Statistische Modellierung und Evaluation

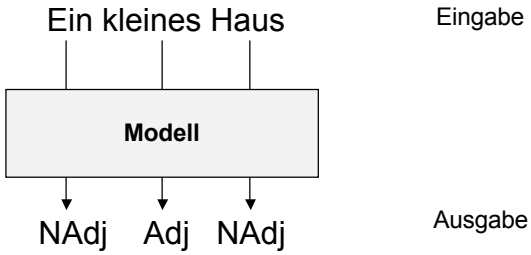
Sebastian Pado
12.12.2006

Computerlinguistik 2006 (Rückblick)

- Ein wichtiges Ziel:
  - Modelle zur automatischen linguistischen Analyse von Text
- Herausforderungen:
  - Mehrsprachigkeit
  - Große Textmengen, gemischte Qualität
- Wünsche an ein Analysemodell:
  - Robust
  - Mit geringem Aufwand zu konstruieren (automatisch?)

Beispielaufgabe: Adjektiverkennung

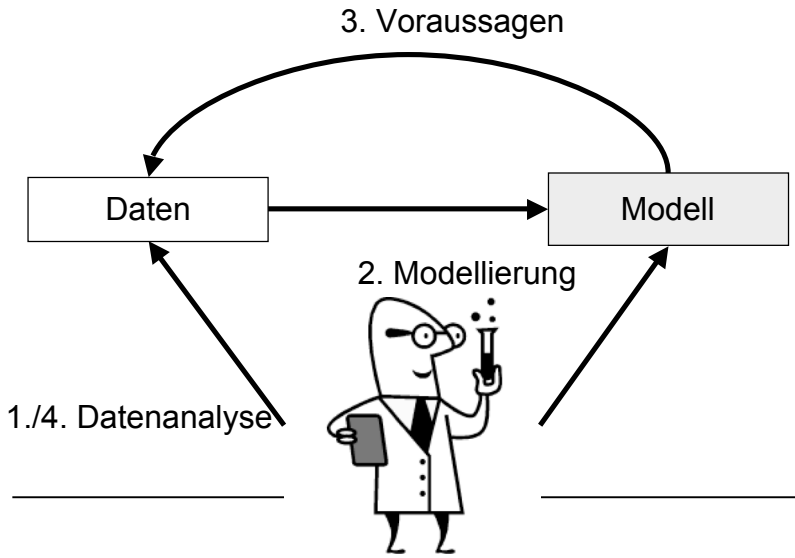
- Handelt es sich bei einem Wort (in einem fortlaufenden Text) um ein Adjektiv?



Modellierung von Adjektiverkennung

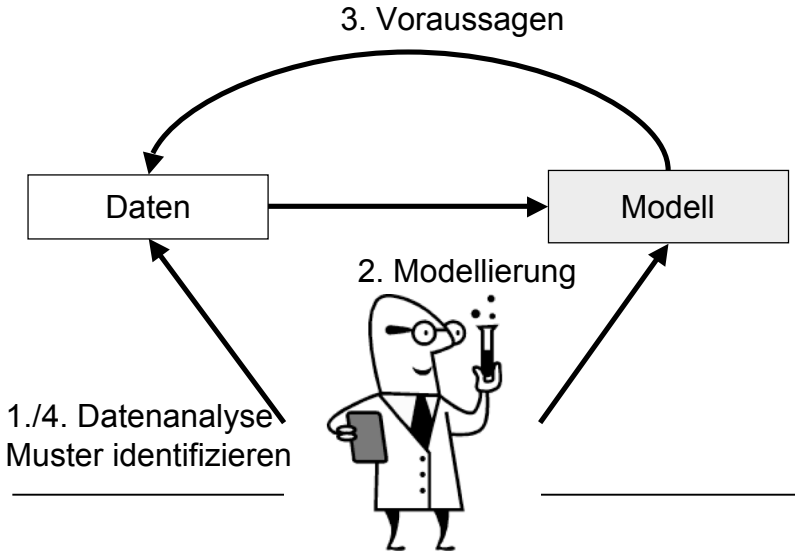
- **Lexikon**basiertes Modell
  - Prüfe, ob Wort in Adjektivliste ist
    - Problem: Adjektive sind offene Wortklasse
- **Daten**basierte Modelle
  - Modell weitgehend automatisch aus Korpus lernbar

Empirische Modellierung



5

Schritt 1: Musteranalyse



6

Datenanlyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

Ich moechte Ihnen fuer Ihren Bericht ueber
den siebenten Bericht ueber
staatliche Beihilfen in
der europaeischen Union danken.

Datenanalyse: Informative Muster für die Adjektiverkennung

Intuitive Frage: **Woran erkenne ich ein Adjektiv**?

- Nächstes Wort ist großgeschrieben (kapitalisiert)
  - `Der **siebente** Bericht`
- Vorheriges Wort ist Artikel
  - `Der **siebente** Bericht`
- Wort selbst ist nicht kapitalisiert
  - `Der **siebente** Bericht`
- Wort selbst ist kein Artikel

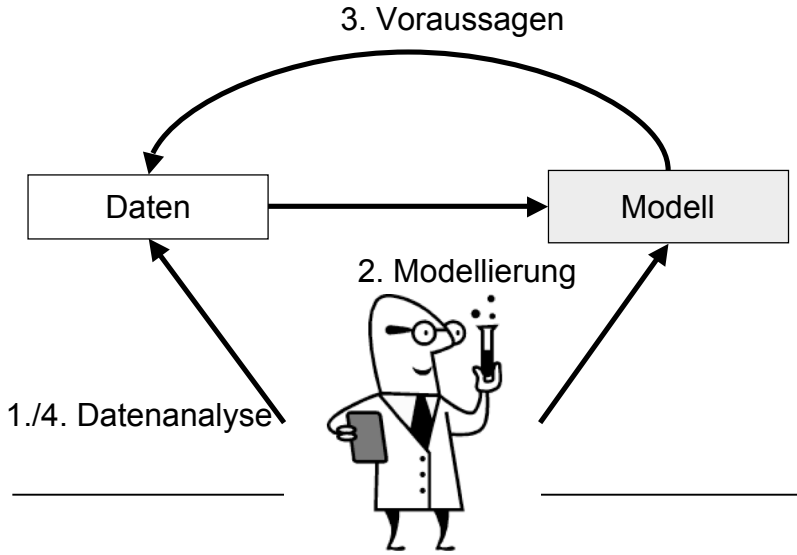
Aufteilung der Daten

Trainingsdaten (Training set)	Testdaten (Test set)

- Meistens zwischen 70% und 90% der Daten
- Grundlage der Modellierung (Schritt 1)

- Überprüfung des Modells an unabhängigen Daten (Schritt 4)

Schritt 2: Modellierung



Modelle aus Mustern

- Option 1: Schreibe explizite Regeln der Form

  `Wenn <Muster>, dann Adj/Nadj`

  Binäres (Symbolisches) Modell

- Option 2: Lerne quantiative Abhängigkeiten zwischen Mustern und "Adjektivheit"

  `<Muster>` ist zu 90% korreliert mit `Adj`

  Statistisches Modell

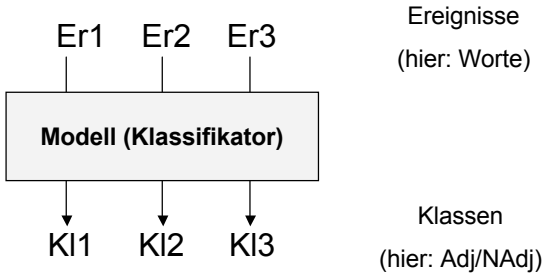
Symbolische Regeln: Implikation

- Nächstes Wort großgeschrieben ⇒ Adj
  - Der Mann (Korrektheitsproblem)
- Nächstes Wort kapitalisiert und Wort kein Artikel ⇒Adj
  - Ich sehe Peter (Korrektheitsproblem)
- Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel ⇒ Adj
  - Große Bedenken (Vollständigkeitsproblem)

1. Es ist schwer, Regeln von Hand zu schreiben, die sowohl korrekt als auch vollständig sind
2. Regeln sind schwer automatisch zu lernen

Adjektiverkennung als Klassifikation

- **Alternative Idee:** Zusammenhang zwischen Mustern und "Adjektivheit" automatisch (maschinell) lernen!
- **Klassifikationsaufgabe**: jedem Eingabe-Ereignis wird eine Klasse (aus kleiner Menge) zugeordnet



Ereignisse
(hier: Worte)

Klassen
(hier: Adj/NAdj)

(Gegenbeispiel: Lemmatisierung)

Einfach(st)e statistische Klassifikation

- **Schritt 1: Training**
- Lese Zusammenhang zwischen Ereignissen und Klassen aus Korpus ab
  - Für jedes Ereignis zählen wir in den Trainingsdaten, wie oft es mit welcher Klasse vorgekommen ist
  - Nötig: Annotierte Korpora

Ereignis	Kl. 1	Kl. 2
Er1	10	20
Er2	2	0
...	...	...

- **Schritt 2: Analyse neuer Daten**
- Jedem Ereignis in neuen Daten wird die häufigste Klasse aus den Trainingdaten zugewiesen

Was ist ein Ereignis? (I)

- Im einfachstes Fall: ein Wort

Trainingskorpus

Ereignis	Ein kleines Haus
Klasse	Nadj Adj Nadj

Modell

Ereignis	Nadj	Adj
ein	1	0
kleines	0	1
Haus	1	0

- Keine gute Idee!
  - Klassifikation eines neuen Satzes:
    - Lexikon!

Ereignis	Ein großes Haus
Klasse	Nadj ?? Nadj

Was ist ein Ereignis? (II)

- Besser: Ereignis ist Kombination von **informativen Mustern** ("**Features**")

Ein kleines Haus

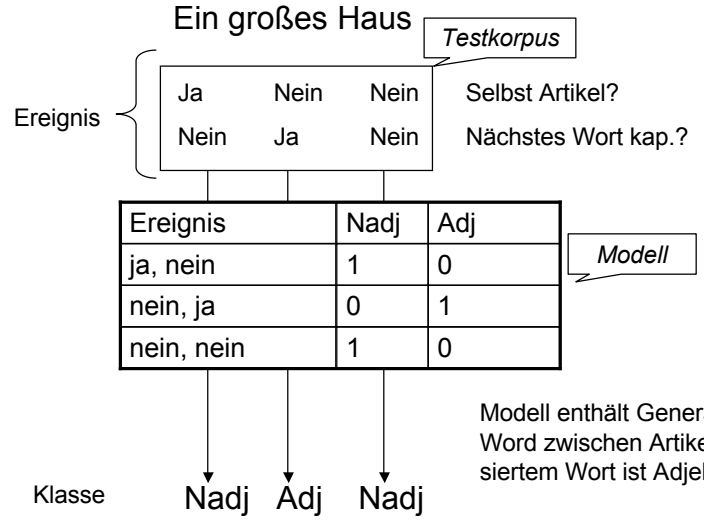
Ereignis	Ja	Nein	Nein	Selbst Artikel?
	Nein	Ja	Nein	Nächstes Wort kap.?
Klasse	Nadj	Adj	Nadj	

Trainingskorpus

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

Modell

Klassifikation eines neuen Satzes



Modell enthält Generalisierung: Word zwischen Artikel und kaptalisiertem Wort ist Adjektiv

Was "bedeutet" das Modell?

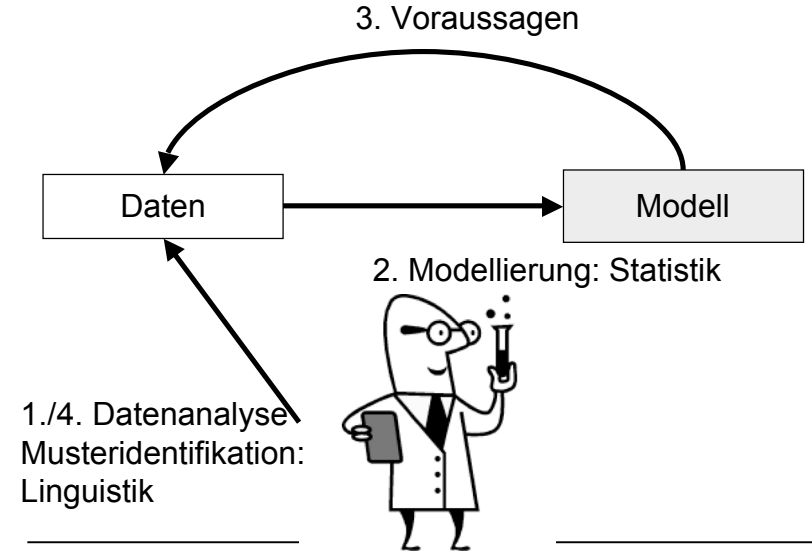
- Features:
  - Selbst Artikel?
  - Nächstes Wort kap.?

Ereignis	Nadj	Adj
ja, nein	1	0
nein, ja	0	1
nein, nein	1	0

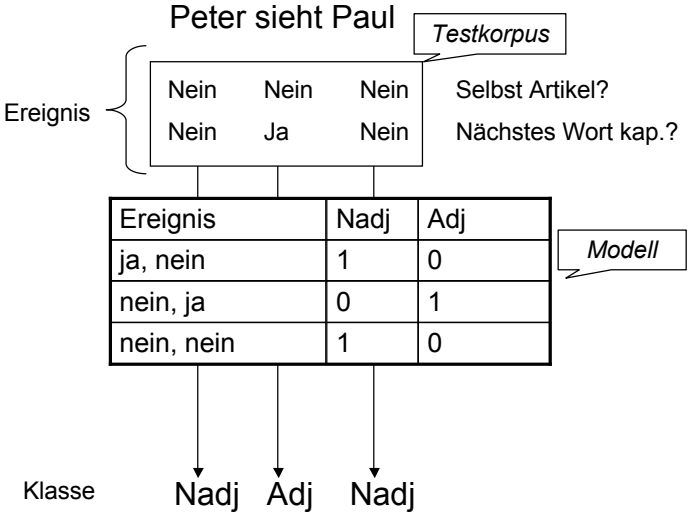
- Zeile 1: Artikel, nächstes Wort nicht kapitalisiert: Kein Adjektiv
- Zeile 2: Kein Artikel, nächstes Wort kapitalisiert: Adjektiv
- Zeile 3: Kein Artikel, nächstes Wort nicht kap.: Kein Adjektiv

- Also doch Regeln: nur eben automatisch gelernte...

Arbeitsteilung



Fehler



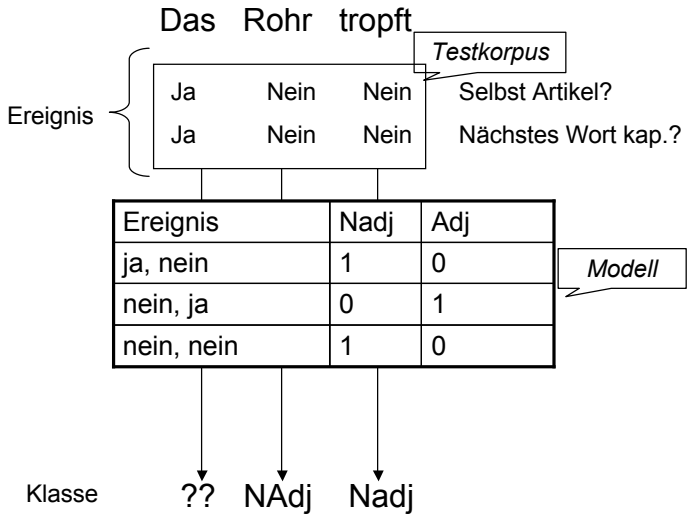
Intelligentere Modelle = mehr Features

- Welche weiteren Muster könnte man ausnutzen, um Adjektive zu identifizieren?
  - Wortendungen (morphologische Information)
    - Adjektive können nur bestimmte Endungen haben
  - Gradpartikel stehen fast immer vor Adjektiven
    - “sehr”, “besonders”, ...
  - Logische Kombination von existierenden Features
    - Voriges Wort Artikel ODER selbst nicht kapitalisiert
  - ...
- Wieso verwendet man nicht einfach alle Features, die einem einfallen?

Größe des Ereignisraumes

- Wieviele Zeilen hat die Modell (Tabelle)?
  - Anzahl möglicher verschiedener Ereignisse
    - **Produkt der Anzahl möglicher Werte** aller Features
      - Beispiel: Selbst Artikel? x Nächstes Wort kapitalisiert = 2 x 2 = 4
      - Lexikalische Features: alleine > 10.000 Were
- Frequenzen in Trainingskorpus werden auf Zeilen (Ereignisse) verteilt
  - Wenn Trainingskorpus < mögliche Ereignisse, sind die meisten Ereignisse ungesehen
    - Modell kann keine Vorhersage machen
- Das “Sparse Data”-Problem

Sparse Data: Beispiel

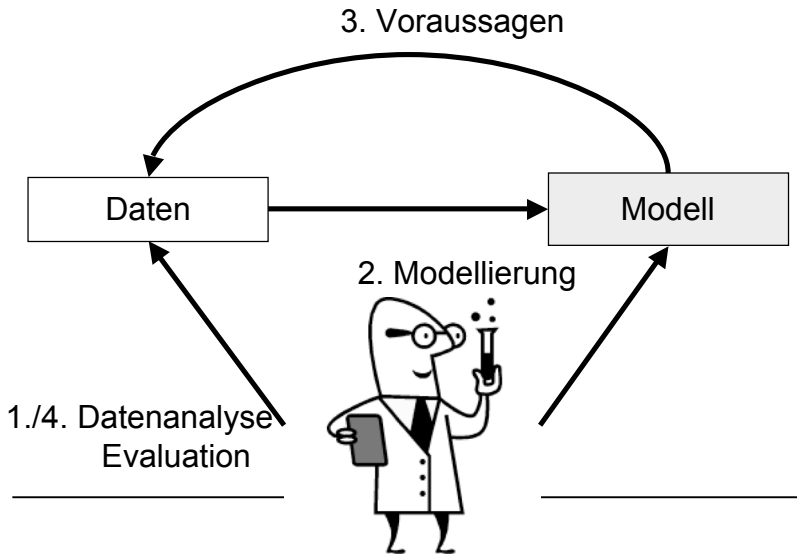


Das Sparse-Data-Dilemma

- Je mehr Features, desto besser die Datenlage für die Entscheidung
- Je mehr Features, auf desto mehr Ereignisse verteilen sich die Trainingsdaten

- Richtlinien für die Identifikation von Features:
  - Wenige gute Features sind besser als viele mittelmäßige
  - Bevorzuge Features mit wenigen Werten

Empirische Modellierung



Evaluation

- Versucht die Frage zu beantworten, wie gut ein Modell ist
  - Es gibt verschiedene Maße für Qualität
- Klassifikation wird oft anhand einer **Konfusionsmatrix** evaluiert

Konfusionsmatrix für Klasse X

Modell \ echt	Echtes X	Echtes nicht-X
Als X klassifiziert	gut	schlecht
Als nicht-X klass.	schlecht	gut

- **Klassenspezifische Evaluation**
- Setzt „echte“ Klasse von Beispielen zu vom Modell zugewiesenen Klassen in Beziehung
  - Zellen enthalten Frequenzen der Ereignisse
  - Richtig klassifiziert: entweder (X,X) oder (~X,~X).

Beispielevaluation

Modell \ echt	Echtes X	Echtes nicht-X
Als X klassifiziert	10	80
Als nicht-X klass.	10	900

Beispielevaluation

Modell \ echt	Echtes X	Echtes nicht-X
Als X klassifiziert	10	80
Als nicht-X klass.	10	900

- Beobachtung 1: Gesamtfrequenzen von Klassen sind als Summen von Spalten/Zeilen ablesbar
  - Im Goldstandard gab es 10+10=20 X, 980 ~X
  - In der Ausgabe des Modells gab es 10+80=90X, 910 ~N

Accuracy / Error

- Einfachstmögliche Evaluation
- Accuracy = (# richtige Ereignisse) / (# alle Ereignisse)
  $$= \frac{(\#\text{ Ereignisse: Vorhersage = Gold-Klasse})}{(\#\text{ alle Ereignisse})}$$
- Error = (1 - Accuracy)
- Evaluiert gesamte Klassifikation (alle Klassen!)

Beispielevaluation

Modell \ echt	Echtes X	Echtes nicht-X
Als X klassifiziert	10	80
Als nicht-X klass.	10	900

- Accuracy = (900+10) / (10+80+10+900) = 910/1000 = 91%
- Error = 9%

Probleme

- Problem 1: Macht keinen Unterschied zwischen (Gold-)Klassen (Adj vs. Nadj)
  - Wieso unterscheiden?
    - In der CL sind die interessanten Klassen oft klein
    - In Gesamtevaluation geht Qualität kleiner Klassen unter
- Lösung: Klassenspezifische Accuracy / Error
    - 10 / 20 korrekt für Adj: 50%
    - 900 / 980 korrekt für Nadj: 91.8%
  - Interpretation nicht ganz einfach

Verschiedene Fehlerarten

- Problem 2: Es gibt zwei Arten von Fehlern
  - Korrektheitsfehler von X ("false positive"):
    - Ein Ereignis ist kein X, wird aber vom Modell als X klassifiziert
  - Vollständigkeitsfehler von X ("false negative"):
    - Ein Ereignis ist ein X, wird aber vom Modell nicht als X klassifiziert
- Wieso unterscheiden?
  - Fehler können unterschiedlich wichtig sein
    - Erkennung von Adjektiven
      - Korrektheitsfehler der Klasse Adj sind meist schlimmer als Vollständigkeitsfehler (Weiterverarbeitung)
    - Qualitätskontrolle von Hardware-Komponenten
      - Vollständigkeitsfehler der Klasse "defekt" schlimm

Fehlerarten in der Konfusionsmatrix

Modell \ echt	Echtes X	Echtes nicht-X
Als X klassifiziert	gut (true positive)	schlecht (Korrektheitsfehler, false positive)
Als nicht-X klass.	schlecht (Vollst.fehler, false negative)	gut (true negative)

- Die beiden Fehlerarten werden oft einzeln gemessen: Precision und Recall

Recall

Modell \ echt	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Welcher Anteil der echten X wurde als X klassifiziert? (Vollständigkeit)
- Werte zwischen 0 und 1 (höher = besser)

Recall für Klasse Adj

Modell \ echt	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Hier: 10/(10+10) = 0.5
- Interpretation: Die Hälfte aller echten Adjektive wurde durch das Modell richtig erkannt

Precision

Modell \ echt	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Welcher Anteil der als X klassifizierten Eriegnisse ist wirklich ein X? (Korrektheit)
- Werte zwischen 0 und 1 (höher = besser)

Precision für Klasse Adj

Modell \ echt	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Hier: 10 / (10+80) = 11%
- Interpretation: Wenn das Modell behauptet, ein Ereignis sei ein Adj, ist das nur 11% der Fälle wahr

Precision und Recall

- Gemeinsame Betrachtung von Precision und Recall characterisiert die "Strategie" eines Modells
  - Hohe Precision, hoher Recall: fast perfekte Klassifikation
  - Niedrige Precision, niedriger Recall: sehr schlechte Klassifikation
  - Hohe Precision, niedriger Recall: "vorsichtiges Modell"
    - Findet nicht alle Ereignisse der Klasse X
    - Klassifiziert fast keine Nicht-Xe als X
  - Niedrige Precision, hoher Recall: "mutiges Modell"
    - Findet fast alle Ereignisse der Klasse X
    - Klassifiziert auch Nicht-Xe als X

Extremfälle..und die Kombination

- Extremfälle
  - Modell klassifiziert alles als X
    - Recall 100%, Precision sehr niedrig
  - Modell klassifiziert nichts als X
    - Recall 0%, Precision nicht definiert (0/0)

- F-Score: Kombination aus P und R:
  - Ein Maß für „Gesamtgüte" der Klassifikation
    - Werte zw. 0 und 1 (höher = besser)
    - Bevorzugt „true positives"

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

F-Score für Klasse Adj

Modell / echt	Echtes Adj	Echtes Nadj
Als Adj klassifiziert	10	80
Als Nadj klass.	10	900

- Precision: 10 / (10+80) = 0.11
- Recall: 10 / (10+10) = 0.5
- F-Score: (2 * 0.5 * 0.11) / (0.5 + 0.11) = 0.18