Machine Learning 1: Evaluation

VL Sprachverarbeitung

Nils Reiter

May 5, 2021

Recap

- Language modeling
  - Given some history, predict the next word
  - Use cases: Smart phone, ...
  - Maximum Likelihood estimation: Easy, but problematic

Neue Glossareinträge I

n-gram

Definition (Bigramm)

Ein Bigramm ist ein Textfragment mit einem Umfang von n. Im Falle eines Bigramms ist $n = 2$ (Trigramm: $n = 3$, etc.). Die Fragmente können sich sowohl aus Wörtern als auch aus Buchstabenfolgen konstituieren. N-Gramme können Focus Words enthalten.

Definition (Chi-Quadrat-Test)

Ein statistischer Test, z.B. zur Feststellung einer stochastischen Unabhängigkeit zweier Ereignisse. In der Sprachverarbeitung wird er angewendet, um Kollokationen als solche zu erkennen.

Neue Glossareinträge II

Sprachmodell

Definition (Markov Assumption)

Dies ist ein Modell zur Errechnung von kommenden Wortfolgen und Vorschlägen dafür, während der Verfassung eines Textes. Dazu werden nur der Kontext des Textes und die vorherigen Worte genutzt um einen Vorschlag zu erstellen. Dazu wird ein n-gram Modell gewählt, welches dann als p(wn | <w1,...,wn-1>) Wahrscheinlichkeit berechnet wird.
Um n zu bestimmen, muss beachtet werden, dass ein höheres n bessere Ergebnisse erzielt, jedoch mehr Speicherplatz und Trainingszeit benötigt. Die gängigsten n-gram Modelle sind Bi- und Trigram Modelle.
Beispiel:
Texteingabe in meinem Handy: (Erstes Wort selbst gewählt: Als) ›Als gut und ohne Erklärung der Frage ich habe noch eine andere Sachen gefunden und ich überlegen mir die nächsten Wochen mal was zusammen machen und mir dann die Zeit aussuchen was wir zusammen mit den Leuten zusammen haben wir das alles gut verstehen‹
In diesem Beispiel kann man sehen, dass die meisten Worte als kurze Kette mit den Worten vorher verständlich gelesen werden können, jedoch wirkt der Text eher wie eine Aneinanderreihung solcher Ketten, als dass er Sinn ergibt.

Neue Glossareinträge III

Definition (Predictive typing)

Predictive Typing signifiziert eine automatisierte maschinelle Vorhersage von bestimmten Wörtern und Ausdrücken, die auf einer Analyse von vorherigen Schreibprozessen basiert. Text- und Satzzeichen werden auf einer virtuellen Tastatur zur Auswahl vorgeschlagen. Das Sprachmodell soll Schreibvorgänge schneller und akkurater gestalten. Beispiele hierfür sind:

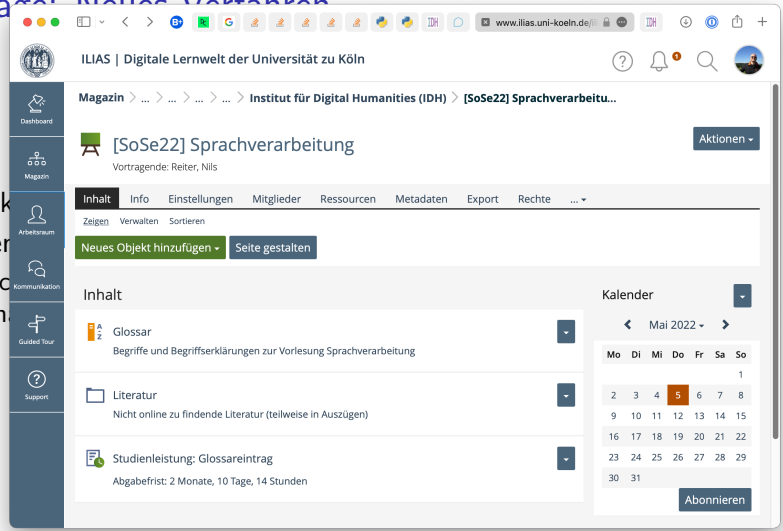
- Automatische Vervollständigung beim Ausfüllen eines Formulars
- Textvorhersage beim Schreiben einer E-Mail
- Vorhersage des nächsten Wortes bei der Eingabe einer SMS

Glossareinträge: Neues Verfahren

- Keine direkten Schreibrechte mehr im Glossar für Kursteilnehmer:innen
- Stattdessen:
  - Einreichen des Beitrags als Übung im entsprechenden Ilias-Objekt
  - Ggf. nach Überarbeitung übertrage ich ins Glossar

Glossareinträge: Neues Verfahren

- Keine direk
- Stattdesser
  - Einreic
  - Ggf. n



Section 1

Evaluation of Machine Learning Systems

Introduction

- So far: Descriptive methods
- Next weeks: Different machine learning strategies
  - Predictive methods: Given a text, predict some properties of it
- Today: Evaluation
- Goal, in general: Predict (linguistic) categories of text
  - Examples: Parts of speech, syntactic relations, semantic roles, word senses, ...

Introduction

- So far: Descriptive methods
- Next weeks: Different machine learning strategies
  - Predictive methods: Given a text, predict some properties of it
- Today: Evaluation
- Goal, in general: Predict (linguistic) categories of text
  - Examples: Parts of speech, syntactic relations, semantic roles, word senses, ...
- Why machine learning?
  - Development in NLP/CL over last 30 years
  - Language phenomena in the wild are complex and context-dependent
  - Rule-based systems difficult to develop and maintain

Evaluation

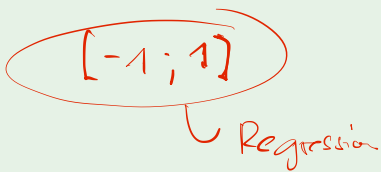
- For today, we consider the actual ML stuff as a black box
- How exactly do we evaluate? How do we measure how good predictions are?

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!

1. »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.«
2. »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.«
3. »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«

Evaluation Strategies

- Manual inspection by the developer: Run the tool, look at the results and decide
  - ⚠ Difficult to reproduce, prone to biases, implicit standards
  - Fast

Evaluation Strategies

- Manual inspection by the developer: Run the tool, look at the results and decide
  - ⚠ Difficult to reproduce, prone to biases, implicit standards
  - Fast
- Manual inspection by an expert: Run the tool, hand it over to an expert and let them decide
  - Difficult to reproduce, expensive
  - More reliable

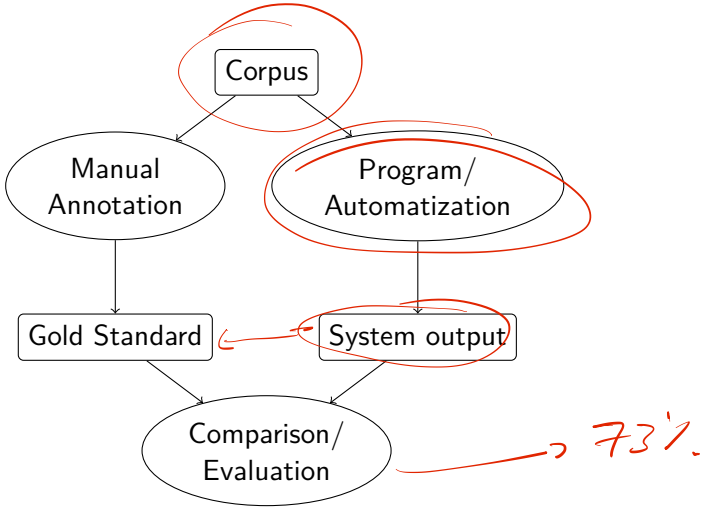
Evaluation Strategies

- Manual inspection by the developer: Run the tool, look at the results and decide
  - ⚠ Difficult to reproduce, prone to biases, implicit standards
  - Fast
- Manual inspection by an expert: Run the tool, hand it over to an expert and let them decide
  - Difficult to reproduce, expensive
  - More reliable
- Plug into an application that benefits from a component: Extrinsic evaluation
  - Need evaluation for the application, impact of component not always clear
  - Realistic evaluation (if it's a realistic application)

Evaluation Strategies

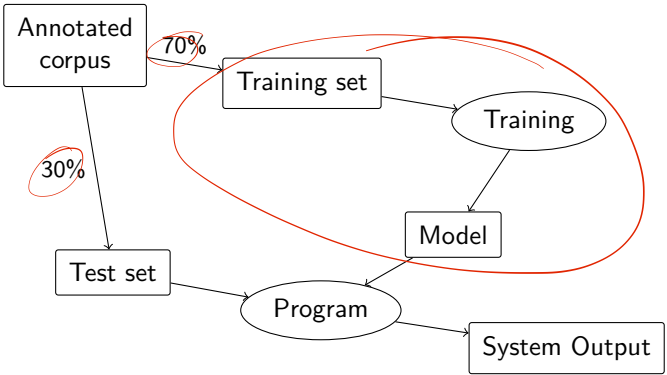
- Manual inspection by the developer: Run the tool, look at the results and decide
  - ⊖ ⚠ Difficult to reproduce, prone to biases, implicit standards
  - ⊕ Fast
- Manual inspection by an expert: Run the tool, hand it over to an expert and let them decide
  - ⊖ Difficult to reproduce, expensive
  - ⊕ More reliable
- Plug into an application that benefits from a component: Extrinsic evaluation
  - ⊖ Need evaluation for the application, impact of component not always clear
  - ⊕ Realistic evaluation (if it's a realistic application)
- Pre-defined reference data set
  - ⊖ Not always available, expensive, time-consuming
  - ⊕ Most reliable, easiest to reproduce
  - ML systems need annotated data anyway

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

Accuracy

Example (Sentiment Analysis)

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

$\frac{1}{4} = 25\%$

Extrinsic Evaluation

- In some cases, GS 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, GS 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

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

$G = [1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0]$ (5, 10)

$|G| = 12$

$S = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$ $|S| = 12$

$A = \frac{10}{12} = \frac{5}{6} = 83\%$

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

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

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

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: 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, 1, 1
Program 2	1, 0, -1, 1, -1, 0, -1, 1

Sentiment Analysis

Different Kinds of Errors

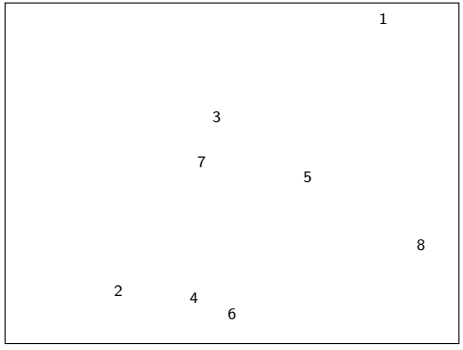


Figure: Visual representation of errors, focussing on -1 class

Sentiment Analysis

Different Kinds of Errors

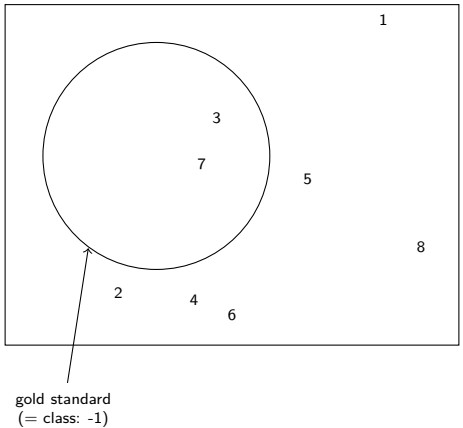


Figure: Visual representation of errors, focussing on -1 class

Sentiment Analysis

Different Kinds of Errors

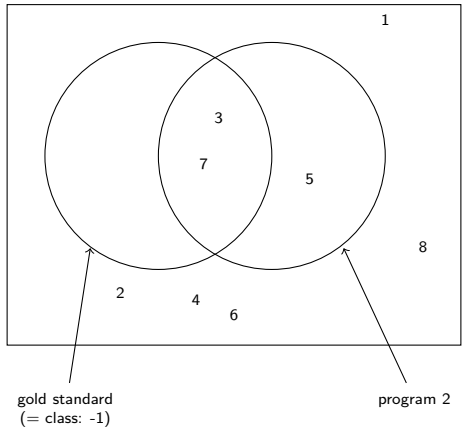
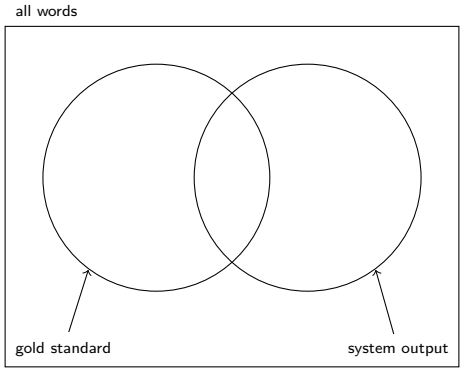
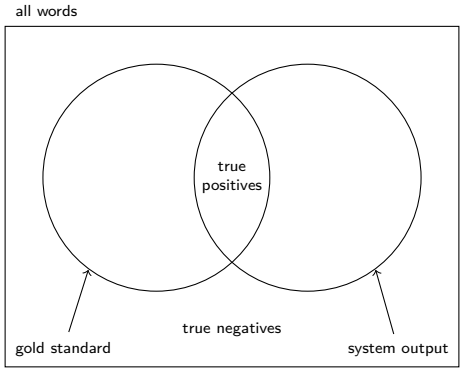


Figure: Visual representation of errors, focussing on -1 class

Different Kinds of Errors



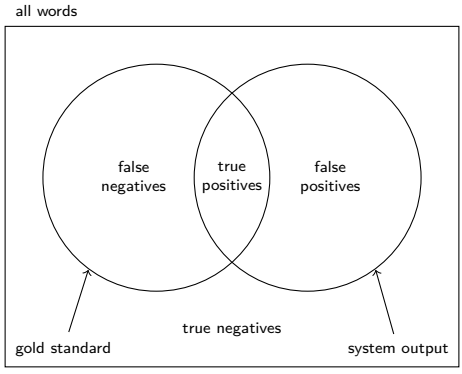
Different Kinds of Errors



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?

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

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

Precision and Recall

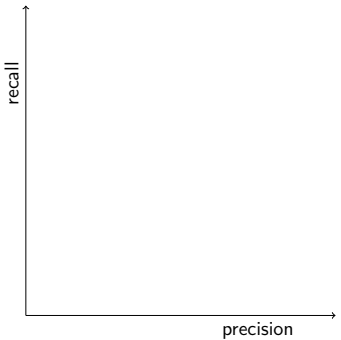
Thresholds

- 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 ⇒ Higher recall, at the cost of precision
  - Higher threshold: Less documents are included ⇒ Higher precision, at the cost of recall

Precision and Recall

Thresholds

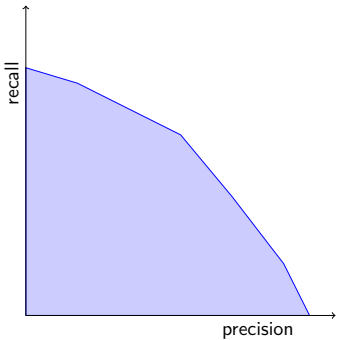
- Sometimes, we have a single parameter that directly controls P and R
  E.g., a threshold for document similarity
  - Lower threshold: More documents are included ⇒ Higher recall, at the cost of precision
  - Higher threshold: Less documents are included ⇒ Higher precision, at the cost of recall
- AUC: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



Precision and Recall

Thresholds

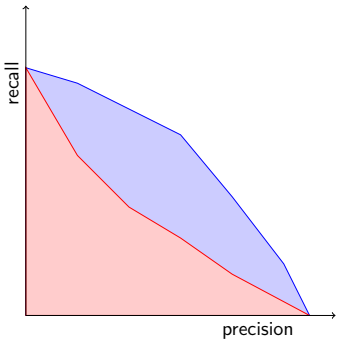
- Sometimes, we have a single parameter that directly controls P and R
  E.g., a threshold for document similarity
  - Lower threshold: More documents are included ⇒ Higher recall, at the cost of precision
  - Higher threshold: Less documents are included ⇒ Higher precision, at the cost of recall
- AUC: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



Precision and Recall

Thresholds

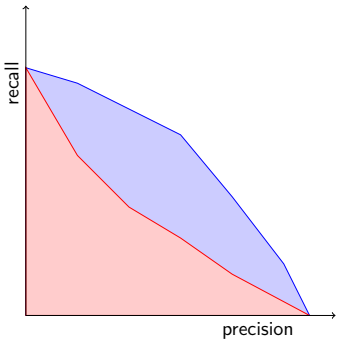
- Sometimes, we have a single parameter that directly controls P and R
  E.g., a threshold for document similarity
  - Lower threshold: More documents are included ⇒ Higher recall, at the cost of precision
  - Higher threshold: Less documents are included ⇒ Higher precision, at the cost of recall
- AUC: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



- AUC(blue) $>$ AUC(red):
  Blue system better

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

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

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

Data Sets for Different Purposes

- Training data set: Count words, estimate probabilities
- Test data set: Simulate application to see how well it works
- Application data set: Do the actual application
  - Usually skipped in research

Data Sets for Different Purposes

- Training data set: Count words, estimate probabilities
- Test data set: Simulate application to see how well it works
- Application data set: Do the actual application
  - Usually skipped in research
- Development data set: Write code, test implementation on dummy examples, fix bugs
- Validation data set: Sometimes used for smoothing or hyperparameter tuning

Data Sets for Different Purposes

- Training data set: Count words, estimate probabilities
- Test data set: Simulate application to see how well it works
- Application data set: Do the actual application
  - Usually skipped in research
- Development data set: Write code, test implementation on dummy examples, fix bugs
- Validation data set: Sometimes used for smoothing or hyperparameter tuning

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

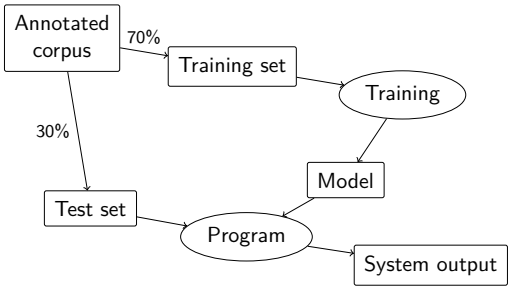


Figure: Percentage split

Generating Purpose-Specific Data Sets

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

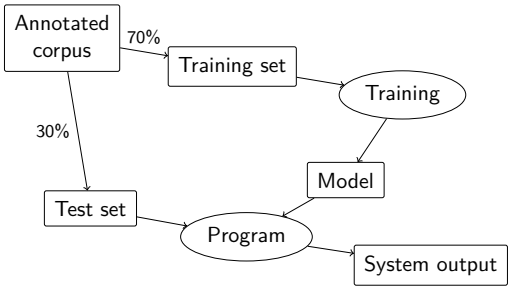
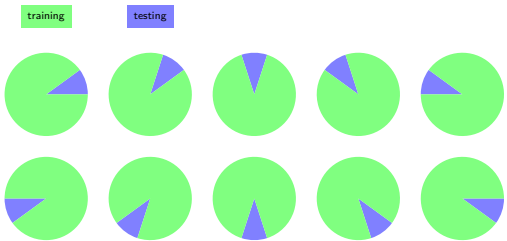


Figure: Percentage split



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 2

Summary

Summary

- Evaluation
  - Baseline: Whatever we compare against (our decision, but needs to be convincing)
  - Intrinsic: Compare against a gold standard
    - Accuracy, precision, recall, f-score
  - Extrinsic: Evaluate downstream application