Recap

- Machine learning: Let the machine figure out which properties are relevant when
- Feature-based ML: Humans define domain-specific features
- Neural ML: Machine *also* figures out which features to use
- Train and test data
- Tabular data as input for machine learning systems
- File formats: CSV/TSV, ARFF
- Basic statistics about features and classes
  - I.e., how often does each feature value appear?

Classification Evaluation

Sprachverarbeitung (VL + Ü)

Nils Reiter

May 2, 2023

UNIVERSITÄT ZU KÖLN

INSTITUT FÜR DIGITAL HUMANITIES UNIVERSITÄT ZU KÖLN

Evaluation

- For today, we consider the actual ML stuff as a black box
- How exactly do we evaluate? How do we measure how good predictions are?

Evaluation

- For today, we consider the actual ML stuff as a black box
- How exactly do we evaluate? How do we measure how good predictions are?

Example (Sentiment Analysis)

- Task: Assign a polarity (positive/neutral/negative) to a linguistic expression
- Linguistic expression: sentences, phrases, documents
  - In this example: Documents

Evaluation

- For today, we consider the actual ML stuff as a black box
- How exactly do we evaluate? How do we measure how good predictions are?

Example (Sentiment Analysis)

- Task: Assign a polarity (positive/neutral/negative) to a linguistic expression
- Linguistic expression: sentences, phrases, documents
  - In this example: Documents
- Classification task: Instances are sorted into previously known categories

Evaluation

- For today, we consider the actual ML stuff as a black box
- How exactly do we evaluate? How do we measure how good predictions are?

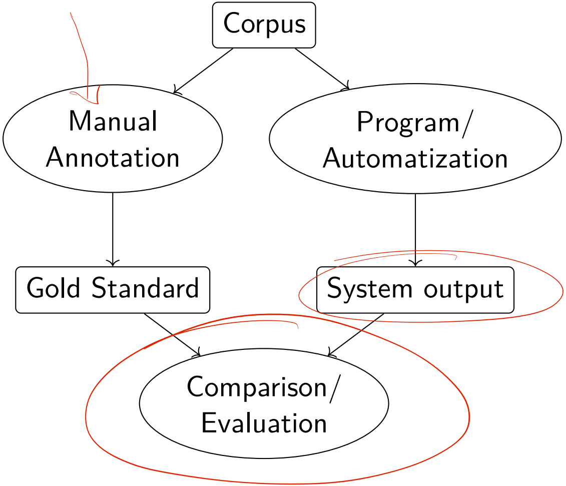
Example (Sentiment Analysis)

- Task: Assign a polarity (positive/neutral/negative) to a linguistic expression
- Linguistic expression: sentences, phrases, documents
  - In this example: Documents
- Classification task: Instances are sorted into previously known categories
- Data set: 100 documents that have labels
  - I.e., we know the result to expect

Annotation Time!

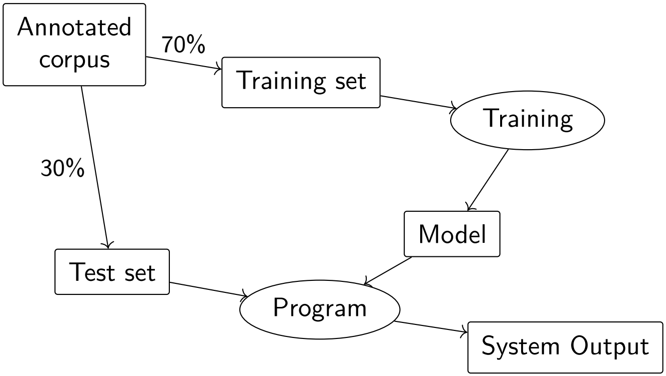
Gefühlt ist die Lage wieder wie kurz nach der Einführung der Kontaktbeschränkungen: die eine Hälfte denkt, jetzt kann man wieder lustig bummeln gehen, die andere Hälfte ist total panisch und zählt Menschen im Park.
Besonders die Senioren werden von den Kontaktbeschränkungen schwer und hart getroffen, obgleich es zu ihrem eigenen Schutz dient. Wir dürfen in dieser schweren Zeit die Seniorinnen und Senioren nicht aus dem Blick verlieren.
Gute Regelung. Kontaktbeschränkungen max. 2 Personen. (Bemerkung: das sind immer die gleichen 2 Personen, sonst macht das keinen Sinn, das bitte noch klarstellen) 1,5 bis 2 m Abstand Wenn immer es geht: #BleibtZuhause Eigener Hausstand OK. https://t.co/zuNpf0pjYr

Experiments



Evaluation

- Goal: Predict the quality on new data
- The program cannot have seen the data, so that it's a realistic test



Evaluation

- Comparison of system output with gold standard
  - »Intrinsic evaluation«
- Two sets of predictions for the items
  - One set from the gold standard
  - One set from the system
- Two aspects to talk about
  - Evaluation metric (how we quantify the performance)
  - Metric interpretation (what we think the metric tells us)

Evaluation

- Comparison of system output with gold standard
  - »Intrinsic evaluation«
- Two sets of predictions for the items
  - One set from the gold standard
  - One set from the system
- Two aspects to talk about
  - Evaluation metric (how we quantify the performance)
  - Metric interpretation (what we think the metric tells us)

Example (Sentiment Analysis)

- Gold standard: `[1, 0, -1, -1]`
- System output: `[1, -1, 1, 0]`
- (positive: 1, neutral: 0, negative: -1)

Extrinsic Evaluation

- In some cases, reference data for a task doesn't exist or can't be created
- Extrinsic evaluation: Evaluate a downstream application
- Compare performance of downstream application
  - Without your component
  - With your component
- Assumptions
  - Your component helps performance of the downstream application
  - We know how to evaluate the downstream task

Extrinsic Evaluation

- In some cases, reference data for a task doesn't exist or can't be created
- Extrinsic evaluation: Evaluate a downstream application
- Compare performance of downstream application
  - Without your component
  - With your component
- Assumptions
  - Your component helps performance of the downstream application
  - We know how to evaluate the downstream task

$$\boxed{\text{Component}} \longrightarrow \boxed{\text{Downstream application}}$$

Section 1

Evaluation Metrics, Part 1

Evaluation

Accuracy and Error Rate

- Accuracy
  - Percentage of correctly classified instances
  - Example above
    - $A = \frac{1}{4} = 0.25 = 25\%$
  - "the higher the better"

Evaluation

Accuracy and Error Rate

- Accuracy
  - Percentage of correctly classified instances
  - Example above
    - $A = \frac{1}{4} = 0.25 = 25\%$
  - "the higher the better"
- Error Rate
  - Percentage of *incorrectly* classified instances
  - Example above
    - $E = \frac{3}{4} = 0.75 = 75\%$
  - "the lower the better"

Evaluation

Accuracy and Error Rate

- Accuracy
  - Percentage of correctly classified instances
  - Example above
    - $A = \frac{1}{4} = 0.25 = 25\%$
  - "the higher the better"
- Error Rate
  - Percentage of *incorrectly* classified instances
  - Example above
    - $E = \frac{3}{4} = 0.75 = 75\%$
  - "the lower the better"
- $A + E = 1$, $E = 1 - A$ and $A = 1 - E$

Accuracy and Error Rate

Examples

- `G = [1, 0, 1]`, `S = [0, 0, 1]`
  - $A = ?$

Accuracy and Error Rate

Examples

- `G = [1, 0, 1], S = [0, 0, 1]`
  - $A = ?$
- `G = ["f", "m", "u", "m", "f"], S = ["m", "f", "u", "m", "f"]`
  - $E = ?$

Accuracy and Error Rate

Examples

- `G = [1, 0, 1], S = [0, 0, 1]`
  - $A = ?$
- `G = ["f", "m", "u", "m", "f"], S = ["m", "f", "u", "m", "f"]`
  - $E = ?$ 2/5 A = 3/5
- We don't need the original data for evaluation, we are just comparing gold standard classes with system output.

Section 2

Metric Interpretation and Use, Part 1

How good are 56% accuracy?

Baseline

- Something to compare with
- Justification for investing research time

Baseline

- Something to compare with
- Justification for investing research time
- Predecessor system
  - E.g., the one from last year
- Competing system
  - E.g., the one from Düsseldorf University
- Very simple system
  - E.g., a single feature decides everything
- Dummy system
  - E.g., if we make random decisions
  - Most common baseline

Baseline

- Something to compare with
- Justification for investing research time
- Predecessor system
  - E.g., the one from last year
- Competing system
  - E.g., the one from Düsseldorf University
- Very simple system
  - E.g., a single feature decides everything
- Dummy system
  - E.g., if we make random decisions
  - Most common baseline

It's allowed to specify multiple baselines

System	Accuracy
Model 1	56
Model 2	53
Model 3	58
Baseline 1	33
Baseline 2	45

Table: Results table in publication

Baseline

A simple solution to the problem

- How well can the task be solved without investing (a lot of) time and work?
- What is a simple solution, and how well does it solve the problem?

Baseline

A simple solution to the problem

- How well can the task be solved without investing (a lot of) time and work?
- What is a simple solution, and how well does it solve the problem?
- Baselines are used for comparison in experiments
- ›Real‹ algorithms should be able to beat the baseline, i.e., achieve higher accuracy
- Baselines have obvious shortcomings, are not expected to work every time
    - Although, sometimes they work surprisingly well

Baseline

Group Exercises

What are reasonable dummy baselines for these tasks?

- Detecting nouns in German texts
- Detecting sentence boundaries
- Detecting fake news
- Detecting the gender of dramatic characters (18-19th century)
- Predict the pos tag of the word after a determiner
- Given a corpus consisting of 'the Universal Declaration of Human Rights', 'Lord of the Rings' and the minutes of the European Parliament. Predict the origin of a random sentence.

Majority Baseline

- Select the most frequent category
- Works well in un-even data distributions
  - I.e., if one category is more frequent than the others
- Can be hard to beat
  - E.g. word sense disambiguation

Random Baseline

- Randomly select a category
- Works well in even distributions
  - I.e., if all categories are equally frequent

Section 3

Evaluation Metric, Part 2

Per Class Evaluation

- Accuracy gives us an overall score
- But we want to know more details:
  - Some classes are more important for applications
  - Error analysis!
- We want to evaluate per class (i.e., per polarity)

Sentiment Analysis

Different Kinds of Errors

Polarity	Document
positive	Awesome movie!
neutral	Great start, boring afterwards. Very good acting.
negative	Boring as hell
...	...

Table: Data set

Sentiment Analysis

Different Kinds of Errors

Polarity	Document
positive	Awesome movie!
neutral	Great start, boring afterwards. Very good acting.
negative	Boring as hell
...	...

Table: Data set

−1

Variant	Output
GS	1, 0, -1, 1, 1, 0, -1, 1
Model 1	1, 0, -1, 1, 1, 0, 1, 1
Model 2	1, 0, -1, 1, -1, 0, -1, 1

Sentiment Analysis

Different Kinds of Errors

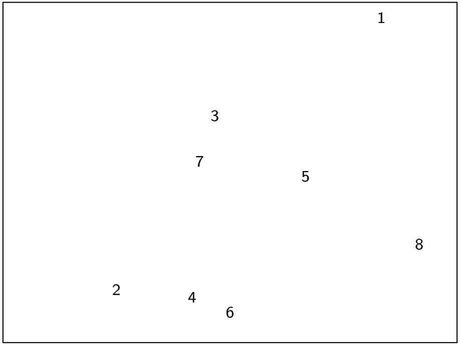


Figure: Visual representation of errors, *focussing on -1 class*

Sentiment Analysis

Different Kinds of Errors

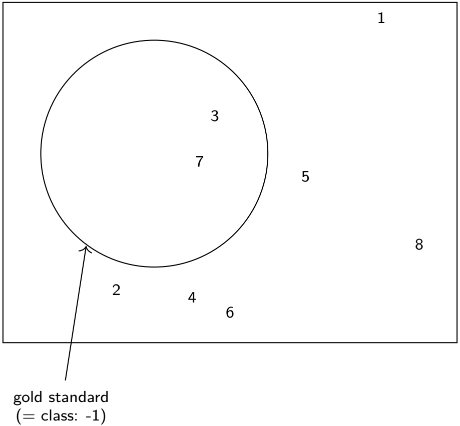


Figure: Visual representation of errors, *focussing on -1 class*

Sentiment Analysis

Different Kinds of Errors

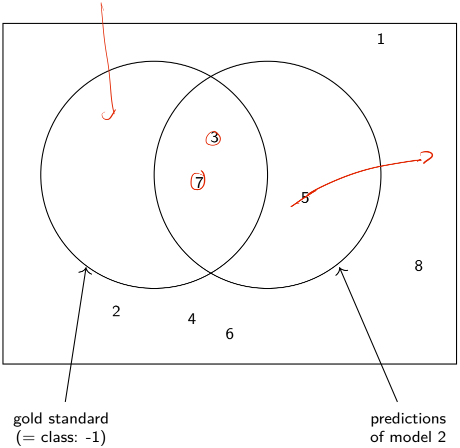
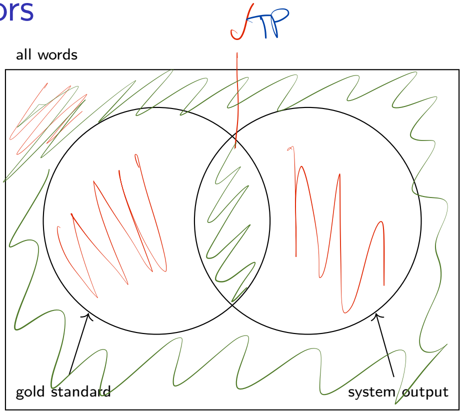
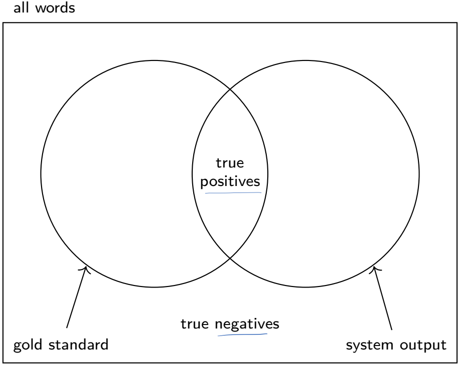


Figure: Visual representation of errors, *focussing on -1 class*

Different Kinds of Errors



Different Kinds of Errors



−1

true positive (tp) Correctly classified as target category

true negative (tn) Correctly classified as not target category

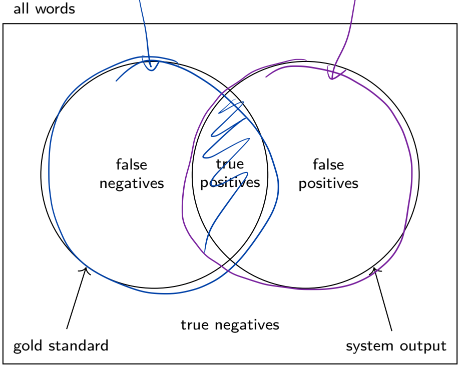
Different Kinds of Errors

Recall

Precision

$$\frac{tp}{tp+fp} = P$$

$$\frac{tp}{tp+fn} = R$$



true positive (tp) Correctly classified as target category

true negative (tn) Correctly classified as not target category

false positive (fp) Incorrectly classified as target category

false negative (fn) Incorrectly classified as not target category

Accuracy, revisited

Accuracy: Percentage of correctly classified instances

$$A = \frac{tp + tn}{tp + tn + fp + fn}$$

Accuracy, revisited

Accuracy: Percentage of correctly classified instances

$$A = \frac{tp + tn}{tp + tn + fp + fn}$$

Error rate: Percentage of incorrectly classified instances

$$E = \frac{fp + fn}{tp + tn + fp + fn}$$

Precision and Recall

Given the documents that the system marked as -1, how many of those are really -1?

Precision and Recall

Given the documents that the system marked as -1, how many of those are really -1?

$$\text{Precision} \quad P = \frac{tp}{tp + fp}$$

Precision and Recall

Given the documents that the system marked as -1, how many of those are really -1?

$$\text{Precision} \quad P = \frac{tp}{tp + fp}$$

How many of the -1 documents did the system find?

Precision and Recall

Given the documents that the system marked as -1, how many of those are really -1?

$$\text{Precision} \quad P = \frac{tp}{tp + fp}$$

How many of the -1 documents did the system find?

$$\text{Recall} \quad R = \frac{tp}{tp + fn}$$

Precision and Recall

- Enumerator: tp

Precision and Recall

- Enumerator: tp
- Precision
  - Denominator: $tp + fp$
  - Number of things that the system labelled as target category (correct and incorrect)
- Recall
  - Denominator: $tp + fn$
  - Number of things that the gold standard contained as target category (what the system should have found)

Precision and Recall

Importance/Weighting

- Weighting between P and R is application-dependent (and difficult to decide!)
- Guiding question: Which kind of error is more severe?

Precision and Recall

Importance/Weighting

- Weighting between P and R is application-dependent (and difficult to decide!)
- Guiding question: Which kind of error is more severe?
- If findings are inspected by humans
  - Precision errors are easy to spot, but recall errors cannot be detected
  - But: humans tend to trust computers

Precision and Recall

Importance/Weighting

- Weighting between P and R is application-dependent (and difficult to decide!)
- Guiding question: Which kind of error is more severe?
- If findings are inspected by humans
  - Precision errors are easy to spot, but recall errors cannot be detected
  - But: humans tend to trust computers
- Severity of consequences

Precision and Recall

Importance/Weighting

- Weighting between P and R is application-dependent (and difficult to decide!)
- Guiding question: Which kind of error is more severe?
- If findings are inspected by humans
  - Precision errors are easy to spot, but recall errors cannot be detected
  - But: humans tend to trust computers
- Severity of consequences

Example (Test performance in a pandemic)

- Individual health: Mistakenly being in quarantine is a severe limitation, and might have economic consequences
- Public health: Find more infections, even if it means a few people are mistakenly put in quarantine

F-Score

- Sometimes, it is convenient to combine precision and recall into a single number
- F-Score is common way to do that
  (it's a fancy way of averaging)
  - β can be used to weight precision and recall differently
  - $\beta = 1$ means equal weighting
- F-Measure corresponds to the harmonic mean

$$F_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R}$$

$$F_1 = 2\frac{PR}{P + R}$$

Section 4

Metric Interpretation and Use, Part 2

$\frac{45+54+78}{3} = 59$

Results in Scientific Papers

System	Class	Precision	Recall	F_1
Model 1	Class -1	45	75	
	Class 0	54	61	
	Class 1	78	12	
	Macro Average	59	49	
	Micro Average	55	56	
Baseline 1	Class -1	0	0	
	Class 0	100	0	
	Class 1	0	0	
	Macro Average	33	0	
	Micro Average	75	0	

Table: Example table with results

Micro- and Macro-Average

- Macro-Average: Arithmetic mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Micro-Average: Weighted arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Takes into account how frequent categories are

$$\frac{50 \cdot 7 + 80 \cdot 1 + 90 \cdot 2}{10}$$

$= 61$

Micro- and Macro-Average

- Macro-Average: Arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Micro-Average: Weighted arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Takes into account how frequent categories are

Class	Freq. $(= w)$	P	R
A	7	50	90
B	1	80	10
C	2	90	20
Macro Average		73	40
Micro Average		61	68

$$\frac{50 + 80 + 90}{3} = 73$$

Section 5

Data Set Organization

Generating Purpose-Specific Data Sets

- Annotated data is expensive and often the bottleneck

Generating Purpose-Specific Data Sets

- Annotated data is expensive and often the bottleneck
- Different ways to use an existing annotated data set

Generating Purpose-Specific Data Sets

- Annotated data is expensive and often the bottleneck
- Different ways to use an existing annotated data set

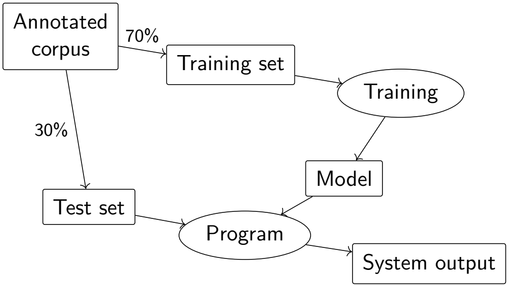


Figure: Percentage split

Generating Purpose-Specific Data Sets

- Annotated data is expensive and often the bottleneck
- Different ways to use an existing annotated data set

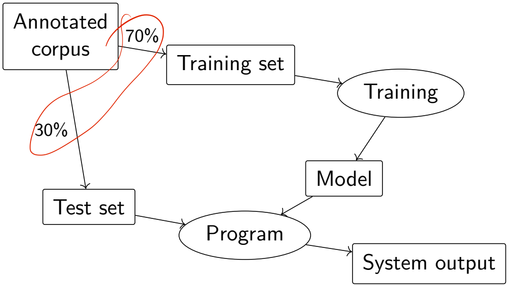
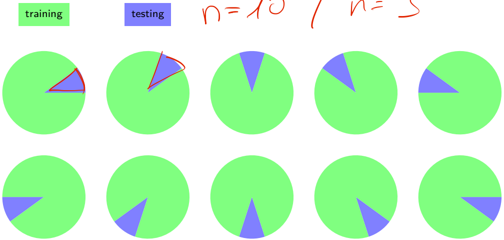


Figure: Percentage split

n=10 / n=3



Calculate P/R/F individually, then average

Figure: Cross Validation

Randomness

- Some test options or algorithms involve random numbers
  - E.g., cross validation
- Results could be unrealistically good, by chance

Randomness

- Some test options or algorithms involve random numbers
  - E.g., cross validation
- Results could be unrealistically good, by chance
- Simple solution: Run the experiments repeatedly (e.g., 1000 times)

Section 6

Summary

Summary

- Evaluation of ML systems is important
  - Because we don't know in advance what works and what does not
- Two components
  - Comparison to a baseline
    - Previous or dummy system
  - Calculation of precision/recall
    - Precision: How many of those marked as category X are truly category X?
    - Recall: How many of those that are category X has the system marked as X?
  - Training/test split or cross validation