

$$F1 \text{ score} = 2 \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$
(CC BY-SA 4.0 by Walber)
FALSE POSITIVES AND FALSE NEGATIVES EQUALLY BAD?

Consider:

- Recognizing cancer
- Suggesting products to buy on e-commerce site
- Identifying human trafficking at the border
- Predicting high demand for ride sharing services
- Predicting recidivism chance
- Approving loan applications

No answer vs wrong answer?
EXTREME CLASSIFIERS

• Identifies every instance as negative (e.g., no cancer):
  ▪ 0% recall (finds none of the cancer cases)
  ▪ 100% false negative rate (misses all actual cancer cases)
  ▪ undefined precision (no false predictions, but no predictions at all)
  ▪ 0% false positive rate (never reports false cancer warnings)

• Identifies every instance as positive (e.g., has cancer):
  ▪ 100% recall (finds all instances of cancer)
  ▪ 0% false negative rate (does not miss any cancer cases)
  ▪ low precision (also reports cancer for all noncancer cases)
  ▪ 100% false positive rate (all noncancer cases reported as warnings)
CONSIDER THE BASELINE PROBABILITY

Predicting unlikely events -- 1 in 2000 has cancer (stats)

<table>
<thead>
<tr>
<th></th>
<th>Cancer</th>
<th>No c.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random predictor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer pred.</td>
<td>3</td>
<td>4998</td>
</tr>
<tr>
<td>No cancer pred.</td>
<td>2</td>
<td>4997</td>
</tr>
<tr>
<td><strong>Never cancer predictor</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer pred.</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No cancer pred.</td>
<td>5</td>
<td>9995</td>
</tr>
</tbody>
</table>

.5 accuracy, .6 recall, 0.001 precision

See also Bayesian statistics
Many classification models produce a number (e.g., "chance of cancer"), need *threshold* to make decision

<table>
<thead>
<tr>
<th></th>
<th>Act. A</th>
<th>Act. not A</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI predicts A</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>AI predicts not A</td>
<td>8</td>
<td>138</td>
</tr>
</tbody>
</table>

Thresholds affects how data is sorted into rows!
AREA UNDER THE CURVE

Turning numeric prediction into classification with threshold ("operating point")
The plot shows the recall precision/tradeoff at different thresholds (the thresholds are not shown explicitly). Curves closer to the top-right corner are better considering all possible thresholds. Typically, the area under the curve is measured to have a single number for comparison.
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES
Same concept, but plotting TPR (recall) against FPR rather than precision. Graphs closer to the top-left corner are better. Again, the area under the (ROC) curve can be measured to get a single number for comparison.
MORE ACCURACY MEASURES FOR CLASSIFICATION PROBLEMS

- Lift
- Break even point
- F1 measure, etc
- Log loss (for class probabilities)
- Cohen's kappa, Gini coefficient (improvement over random)
MEASURING PREDICTION ACCURACY FOR REGRESSION AND RANKING TASKS

(The Data Scientists Toolbox)
## Confusion Matrix for Regression Tasks?

<table>
<thead>
<tr>
<th>Rooms</th>
<th>Crime Rate</th>
<th>...</th>
<th>Predicted Price</th>
<th>Actual Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.01</td>
<td>...</td>
<td>230k</td>
<td>250k</td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>...</td>
<td>530k</td>
<td>498k</td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>...</td>
<td>210k</td>
<td>211k</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>...</td>
<td>219k</td>
<td>210k</td>
</tr>
</tbody>
</table>
Confusion Matrix does not work, need a different way of measuring accuracy that can distinguish "pretty good" from "far off" predictions
REGRESSION TO CLASSIFICATION

<table>
<thead>
<tr>
<th>Rooms</th>
<th>Crime Rate</th>
<th>...</th>
<th>Predicted Price</th>
<th>Actual Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.01</td>
<td>...</td>
<td>230k</td>
<td>250k</td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>...</td>
<td>530k</td>
<td>498k</td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>...</td>
<td>210k</td>
<td>211k</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>...</td>
<td>219k</td>
<td>210k</td>
</tr>
</tbody>
</table>

Was the price below 300k?

Which price range is it in: [0-100k], [100k-200k], [200k-300k], ...
### COMPARING PREDICTED AND EXPECTED OUTCOMES

Mean Absolute Percentage Error

\[
\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|
\]

\((A_t \text{ actual outcome, } F_t \text{ predicted outcome, for row } t)\)

Compute relative prediction error per row, average over all rows

<table>
<thead>
<tr>
<th>Rooms</th>
<th>Crime Rate</th>
<th>Predicted Price</th>
<th>Actual Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.01</td>
<td>230k</td>
<td>250k</td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>530k</td>
<td>498k</td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>210k</td>
<td>211k</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>219k</td>
<td>210k</td>
</tr>
</tbody>
</table>

\[
\text{MAPE} = \frac{1}{4} \left( \frac{20}{250} + \frac{32}{498} + \frac{1}{211} + \frac{9}{210} \right) = \frac{1}{4} \left( 0.08 + 0.064 + 0.005 + 0.043 \right) = 0.048
\]
AGAIN: COMPARE AGAINST BASELINES

Accuracy measures in isolation are difficult to interpret

Report baseline results, reduction in error
BASELINES FOR REGRESSION PROBLEMS

Baselines for house price prediction?
OTHER MEASURES FOR REGRESSION MODELS

- Mean Absolute Error (MAE) = \( \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t| \)
- Mean Squared Error (MSE) = \( \frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2 \)
- Root Mean Square Error (RMSE) = \( \sqrt{\frac{\sum_{t=1}^{n} (A_t - F_t)^2}{n}} \)
- \( R^2 = \) percentage of variance explained by model
- ...
EVALUATING RANKINGS

Ordered list of results, true results should be ranked high

Common in information retrieval (e.g., search engines) and recommendations

<table>
<thead>
<tr>
<th>Rank</th>
<th>Product</th>
<th>Correct?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Juggling clubs</td>
<td>true</td>
</tr>
<tr>
<td>2</td>
<td>Bowling pins</td>
<td>false</td>
</tr>
<tr>
<td>3</td>
<td>Juggling balls</td>
<td>false</td>
</tr>
<tr>
<td>4</td>
<td>Board games</td>
<td>true</td>
</tr>
<tr>
<td>5</td>
<td>Wine</td>
<td>false</td>
</tr>
<tr>
<td>6</td>
<td>Audiobook</td>
<td>true</td>
</tr>
</tbody>
</table>

MAP@K = precision in first $K$ results

Averaged over many queries

MAP@1 = 1, MAP@2 = 0.5, MAP@3 = 0.33, ...

Remember to compare against baselines! Baseline for shopping recommendations?
OTHER RANKING MEASURES

- Mean Reciprocal Rank (MRR) (average rank for first correct prediction)
- Average precision (concentration of results in highest ranked predictions)
- MAR@K (recall)
- Coverage (percentage of items ever recommended)
- Personalization (how similar predictions are for different users/queries)
- Discounted cumulative gain
- ...

7.9
Good discussion of tradeoffs at https://medium.com/swlh/rank-aware-recons-evaluation-metrics-5191bba16832
Highly problem dependent:

- Classify text into positive or negative -> classification problem
- Determine truth of a statement -> classification problem
- Translation and summarization -> comparing sequences (e.g. ngrams) to human results with specialized metrics, e.g., BLEU and ROUGE
- Modeling text -> how well its probabilities match actual text, e.g., likelihood or perplexity
ANALOGY TO SOFTWARE TESTING

(this gets messy)
SOFTWARE TESTING

- Program $p$ with specification $s$
- Test consists of
  - Controlled environment
  - Test call, test inputs
  - Expected behavior/output (oracle)

```java
assertEquals(4, add(2, 2));
assertEquals(??, factorPrime(15485863));
```

Testing is complete but unsound: Cannot guarantee the absence of bugs
Software testing can be applied to many qualities:

- Functional errors
- Performance errors
- Buffer overflows
- Usability errors
- Robustness errors
- Hardware errors
- API usage errors

"Testing shows the presence, not the absence of bugs" -- Edsger W. Dijkstra 1969
<table>
<thead>
<tr>
<th>Rooms</th>
<th>Crime Rate</th>
<th>...</th>
<th>Actual Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.01</td>
<td>...</td>
<td>250k</td>
</tr>
<tr>
<td>4</td>
<td>.01</td>
<td>...</td>
<td>498k</td>
</tr>
<tr>
<td>2</td>
<td>.03</td>
<td>...</td>
<td>211k</td>
</tr>
<tr>
<td>2</td>
<td>.02</td>
<td>...</td>
<td>210k</td>
</tr>
</tbody>
</table>

Fail the entire test suite for one wrong prediction?
THE ORACLE PROBLEM

*How do we know the expected output of a test?*

- Manually construct input-output pairs (does not scale, cannot automate)
- Comparison against gold standard (e.g., alternative implementation, executable specification)
- Checking of global properties only -- crashes, buffer overflows, code injections
- Manually written assertions -- partial specifications checked at runtime

```java
assertEquals(??, factorPrime(15485863));
```
AUTOMATED TESTING / TEST CASE GENERATION

- Many techniques to generate test cases
- Dumb fuzzing: generate random inputs
- Smart fuzzing (e.g., symbolic execution, coverage guided fuzzing): generate inputs to maximally cover the implementation
- Program analysis to understand the shape of inputs, learning from existing tests
- Minimizing redundant tests
- Abstracting/simulating/mocking the environment
- Typically looking for crashing bugs or assertion violations
IS LABELED VALIDATION DATA SOLVING THE ORACLE PROBLEM?

```python
assertEquals(250000, model.predict([3, .01, ...]));
assertEquals(498000, model.predict([4, .01, ...]));
```
DIFFERENT EXPECTATIONS FOR PREDICTION ACCURACY

- Not expecting that all predictions will be correct (80% accuracy may be very good)
- Data may be mislabeled in training or validation set
- There may not even be enough context (features) to distinguish all training outcomes

- Lack of specifications
- A wrong prediction is not necessarily a bug
ANALOGY OF PERFORMANCE TESTING?

- Performance tests are not precise (measurement noise)
  - Averaging over repeated executions of the same test
  - Commonly using diverse benchmarks, i.e., multiple inputs
  - Need to control environment (hardware)
- No precise specification
  - Regression tests
  - Benchmarking as open-ended comparison
  - Tracking results over time

@Test(timeout=100)
public void testCompute() {
    expensiveComputation(...);
}
MACHINE LEARNING IS REQUIREMENTS ENGINEERING

(my pet theory)

see also https://medium.com/@ckaestne/machine-learning-is-requirements-engineering-8957ae55ef4
VALIDATION VS VERIFICATION

Diagram showing the relationship between Problem and Needs, Requirements Engineering, Validation, Verification, Requirements/Specification, Coding, and Implementation.
VALIDATION VS VERIFICATION

- Validation:
  - Interviews, Req. Synthesis
  - Specification → More Specifications
  - Coding → Implementation

- Verification:
  - Machine Learning
  - ML Model → More Specifications (e.g. Fairness, Safeguards)
  - Generate/Pickle → ML Model
  - Coding → Implementation

Requirements Analysis

---

9.3
see explanation at https://medium.com/@ckaestne/machine-learning-is-requirements-engineering-897ae5eff4
EXAMPLE AND DISCUSSION

Model learned from gathered data (~ interviews, sufficient? representative?)

Cannot equally satisfy all stakeholders, conflicting goals; judgement call, compromises, constraints

Implementation is trivial/automatically generated

Does it meet the users' expectations?

Is the model compatible with other specifications? (fairness, robustness)

What if we cannot understand the model? (interpretability)

IF age between 18–20 and sex is male THEN predict arrest
ELSE IF age between 21–23 and 2–3 prior offenses THEN predict arrest
ELSE IF more than three priors THEN predict arrest
ELSE predict no arrest

9.4
TERMINOLOGY SUGGESTIONS

- Avoid term *model bug*, no agreement, no standardization
- *Performance* or *accuracy* are better fitting terms than *correct* for model quality
- Careful with the term *testing* for measuring *prediction accuracy*, be aware of different connotations
- *Verification/validation* analogy may help frame thinking, but will likely be confusing to most without longer explanation
CURATING VALIDATION DATA

(Learning from Software Testing?)
HOW MUCH VALIDATION DATA?

- Problem dependent
- Statistics can give confidence interval for results
  - e.g. Sample Size Calculator: 384 samples needed for ±5% confidence interval (95% conf. level; 1M population)
- Experience and heuristics. Example: Hulten's heuristics for stable problems:
  - 10s is too small
  - 100s sanity check
  - 1000s usually good
  - 10000s probably overkill
  - Reserve 1000s recent data points for evaluation (or 10%, whichever is more)
  - Reserve 100s for important subpopulations
SOFTWARE TESTING ANALOGY: TEST ADEQUACY
SOFTWARE TESTING ANALOGY: TEST ADEQUACY

- Specification coverage (e.g., use cases, boundary conditions):
  - No specification!
  - ~> Do we have data for all important use cases and subpopulations?
  - ~> Do we have representatives data for all output classes?
- White-box coverage (e.g., branch coverage)
  - All path of a decision tree?
  - All neurons activated at least once in a DNN? (several papers "neuron coverage")
  - Linear regression models??
- Mutation scores
  - Mutating model parameters? Hyper parameters?
  - When is a mutant killed?

Does any of this make sense?
VALIDATION DATA REPRESENTATIVE?

- Validation data should reflect usage data
- Be aware of data drift (face recognition during pandemic, new patterns in credit card fraud detection)
- "Out of distribution" predictions often low quality (it may even be worth to detect out of distribution data in production, more later)
INDEPENDENCE OF DATA: TEMPORAL

Attempt to predict the stock price development for different companies based on twitter posts

Data: stock prices of 1000 companies over 4 years and twitter mentions of those companies

Problems of random train--validation split?
The model will be evaluated on past stock prices knowing the future prices of the companies in the training set. Even if we split by companies, we could observe general future trends in the economy during training.
INDEPENDENCE OF DATA: TEMPORAL
The curve is the real trend, red points are training data, green points are validation data. If validation data is randomly selected, it is much easier to predict, because the trends around it are known.
INDEPENDENCE OF DATA: RELATED DATAPoints

Kaggle competition on detecting distracted drivers

Relation of datapoints may not be in the data (e.g., driver)

https://www.fast.ai/2017/11/13/validation-sets/
Many potential subtle and less subtle problems:

- Sales from same user
- Pictures taken on same day
"Call mom" "What's the weather tomorrow?" "Add asafetida to my shopping list"
There Is a Racial Divide in Speech-Recognition Systems, Researchers Say: Technology from Amazon, Apple, Google, IBM and Microsoft misidentified 35 percent of words from people who were black. White people fared much better. -- NYTimes March 2020
NOT ALL INPUTS ARE EQUAL

some random mistakes vs rare but biased mistakes?

- A system to detect when somebody is at the door that never works for people under 5ft (1.52m)
- A spam filter that deletes alerts from banks

Consider separate evaluations for important subpopulations; monitor mistakes in production
IDENTIFY IMPORTANT INPUTS

Curate Validation Data for Specific Problems and Subpopulations:

- *Regression testing:* Validation dataset for important inputs ("call mom") -- expect very high accuracy -- closest equivalent to **unit tests**
- *Uniformness/fairness testing:* Separate validation dataset for different subpopulations (e.g., accents) -- expect comparable accuracy
- *Setting goals:* Validation datasets for challenging cases or stretch goals -- accept lower accuracy

Derive from requirements, experts, user feedback, expected problems etc. Think *blackbox testing.*
IMPORTANT INPUT GROUPS FOR CANCER DETECTION?
BLACK-BOX TESTING TECHNIQUES AS INSPIRATION?

- Boundary value analysis
- Partition testing & equivalence classes
- Combinatorial testing
- Decision tables

Use to identify subpopulations (validation datasets), not individual tests.
AUTOMATED (RANDOM) TESTING

(if it wasn't for that darn oracle problem)
Many techniques to generate test cases
- Dumb fuzzing: generate random inputs
- Smart fuzzing (e.g., symbolic execution, coverage guided fuzzing): generate inputs to maximally cover the implementation
- Program analysis to understand the shape of inputs, learning from existing tests
- Minimizing redundant tests
- Abstracting/simulating/mocking the environment
AUTOMATED TEST DATA GENERATION?

- Completely random data generation (uniform sampling from each feature's domain)
- Using knowledge about feature distributions (sample from each feature's distribution)
- Knowledge about dependencies among features and whole population distribution (e.g., model with probabilistic programming language)
- Mutate from existing inputs (e.g., small random modifications to select features)

- But how do we get labels?
RECALL: THE ORACLE PROBLEM

How do we know the expected output of a test?

```java
assertEquals(??, factorPrime(15485863));
```

- Manually construct input-output pairs (does not scale, cannot automate)
- Comparison against gold standard (e.g., alternative implementation, executable specification)
- Checking of global properties only -- crashes, buffer overflows, code injections
- Manually written assertions -- partial specifications checked at runtime
MACHINE LEARNED MODELS = UNTESTABLE SOFTWARE?

- Manually construct input-output pairs (does not scale, cannot automate)
  - too expensive at scale
- Comparison against gold standard (e.g., alternative implementation, executable specification)
  - no specification, usually no other "correct" model
  - comparing different techniques useful? (see ensemble learning)
- Checking of global properties only -- crashes, buffer overflows, code injections
  - ??
- Manually written assertions -- partial specifications checked at runtime
  - ??
INVARIANTS IN MACHINE LEARNED MODELS?
EXAMPLES OF INVARIANTS

- Credit rating should not depend on gender:
  - $\forall x. f(x[\text{gender} \leftarrow \text{male}]) = f(x[\text{gender} \leftarrow \text{female}])$

- Synonyms should not change the sentiment of text:
  - $\forall x. f(x) = f(\text{replace}(x, "is not", "isn't"))$

- Negation should swap meaning:
  - $\forall x \in "X \text{ is } Y". f(x) = 1 - f(\text{replace}(x, " is ", " is not "))$

- Robustness around training data:
  - $\forall x \in \text{ training data. } \forall y \in \text{ mutate}(x, \delta). f(x) = f(y)$

- Low credit scores should never get a loan (sufficient conditions for classification, "anchors"):
  - $\forall x. \text{score} < 649 \Rightarrow \neg f(x)$

Identifying invariants requires domain knowledge of the problem!
Formal description of relationships among inputs and outputs (*Metamorphic Relations*)

In general, for a model $f$ and inputs $x$ define two functions to transform inputs and outputs $g_I$ and $g_O$ such that:

\[ \forall x. f(g_I(x)) = g_O(f(x)) \]

e.g. $g_I(x) = \text{replace}(x, " is ", " is not ")$ and $g_O(x) = \neg x$
ON TESTING WITH INVARIANTS/ASSERTIONS

- Defining good metamorphic relations requires knowledge of the problem domain
- Good metamorphic relations focus on parts of the system
- Invariants usually cover only one aspect of correctness
- Invariants and near-invariants can be mined automatically from sample data (see specification mining and anchors)

Further reading:

INVARIANT CHECKING ALIGNS WITH REQUIREMENTS VALIDATION
AUTOMATED TESTING / TEST CASE GENERATION

- Many techniques to generate test cases
- Dumb fuzzing: generate random inputs
- Smart fuzzing (e.g., symbolic execution, coverage guided fuzzing): generate inputs to maximally cover the implementation
- Program analysis to understand the shape of inputs, learning from existing tests
- Minimizing redundant tests
- Abstracting/simulating/mocking the environment
- Typically looking for crashing bugs or assertion violations
APPROACHES FOR CHECKING IN VARIANTS

- Generating test data (random, distributions) usually easy
- For many techniques gradient-based techniques to search for invariant violations (see adversarial ML)
- Early work on formally verifying invariants for certain models (e.g., small deep neural networks)

ONE MORE THING: SIMULATION-BASED TESTING

- Derive input-output pairs from simulation, esp. in vision systems
- Example: Vision for self-driving cars:
  - Render scene -> add noise ->
    recognize -> compare recognized result with simulator state
- Quality depends on quality of the simulator and how well it can produce inputs from outputs:
  - examples: render picture/video, synthesize speech, ...
  - Less suitable where input-output relationship unknown, e.g., cancer detection, housing price prediction, shopping recommendations

CONTINUOUS INTEGRATION FOR MODEL QUALITY
CONTINUOUS INTEGRATION FOR MODEL QUALITY

- Testing script
  - Existing model: Implementation to automatically evaluate model on labeled training set; multiple separate evaluation sets possible, e.g., for critical subcommunities or regressions
  - Training model: Automatically train and evaluate model, possibly using cross-validation; many ML libraries provide built-in support
  - Report accuracy, recall, etc. in console output or log files
  - May deploy learning and evaluation tasks to cloud services
  - Optionally: Fail test below quality bound (e.g., accuracy < .9; accuracy < accuracy of last model)
- Version control test data, model and test scripts, ideally also learning data and learning code (feature extraction, modeling, ...)
- Continuous integration tool can trigger test script and parse output, plot for comparisons (e.g., similar to performance tests)
- Optionally: Continuous deployment to production server
DASHBOARDS FOR MODEL EVALUATION RESULTS
Test Data Performance

- **Threshold:** 0.0584 - 0.288

**Performance:**

- **Precision-Recall:**
  - Precision: 0.99
  - Recall: 0.94

- **ROC**
  - TP Rate: 0.29
  - FP Rate: 0.92

**Confusion Matrix**

- **Positive label:** true
- **Actual:**
  - YES: 0.21
  - NO: 0.79
- **Predicted:**
  - YES: 0.21
  - NO: 0.79

**Calibration**

- **Reliability**
  - Fraction of positives vs. mean predicted value

The reliability diagram shows how reliable (or "well-calibrated") the model's probability estimates are when evaluated on the test data. For example, a well-calibrated (binary) model should classify the samples such that among the samples to which it gives a probability close to 0.8 of belonging to the positive class, approximately 80% of those samples actually belong to the positive class. More Info

- **A Perfectly Calibrated Model**
- **This Model (Before Calibration)**
- **This Model (After Calibration)**

**AUC:** 0.7936

**Error:** 0.4907
SPECIALIZED CI SYSTEMS

Renggli et. al, Continuous Integration of Machine Learning Models with ease.ml/ci: Towards a Rigorous Yet Practical Treatment, SysML 2019
DASHBOARDS FOR COMPARING MODELS

SUMMARY

- Model prediction accuracy only one part of system quality
- Select suitable measure for prediction accuracy, depending on problem (recall, MAPE, AUC, MAP@K, ...)
- Ensure independence of test and validation data
- Software testing is a poor analogy (model bug); validation may be a better analogy
- Still learn from software testing
  - Carefully select test data
  - Not all inputs are equal: Identify important inputs (inspiration from blackbox testing)
- Automated random testing
  - Feasible with invariants (e.g. metamorphic relations)
  - Sometimes possible with simulation
- Automate the test execution with continuous integration