

# Recap

- ▶ Probability theory
  - ▶ Probability: Fraction of positive over all possible events
  - ▶ Conditional probability: Restrict the space of possible events
- ▶ Naive Bayes
  - ▶ Probability-based classification algorithm
  - ▶ Assumes feature independence (therefore: “naive”)
  - ▶ Still used in many applications
    - ▶ E.g., spam classification



UNIVERSITÄT  
ZU KÖLN

# Machine Learning 2: Evaluation

## VL Sprachliche Informationsverarbeitung

Nils Reiter

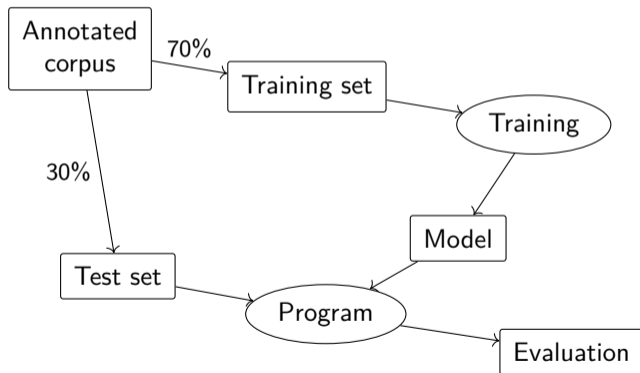
`nils.reiter@uni-koeln.de`

November 28, 2024

Winter term 2024/25

## Training and Testing


- ▶ Goal: Apply the model on new data (and estimate its performance then)
- ▶ The program cannot have seen the data, so that it is a realistic test




# Evaluation

- ▶ We *always* want to know how well a model works
- ▶ Straightforward evaluation: Comparison with a gold standard
- ⚠ ML evaluation focusses on reliability – validity often not discussed

# Evaluation

- ▶ We *always* want to know how well a model works
- ▶ Straightforward evaluation: Comparison with a gold standard
- ▶  ML evaluation focusses on reliability – validity often not discussed
- ▶ Most simple metric: Accuracy
  - ▶ Percentage of correctly classified instances (the higher the better)
  - ▶ Inverse: Error rate (percentage of incorrectly classified instances)

# Evaluation

- ▶ We *always* want to know how well a model works
- ▶ Straightforward evaluation: Comparison with a gold standard
- ▶  ML evaluation focusses on reliability – validity often not discussed
- ▶ Most simple metric: Accuracy
  - ▶ Percentage of correctly classified instances (the higher the better)
  - ▶ Inverse: Error rate (percentage of incorrectly classified instances)
- ▶ What could be problems with this metric?

# 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

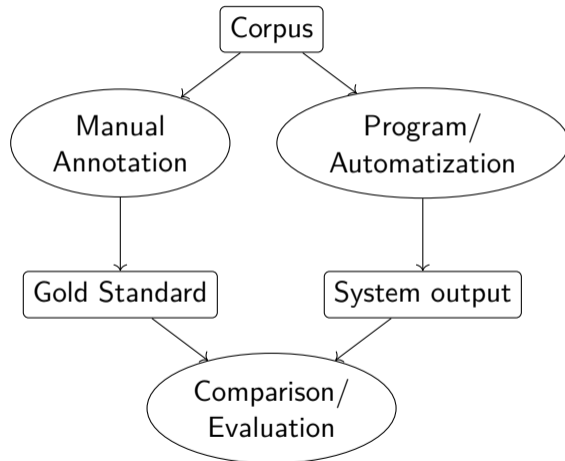
# Evaluation

- ▶ 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)

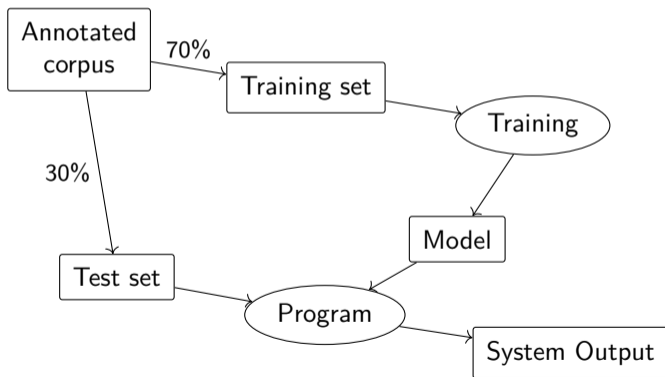
- ▶ Task: Assign a polarity (positive/neutral/negative) to a linguistic expression
- ▶ Linguistic expression: sentences, phrases, documents
  - ▶ In this example: Documents
- ▶ Classification task: Instances are sorted into previously known categories
- ▶ Data set: 100 documents that have labels
  -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

# 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

# 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

## Example (Sentiment Analysis)

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

▶  $G = [1, 0, 1], S = [0, 0, 1]$

▶  $A = \frac{1}{3}$

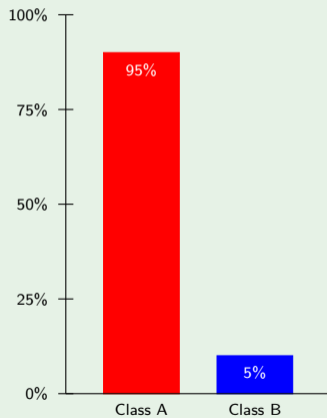
▶  $G = ["f", "m", "u", "m", "f"], S = ["m", "f", "u", "m", "f"]$

▶  $E = \frac{2}{5}$

(We don't need the original data for evaluation, we are just comparing gold standard classes with system output.)

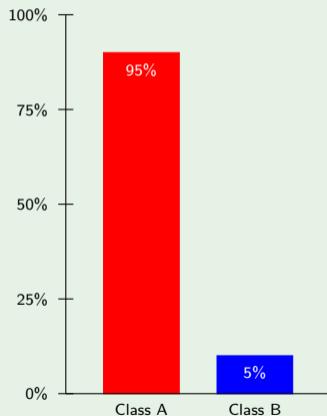
# Pitfalls

## Example (Uneven Distributions)



# Pitfalls

## Example (Uneven Distributions)



- ▶ What's the accuracy if the system assigns all items to category A?

# Pitfalls

## Example (Multiple Classes)

↓ System / Gold →	A	B	C
A	5	3	1
B	4	1	2
C	2	3	9

Table: Confusion Matrix

- ▶ What's the accuracy of the system?

# 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**

# Sentiment Analysis

## Different Kinds of Errors

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

Table: Gold Standard

# Sentiment Analysis

## Different Kinds of Errors

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

Table: Gold Standard

Variant	Output
GS	1, 0, -1, 1, 1, 0, -1, 1
Program 1	1, 0, -1, 1, 1, 0, <b>1</b> , 1
Program 2	1, 0, -1, 1, <b>-1</b> , 0, -1, 1



# Sentiment Analysis

## Different Kinds of Errors

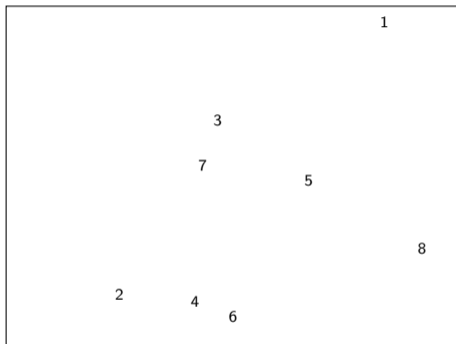


Figure: Visual representation of errors, focussing on -1 class

# Sentiment Analysis

## Different Kinds of Errors

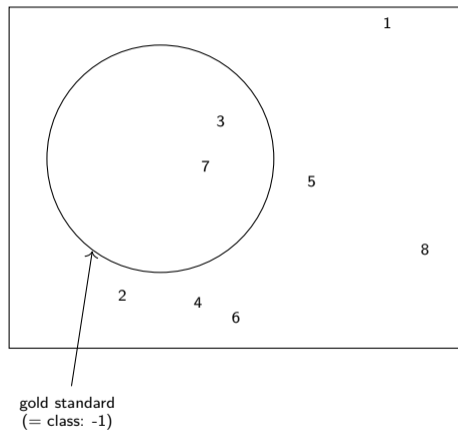


Figure: Visual representation of errors, focussing on -1 class

# Sentiment Analysis

## Different Kinds of Errors

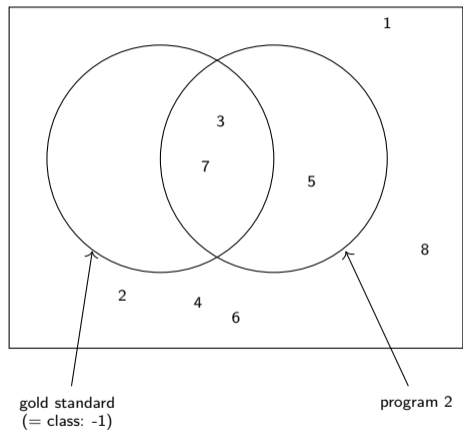
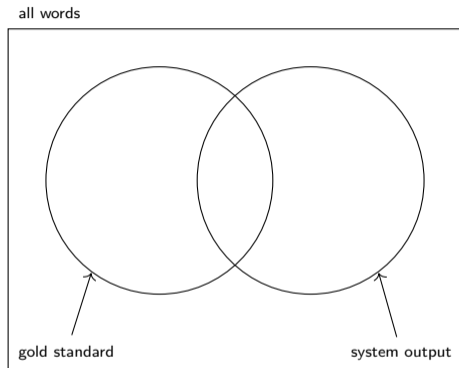
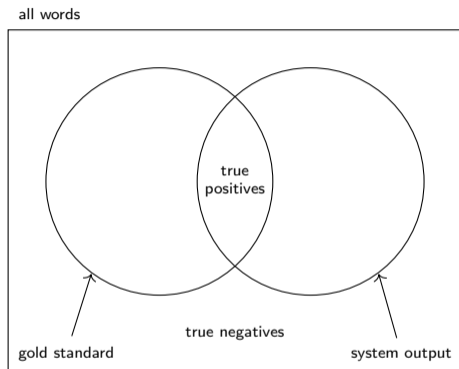


Figure: Visual representation of errors, focussing on -1 class

# Different Kinds of Errors



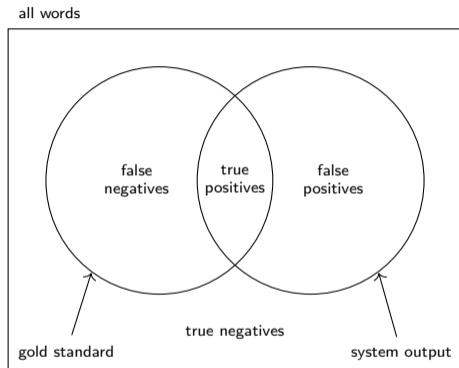
## Different Kinds of Errors



true positive (tp) Correctly classified as target category

true negative (tn) Correctly classified as not target category

## Different Kinds of Errors



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?

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

## Precision and Recall

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

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

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?

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

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

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

# 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)
  - ▶  $\beta$  can be used to weight precision and recall differently
  - ▶  $\beta = 1$  means equal weighting
- ▶ F-Measure corresponds to the harmonic mean

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

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

# 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 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
- ▶ Can be hard to beat
  - ▶ E.g. word sense disambiguation

# Validity

- ▶ Reliability: Is the model accurate, stable and predictable?
  - ▶ Measured using precision, recall, IAA, ...
- ▶ Validity: Does our model measure what we think it measures?
  - ▶ Not a metric, but an argumentation
  - ▶ Data sources
  - ▶ Knowledge about the domain and the relevant concept(s)
  - ▶ Common practices, authority



# Precision and Recall

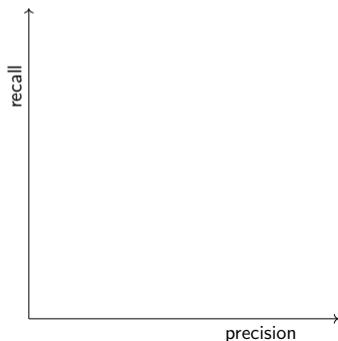
## Thresholds

- ▶ Sometimes, we have a single parameter that directly controls P and R  
E.g., a threshold for document similarity
  - ▶ Lower threshold: More documents are included  $\Rightarrow$  Higher recall, at the cost of precision
  - ▶ Higher threshold: Less documents are included  $\Rightarrow$  Higher precision, at the cost of recall

# Precision and Recall

## Thresholds

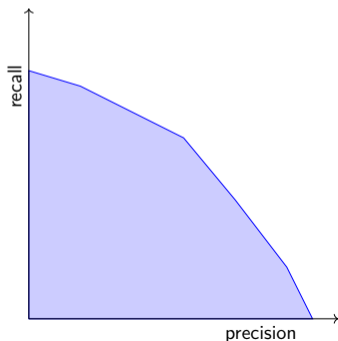
- ▶ Sometimes, we have a single parameter that directly controls P and R  
E.g., a threshold for document similarity
  - ▶ Lower threshold: More documents are included  $\Rightarrow$  Higher recall, at the cost of precision
  - ▶ Higher threshold: Less documents are included  $\Rightarrow$  Higher precision, at the cost of recall
- ▶ AUC: Area under curve



# Precision and Recall

## Thresholds

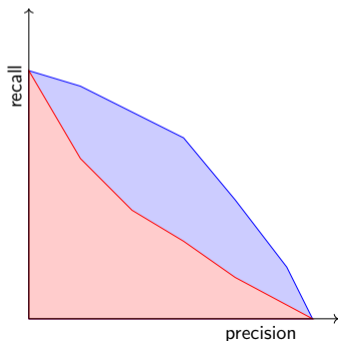
- ▶ Sometimes, we have a single parameter that directly controls P and R  
E.g., a threshold for document similarity
  - ▶ Lower threshold: More documents are included  $\Rightarrow$  Higher recall, at the cost of precision
  - ▶ Higher threshold: Less documents are included  $\Rightarrow$  Higher precision, at the cost of recall
- ▶ AUC: Area under curve



# Precision and Recall

## Thresholds

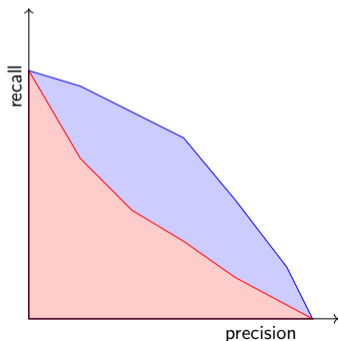
- ▶ Sometimes, we have a single parameter that directly controls P and R  
E.g., a threshold for document similarity
  - ▶ Lower threshold: More documents are included  $\Rightarrow$  Higher recall, at the cost of precision
  - ▶ Higher threshold: Less documents are included  $\Rightarrow$  Higher precision, at the cost of recall
- ▶ AUC: Area under curve



# Precision and Recall

## Thresholds

- ▶ Sometimes, we have a single parameter that directly controls P and R  
E.g., a threshold for document similarity
  - ▶ Lower threshold: More documents are included  $\Rightarrow$  Higher recall, at the cost of precision
  - ▶ Higher threshold: Less documents are included  $\Rightarrow$  Higher precision, at the cost of recall
- ▶ AUC: Area under curve



- ▶  $AUC(\text{blue}) > AUC(\text{red})$ :  
Blue system better

# References I