



Einführung in die Computerlinguistik Maschinelles Lernen

Dr. Benjamin Roth & Annemarie Friedrich
Centrum für Informations- und Sprachverarbeitung
LMU München
WS 2016/2017

Maschinelles Lernen

- Einführung (heute)
 - Attribute
 - Klassifikation vs. Clustering
 - Evaluation von Algorithmen
- Algorithmen
 - K-means
 - Entscheidungsbäume
 - Naive Bayes (& Sprachmodelle)

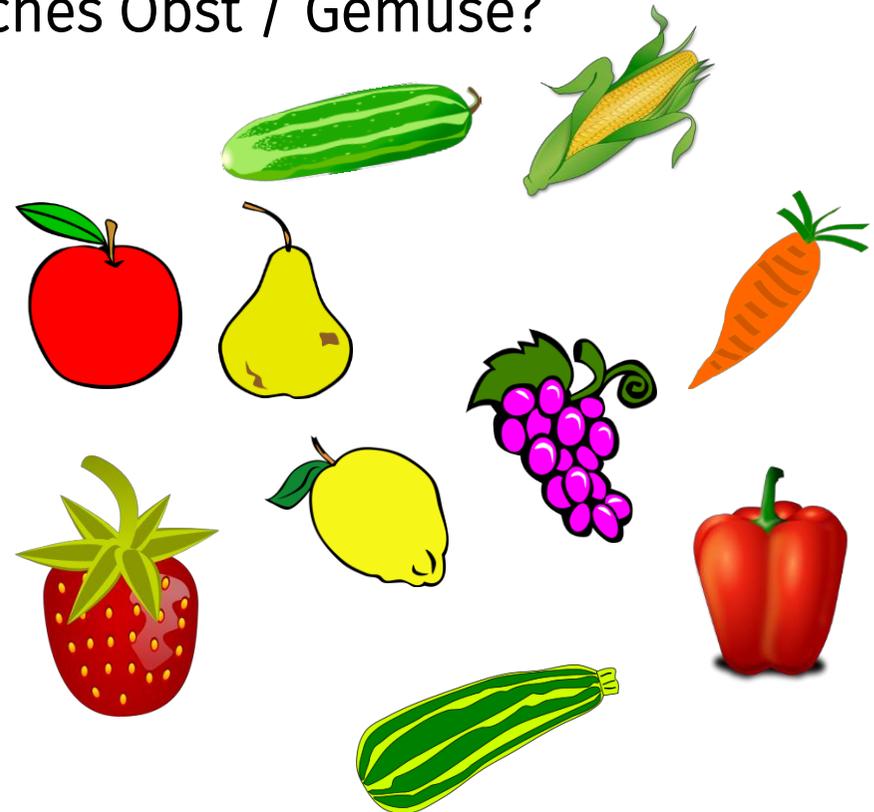
Klassifikation

Mustererkennung (pattern recognition)



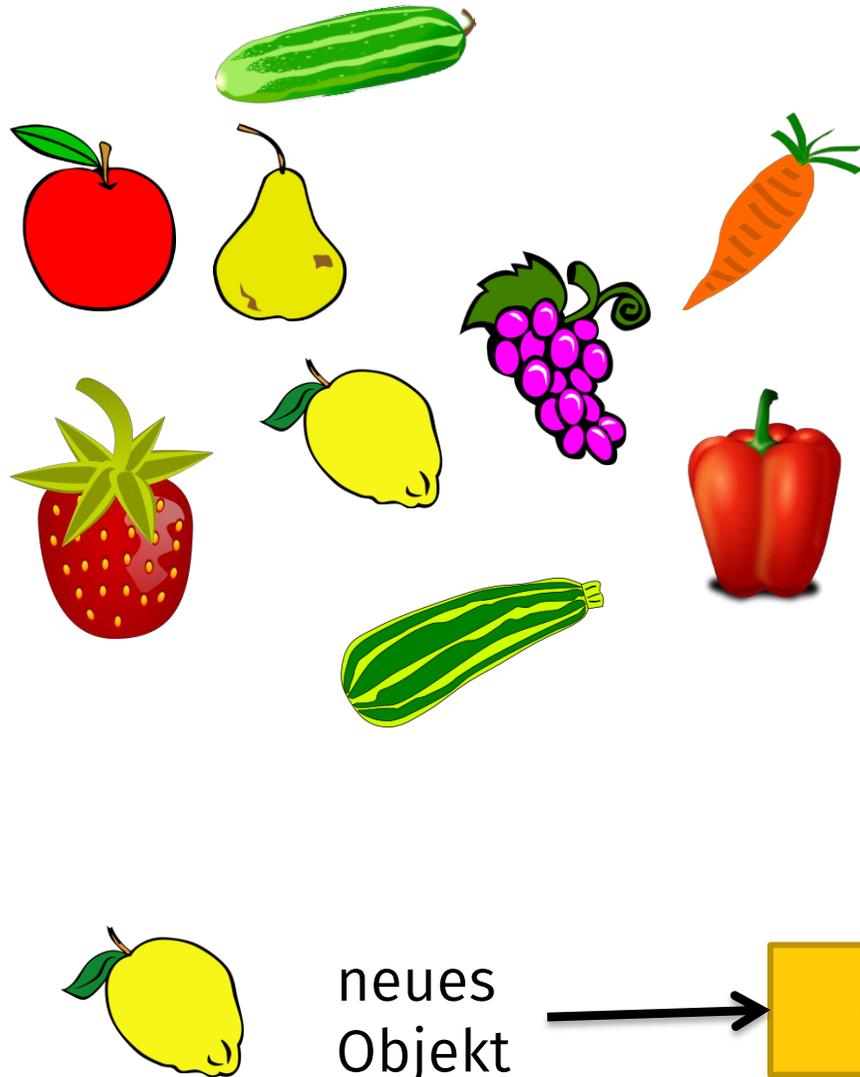
Kamera

Welches Obst / Gemüse?



Was sind mögliche Erkennungsmerkmale?

Mustererkennung (pattern recognition)



Merkmale/Attribute:

- Farbe
- Größe
- Form
- ...

Beispiele:
Attribute/Werte + richtige Klasse

Algorithmus
(Maschinelles Lernen)

neues
Objekt

Classifier

Zitrone (80%)
Birne (20%)

Maschinelles Lernen

- **Prinzip:** lerne aus Mustern von linguistischer Information ...
 - Sprachmodelle (welche Sätze sind okay in einer natürlichen Sprache?)
 - Parse-Bäume (finde automatisch den besten Baum für einen Satz)
 - Übersetzungsmodelle
 - Textkategorien (Text → Nachrichten, Prosa, Gedicht, Spam, e-Mail, ... ?)
- **Grundsätzlich:** je zahlreicher und informativer die Daten, desto besser

Klassifikation

- **überwachtes** Lernen (supervised)
- Algorithmus lernt Funktion $x \rightarrow y$ aus annotierten Daten $\langle x_i = \text{Merkmale (features)}, y_i = \text{Klasse} \rangle$

Beispiel:

Klassen: e-Mail vs. Spam

Merkmale: Wörter der Mail (lemmatisiert)

Klassifikator lernt z.B., dass Vorkommen von „hot girl“ eher auf Spam hinweist.

Training und Evaluation

- **Trainingsdaten** → werden benutzt, um den Klassifikator zu trainieren (zu erstellen)
- **Testdaten** → Klassifikator wird auf diese Daten angewendet und Ergebnis der Vorhersage wird mit den tatsächlichen Labels verglichen
- oft auch noch: **Entwicklungsdaten** (development set)
- **keine Überschneidung zwischen diesen Datensets!!**

Evaluation

- **Testdaten:** Auswertung des Klassifikationsergebnisses
 - hier: 100 Test-Instanzen

automatisch zugewiesene
Klasse

		e-Mail	Spam	
tatsächliche Klasse	e-Mail	58	21	← richtig
	Spam	9	12	← falsch

- **Accuracy:** % der Testinstanzen, die richtig klassifiziert wurden
 - Accuracy im Beispiel?

Precision

- Wie sehr kann ich dem Klassifikator trauen, wenn er eine Mail als Spam klassifiziert?
- **Precision** wird pro Klasse berechnet

- $Precision(K) = \frac{\text{richtig als } K \text{ klassifizierte Instanzen}}{\text{alle als } K \text{ klassifizierten Instanzen}}$

- Precision(e-Mail)?

- Precision(Spam)? automatisch zugewiesene Klasse

		e-Mail	Spam
tatsächliche Klasse	e-Mail	58	21
	Spam	9	12

Recall

- Wie viele der tatsächlichen Spam-Mails findet der Klassifikator?
- **Recall** wird pro Klasse berechnet

- $Recall(K) = \frac{\text{richtig als } K \text{ klassifizierte Instanzen}}{\text{alle Instanzen, die tatsächlich } K \text{ sind}}$

- Recall(e-Mail)?

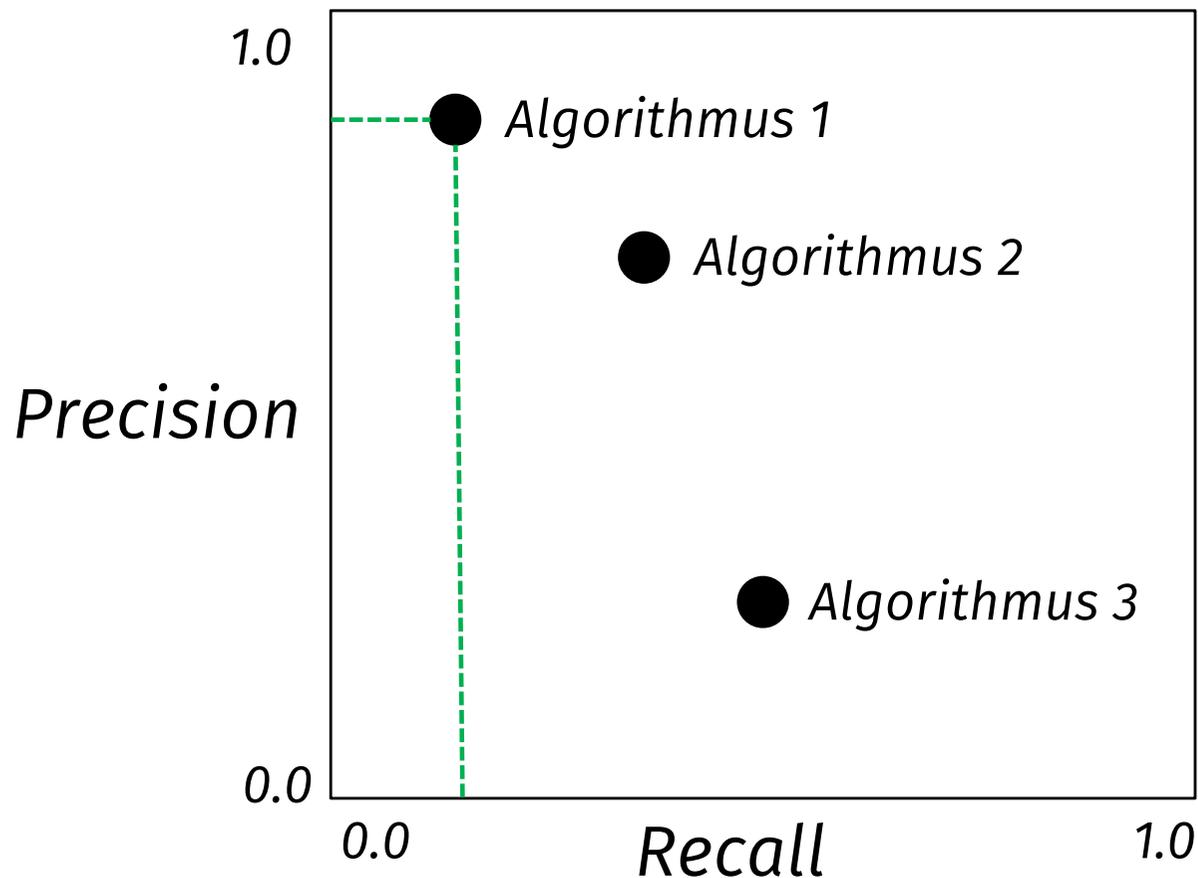
- Recall(Spam)?

automatisch zugewiesene
Klasse

		e-Mail	Spam
tatsächliche Klasse	e-Mail	58	21
	Spam	9	12

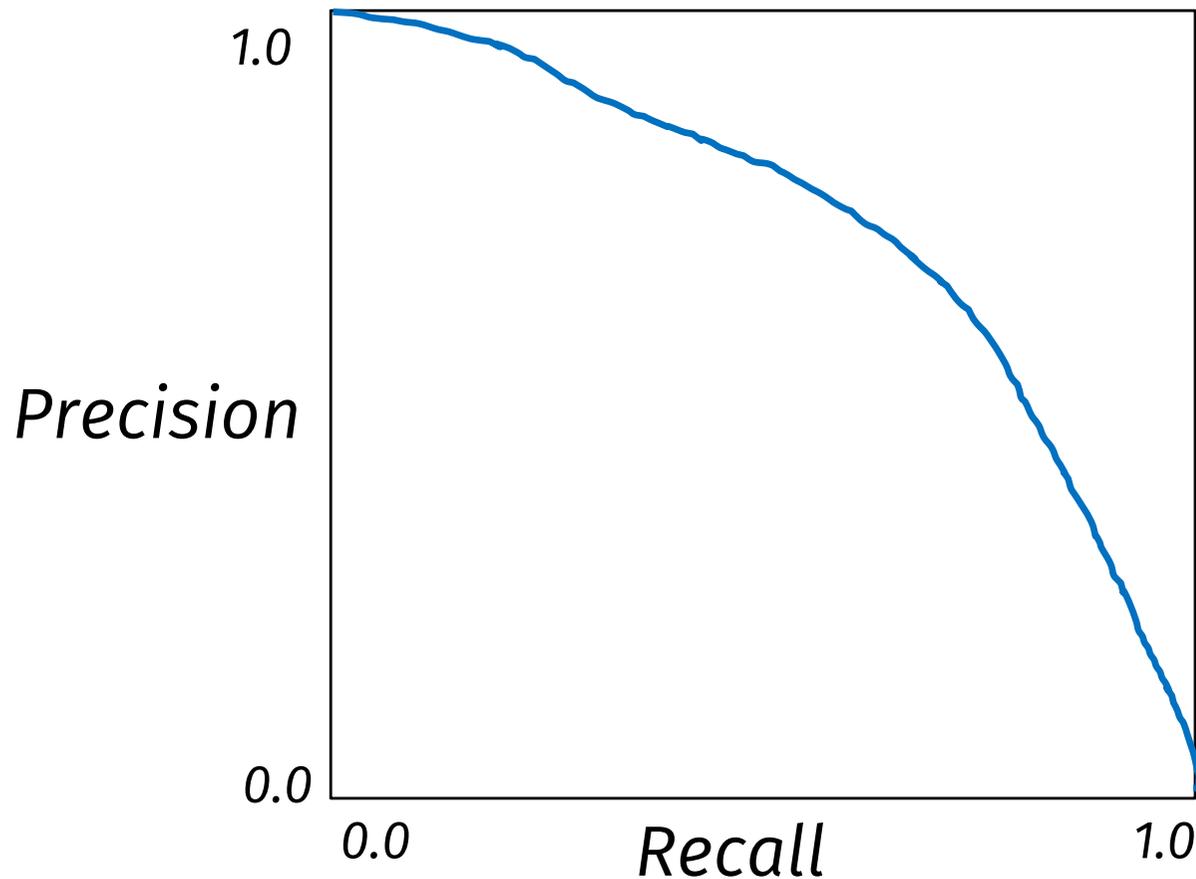
Precision-Recall Trade-Off

- verschiedene Algorithmen oder verschiedene Einstellungen resultieren in unterschiedlichen Werten für Precision + Recall



Precision-Recall Trade-Off

- verschiedene Algorithmen oder verschiedene Einstellungen resultieren in unterschiedlichen Werten für Precision + Recall



F1-Measure

- ein Score, der zur Evaluation genutzt werden kann
- harmonischer Mittelwert von Precision und Recall

- $$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

- Alternativen: Precision und Recall können auch anders gewichtet werden (F_β)