Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

CAS Practical Machine Learning
Introduction

Evaluation

Prof. Dr. Jürgen Vogel (juergen.vogel@bfh.ch)

How Good is the Machine Learning System?

- returned result is good if it solves the problem at hand
  - may be qualitative or quantitative
  - may be subjective (user need, context, and preferences)
  - may change over time
  - also depends on factors such as credibility, specificity, exhaustivity, recency, clarity, interpretability… of the result
- thus, the ML system needs to be assessed in “real-life” situations
  - often with user involvement
  - similar methods as with user requirements research
    - usability tests, interviews, field studies, log analysis, …
  - but takes time and is costly
- alternative: pre-defined test settings with quantitative evaluation to allow for automated testing

Metrics

Evaluation

Evaluation Metrics for Correctness

- Success
  = result is correct
  - success rate = #correct results / | test set |
  - aka accuracy
- Error
  = result is incorrect
  - error rate = #errors / | test set |

Generalized Success Rate (Accuracy)

- our ML system takes some test data D as input and produces some results

  $D \rightarrow \{r'_1, ..., r'_n\}$

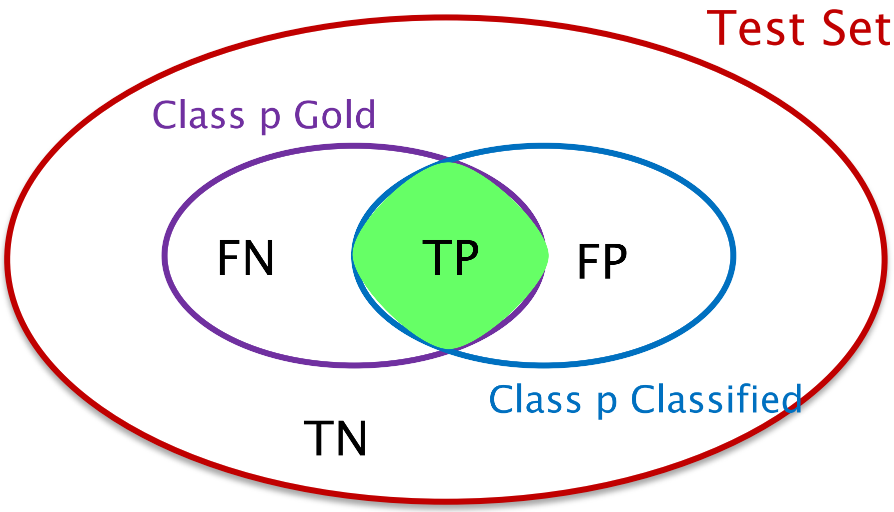
  - e.g., if r'_i are from a list of predefined labels, we call this classification
- the test data also includes the expected result ("gold standard")

  $D \rightarrow \{r_1, ..., r_n\}$
- for the test setting, we define some comparison function(s)

  c(r, r') = 1 if r = r', 0 else
- then we can calculate the success rate SR as

$$SR = \frac{1}{n}\sum_{i=1}^{n} c(r_i, r'_i)$$

Precision and Recall for Binary Classification



	positive gold	negative gold
positive classified	true positive (TP)	false positive (FP)
negative classified	false negative (FN)	true negative (TN)

Precision

- P = TP / | Class p Classified |
- Fraction of items in Class p classified that are also Class p in the gold standard
- Provides a measure of the "degree of soundness" of the system

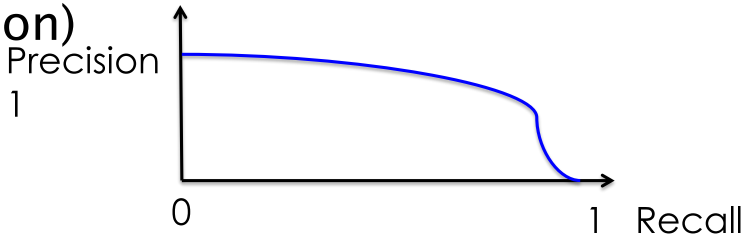
Recall

- R = TP / | Class p Gold |
- Fraction of Class p items in the gold standard that are also classified as Class p
- Provides a measure of the "degree of completeness" of the system

Precision vs. Recall

- There is often a trade-off between Precision and Recall
  - improving the algorithm towards one weakens the other
  - why?

- Can get maximum recall (but low precision) by classifying all items as Class p!
  - Recall is a non-decreasing function of the number of docs retrieved
  - Precision may be computed at different levels of recall



- which one to emphasize depends on the usage scenario, e.g., in IR
  - Precision-oriented users
    - Web surfers
  - Recall-oriented users
    - Professional searchers, legals, intelligence analysts

Precision vs. Recall

In an attempt to measure the overall quality:

F-measure

- combined measure that assesses the tradeoff between precision and recall (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \beta^2 = \frac{1-\alpha}{\alpha}$$

  - values of $\beta<1$ emphasize precision
  - values of $\beta>1$ emphasize recall
- in most cases, the balanced F-measure is used
  - F1 = 2 * precision * recall / (precision + recall)
  - i.e., with $\beta = 1$ or $\alpha = ½$

Other Metrics

- the generalization of our binary classifier result matrix (classification result vs. gold standard) is called a confusion matrix
  - many different metrics can be derived from this (see https://en.wikipedia.org/wiki/Confusion_matrix)
  - other widely used metrics include ROC, K-S, gain/lift, …
- for specific ML problems and algorithms, many additional metrics exist
- also important: operational performance metrics, e.g.,
  - training/classification time
  - processing data/time unit
  - data exchange data/time unit
- depending on the task at hand, it may be necessary to define your own metric(s)

Automated Evaluation

Evaluation

Automated Evaluation Workflow

How can we automate evaluation?

1. define a controlled test set (benchmark)
   - collection of data
   - one or more tasks to be solved by the ML system
   - expected results
     - created by (typically several) domain experts
     - "gold standard"
2. execute ML system for test set
3. compare computed results against expected results
   - depending on the task, the result can be
     - correct or not
       - face detection: the face has been detected or not
     - partially correct
       - face detection: 2 out of 3 faces in the picture have been detected
     - better (or equal or worse) than another method

Evaluation Goals

Compare a solution with…

- different configuration options
- alternative solutions
- a basic solution (“baseline”)
- the industry and/or academic leader ("state-of-the-art")
- human performance ("gold standard")
- itself over time

Data for Training and Testing

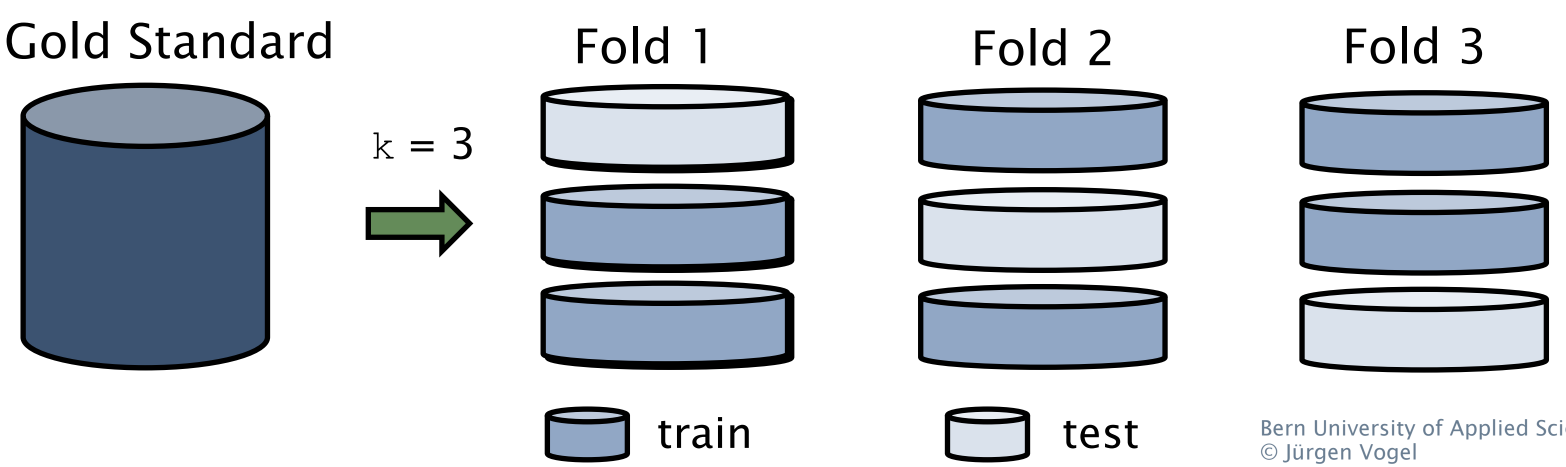
Evaluation

Using Data for Training and Testing

- ML methods usually require fine-tuning for good quality, e.g.,
  - K-Means Clustering: K
- for this we need: **training**
  = execute method on training data and adapt until satisfied
- as a final step: **test**
  = execute method on test data and obtain evaluation metric
- CAREFUL: never ever use the same data for training and testing!
  1. do not test training data
  2. do not train on test data
  - why?
- BUT: gold standard is often small
  - expensive to create
  - needs to be divided into training and test data

K-Fold Cross Validation

- how to split gold standard data into test and training set such that
  - we have enough training data?
  - our test results are not biased?
- k-fold cross validation
  - split data into k folds
  - use $(k-1)$ for training, 1 for testing
  - repeat k times
  - average results
- good value for k: 10



Dataset Challenges

Potential problems: is the dataset

- correct?
- large enough?
- representative?
- causing overfitting?

Standard Datasets

- for many application domains, large datasets are available
  - not all free but still cost saving
  - allows to compare approaches in a larger community
- where to search
  - Wikipedia
  - kaggle (https://www.kaggle.com/) and other ML sites
  - research groups at Universities
  - conference series
  - research articles
  - data collecting companies and public administrations