pagina
Publikations
technologien

VERSCHLAGWORTUNG VON E-BOOKS MIT HILFE VON MACHINE LEARNING-VERFAHREN

16.11.2018 / Markupforum Stuttgart

Kai Weber / @fruehlingstag

Kai Weber

- **Ausbildung:** Buchhändler, Buchhandlung Reuffel, Koblenz
- **Studium:**
  - [Magister Artium:] Allg. u. Vergl. Literaturwissenschaft, Buchwissenschaft, Bohemistik und Deutsch als Fremdsprache, Johannes-Gutenberg-Universität, Mainz
  - [Master of Computer Science:] Fernstudium Informatik, Hochschule Trier
- **Beruf:**
  - e-Publishing / e-Business, Ernst Reinhardt Verlag, München
  - Projektmanager E-Book, Koch, Neff & Volckmar GmbH, Stuttgart
  - Softwareentwickler Digitale Verlagsauslieferung, Koch, Neff & Oetinger Verlagsauslieferung GmbH, Stuttgart
  - Softwareentwickler Digital Humanities, pagina GmbH Publikationstechnologien, Tübingen

Entfremdung ist Folge von Desinteresse an der Technologie, die uns umgibt

»Modern man lives isolated in his artificial environment, not because the artificial is evil as such, but because of his lack of comprehension of the forces which make it work – of the principles which relate his gadgets to the forces of nature, to the universal order. It is not central heating which makes his existence 'unnatural', but his refusal to take an interest in the principles behind it. By being entirely dependent on science, yet closing his mind to it, he leads the life of an urban barbarian.«

Arthur Koestler: The Act of Creation

(„Der moderne Mensch lebt isoliert in seiner künstlichen Umgebung, nicht weil das Künstliche ein Übel an sich wäre, sondern weil er die Kräfte, die das Künstliche funktionieren lassen, nicht versteht – die Prinzipien, welche seine Geräte mit den Naturkräften, mit der universellen Ordnung, verbinden. Es ist nicht die Zentralheizung, die seine Existenz „unnatürlich" werden lässt, sondern seine Weigerung, sich für die dahinterstehenden Prinzipien zu interessieren. Dadurch, dass er völlig von den Naturwissenschaften abhängig ist und zugleich seinen Geist vor ihnen verschließt, lebt er das Leben eines urbanen Barbaren.")

pagina
Publikations
technologien

… also sollten wir uns für Maschinenlernen und künstliche Intelligenz interessieren,
um uns vor der Entfremdung zu bewahren!

pagina
Publikations
technologien

Ziele dieses Vortrags

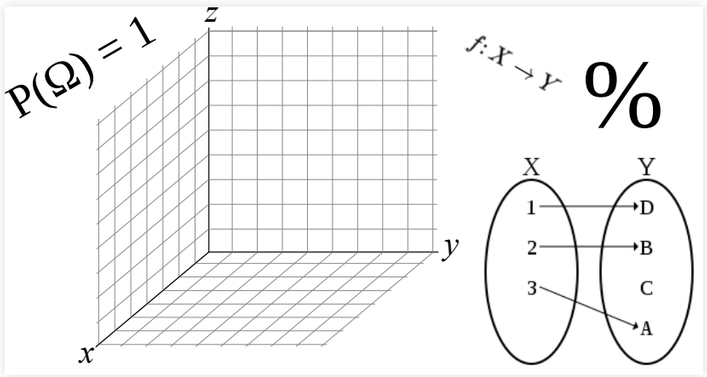
- Vorstellen eines Beispiels zur Anwendung von Maschinenlernen im Verlagskontext
- Aufzeigen einer typischen Datenvorbereitung bei Texten in natürlicher Sprache
- Erklären von zwei für die Textklassifikation gut geeigneten Maschinenlernalgorithmen:
  - konkret, mit Beispielzahlen
  - ohne mathematische Formeln
- Auch Nichtinformatikerinnen und Nichtmathematikern eine Vorstellung davon vermitteln, wie Maschinenlernen funktioniert

Voraussetzungen

Maschinenlernen verstehen ohne Informatik und Mathematik, echt jetzt?

Nun ja, eine Vorstellung von … ist schon nötig oder zumindest hilfreich:

- Wahrscheinlichkeiten
- Prozentrechnung
- zwei- und dreidimensionale Koordinatensysteme
- Funktionsbegriff (Abbildung Eingabe- auf Ausgabewert)



Ziemlich niedrigschwellig, nicht wahr?

Ziel meines Maschinenlern-Vorhabens

- Deutschsprachige E-Books (EPUB) automatisch auf VLB-Warengruppe und/oder Thema-Kategorie abbilden

VLB-Warengruppen

- codiert als dreistellige Ziffer
- verpflichtend im Verzeichnis lieferbarer Bücher (VLB)
- jedes Buch wird genau einer WG zugeordnet
- Nummerkreise gruppieren Inhalte

1	**Belletristik**
110	Erzählende Literatur
111	Hauptwerk vor 1945
112	Gegenwartsliteratur (ab 1945)
113	Historische Romane und Erzählungen
114	Märchen, Sagen, Legenden
115	Anthologien
116	Romanhafte Biographien
117	Briefe, Tagebücher
118	Essays, Feuilleton, Literaturkritik, Interview
119	Aphorismen
120	Spannung
121	Krimis, Thriller, Spionage
122	Historische Kriminalromane
123	Horror
130	Science Fiction, Fantasy
131	Science Fiction
132	Fantasy
133	Fantastische Literatur

6	**Naturwissenschaften, Medizin, Informatik, Technik**
610	Naturwissenschaften allgemein
620	Mathematik
621	Allgemeines, Lexika
622	Grundlagen
623	Arithmetik, Algebra
624	Geometrie
626	Analysis
627	Wahrscheinlichkeitstheorie, Stochastik, Mathematische Statistik
629	Sonstiges
630	Informatik, EDV
631	Allgemeines, Lexika
632	Informatik
633	Programmiersprachen
634	Betriebssysteme, Benutzeroberflächen
635	Anwendungs-Software
636	Datenkommunikation, Netzwerke
637	Internet
638	Hardware
639	Sonstiges

Thema-Klassifikation

Code	Überschrift	?
D	Biografien, Literatur, Literaturwissenschaft…	
DN	Biografien und Sachliteratur… Bei DN* Codes möglichst auch näheren Inhalt mit einem zusätzlichen Code bestimmen, z.B. SFH „Golf" mit DNBS. S. auch: FC Biografischer Roman…	✻
DNB	Biografien: allgemein…	
DNBF	Biografien: Kunst und Unterhaltung… Hier einzuordnen: Prominenten-Biografien	✻
DNBF1	Autobiografien: Kunst und Unterhaltung Hier einzuordnen: Autobiografien von Prominenten	✻

- internationale Klassifikation (EDItEUR.org)
- alphanumerische Codes von variabler Länge
- jedem Buch können mehrere Codes zugewiesen werden


ePUB
Klassifikator
WG-Metadaten

Maschinenlernen: kurzer schematischer Ablauf

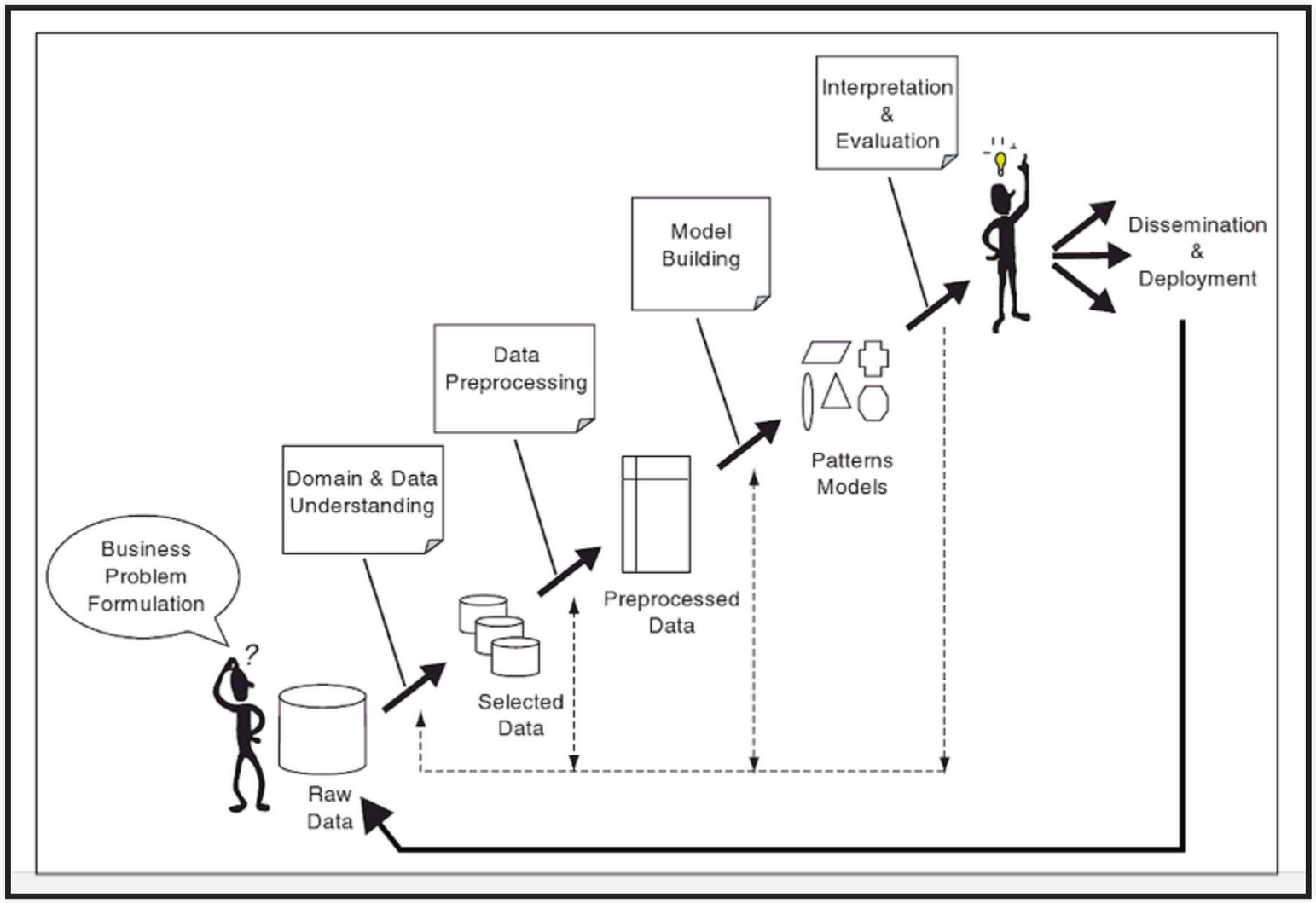
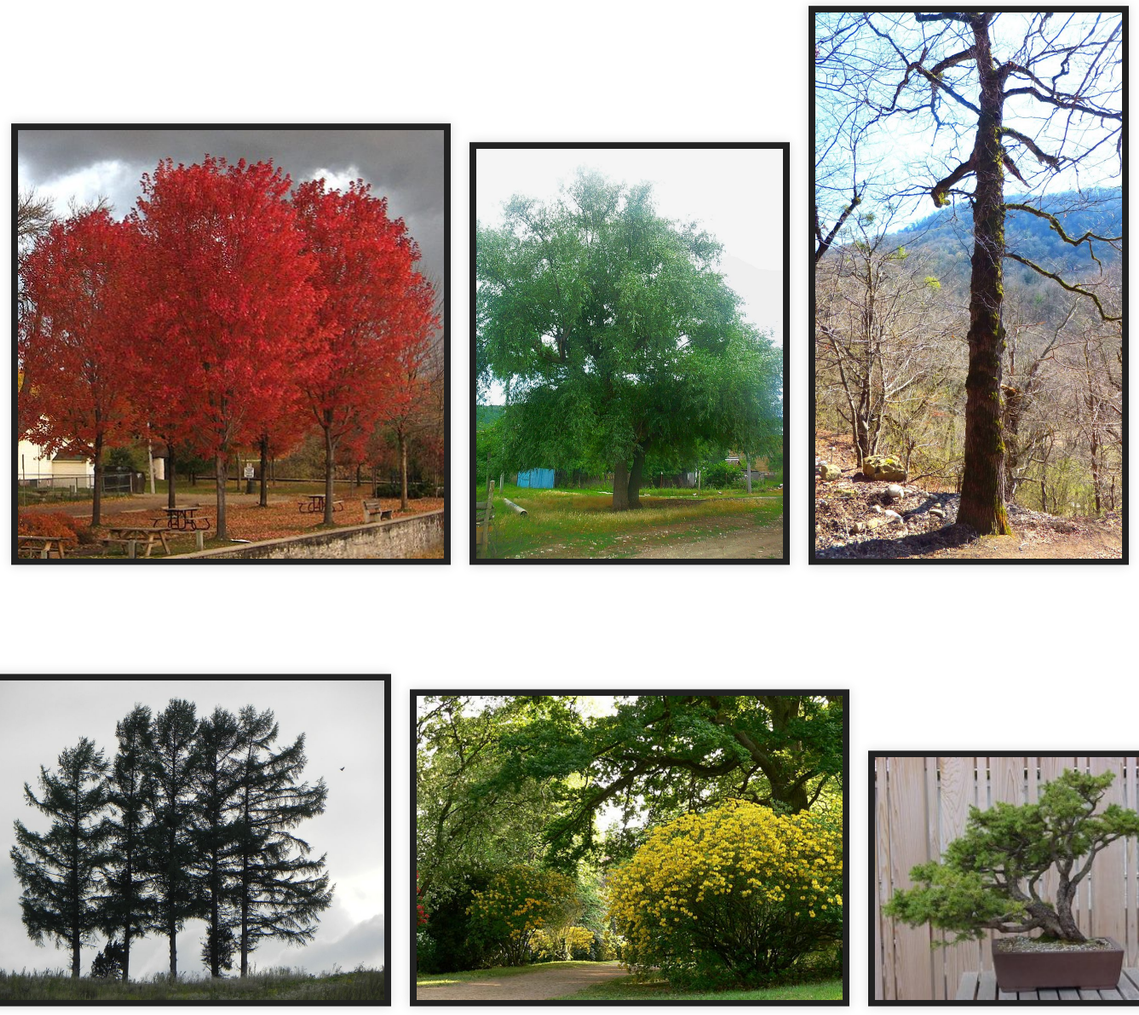


Abb. aus Pavel Brazdil et al.: Metalearning. Applications to Data Mining, Berlin/Heidelberg 2009, S. 3

Modellbegriff

Ein Modell repräsentiert den gegenwärtigen Wissensstand, der durch Beobachtung von Beispielen erreicht wurde.



Bildnachweise: (1) Alethe – Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=5076388 * (2) matin fattahi, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=54714067 * (3) **Дмитрий** С**кляренко**, CC BY 3.0 https://commons.wikimedia.org/wiki/File:-_panoramio_(6749).jpg * (4) pakku, CC BY 3.0 * (5) Arieswings – Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=49034785 * (6) CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=559431

Ausgangsmaterial

- ca. 6900 E-Books aus ca. 70 Verlagen mit bekannten, von Verlagen vergebenen VLB-Warengruppen (ca. 150 Warengruppen kommen im Korpus vor)
- ca. 1300 E-Books verfügen auch über Thema-Klassifikationen

Datenvorverarbeitung (1)

- Zerlegen der E-Book-Texte in Sätze, Wörter, Lemmata
- Wortartenerkennung („Part-of-Speech-Tagging“)
- Zwischenspeicherung des Gesamtwortschatzes in effizienter Datenstruktur („Lucene-Index“) – das erleichtert spätere Berechnungen der „Wichtigkeit“ einzelner Wörter

Datenvorverarbeitung (1a)

Text aus HTML extrahieren:

```
...
<head>
    <link href="../Styles/style001.css" rel="stylesheet" type="text/css"/>
    <title>Effi Briest</title>
</head>
<body>
    <h1 id="heading_id_10"><span class="bold">Zehntes Kapitel</span></h1>
    <p>Innstetten war erst sechs Uhr früh von Varzin zurückgekommen und hatte sich, Rollos Liebkosungen abwehrend,
    <p>»Die gnädige Frau schläft noch.«</p>
    <p>»Aber es ist ja schon spät. Ist etwas passiert?«</p>
    <p>»Ich weiß es nicht; ich weiß nur, Johanna hat die Nacht über im Zimmer der gnädigen Frau schlafen müssen.«</
    <p>»Nun, dann schicke Johanna.«</p>
    [...]
</body>
```

```
Zehntes Kapitel

   Innstetten war erst sechs Uhr früh von Varzin zurückgekommen und hatte sich, Rollos Liebkosungen abwehrend, so leise wie m
   Er machte sich's hier bequem und duldete nur, dass ihn Friedrich mit einer Reisedecke zudeckte. »Wecke mich um neun.« Und
   geweckt worden. Er stand rasch auf und sagte: »Bringe das Frühstück.«
   »Die gnädige Frau schläft noch.«
   »Aber es ist ja schon spät. Ist etwas passiert?«
   »Ich weiß es nicht; ich weiß nur, Johanna hat die Nacht über im Zimmer der gnädigen Frau schlafen müssen.«
   »Nun, dann schicke Johanna.«
```

pagina Publikations technologien

Datenvorverarbeitung (1b)

Zerlegung des Textes in Sätze, dabei möglichst Punkte hinter Abkürzungen (Dr., bzw., usw.) nicht mit einem Satzende verwechseln.

1	Zehntes Kapitel
2	Innstetten war erst sechs Uhr früh von Varzin zurückgekommen und hatte sich, Rollos Liebkosungen abwehrend, so leise wie möglich in sein Zimmer zurückgezogen.
3	Er machte sich’s hier bequem und duldete nur, dass ihn Friedrich mit einer Reisedecke zudeckte.
4	»Wecke mich um neun.«
5	Und um diese Stunde war er denn auch geweckt worden.

Datenvorverarbeitung (1c)

Zerlegung des Textes in Wörter (Tokenisierung)

1	Zehntes
2	Kapitel
3	Innstetten
4	war
5	erst
6	sechs
7	Uhr
8	früh
9	von
10	Varzin

Rückführung von flektierten Wörtern auf ihre Wörterbuchform (Lemmatisierung)

1	zehnt
2	Kapitel
3	Innstetten
4	sein
5	erst
6	sechs
7	Uhr
8	früh
9	von
10	Varzin

Ermittlung der Wortarten (Part-of-Speech-Tagging)

1	Nomen (NN) [fälschlicherweise!]
2	Nomen (NN)
3	Eigenname (NE)
4	finites Hilfsverb (VAFIN)
5	Adverb (ADV)
6	Kardinalzahl (CARD)
7	Nomen (NN)
8	determinierendes Adjektiv (ADJD)
9	Präposition (APPR)
10	Eigenname (NE)

Datenvorverarbeitung (2)

- Erstellen sog. Wortvektoren (Tabellen) aus dem Gesamtvokabular der E-Books (ca. 3.000.000 „Lemmata“)
  - Ausfiltern sehr häufig vorkommender Wörter („Stopwords“)
  - Auswahl der ca. 15.000-25.000 wichtigsten Wörter
  - evtl. auch Ermittlung von Oberbegriffen (Hyperonymen)

E-Book	Vorkommen des Wortes …					WG
	…	einstimmen	energetisieren	entspringen	erhebend	…	
#1		ja	nein	nein	nein		112
#2		nein	nein	nein	ja		260
#3		ja	nein	nein	nein		973

Datenvorverarbeitung (3)

- Berechnung der Häufigkeit bestimmer grammatischer oder statistischer Phänomene pro E-Book, z. B.
  - Häufigkeit von Adjektiven, Verb(form)en, Pronomina, usw.
  - Häufigkeit von Passivsätzen
  - durchschnittliche Wort- und Satzlänge
  - Schachtelungstiefe des Inhaltsverzeichnisses
  - Anzahl enthaltener Abbildungen

E-Book	…	% Adjektive	% Personalpronomen	Ø Wortlänge	IHVZ-Tiefe	Anz. Abb.	WG
#1	…	6,28	10,48	5,08	1	14	112
#2	…	5,57	11,41	4,68	1	2	260
#3	…	7,58	6,01	5,70	1	4	973

Datenvorverarbeitung (4)

Alle bisher ermittelten Daten werden zu einem „Attributvektor“ (= einer Tabelle aller Dokumenteigenschaften) vereinigt.

E-Book	Wortvektor				sonstige Eigenschaften			WG
	entspringend	erhebend	festfahren	...	% Adjektive	Anz. Abb.	...	
#1	nein	nein	ja	...	6,28	14	...	112
#2	nein	ja	nein	...	5,57	2	...	260
#3	nein	nein	nein	...	7,58	4	...	973

Ist der Attributvektor zu groß (d. h. zu viele Spalten in der Tabelle), können die „informationshaltigsten“ Attribute berechnet werden und nur mit diesen weitergearbeitet werden.

pagina
Publikations
technologien

Textklassifikation mit Naïve Bayes

- Lernverfahren auf Basis von Wahrscheinlichkeitsrechnung
- benannt nach dem englischen Mathematiker Thomas Bayes, dessen *Satz von Bayes* die theoretische Grundlage für das Verfahren liefert
- wird bereits seit den 1950er Jahren erforscht
- wird in der Textklassifikation schon lange erfolgreich verwendet (z. B. in der Spamklassifikation)

Benötigte Wahrscheinlichkeitswerte für Naïve Bayes

1. A-priori-Wahrscheinlichkeit: „Wie wahrscheinlich ist es, dass ein E-Book zu einer bestimmten Warengruppe gehört, wenn ich den Inhalt des E-Books gar nicht beachte?“
2. Bedingte Wahrscheinlichkeiten bereits bekannter E-Books: „Ich habe 283 E-Books, die zur Warengruppe 260 gehören und von denen ein gewisser Prozentsatz das Wort *Meer*, ein anderer Prozentsatz das Wort *Insel* enthält. Wie wahrscheinlich ist es, dass eines der bekannten E-Book der Warengruppe 260 diese Wörter enthält?
3. A-posteriori-Wahrscheinlichkeit: Ich habe ein neues E-Book erhalten, von dem ich nicht weiß, zu welcher Warengruppe es gehört. Ich habe gesehen, dass das Buch die Wörter *Brief*, *sehen* und *stark* enthält. Zu welcher Warengruppe gehört das Buch am wahrscheinlichsten?

Mit dem *Satz von Bayes* lässt sich die A-posteriori-Wahrscheinlichkeit aus den beiden erstgenannten Wahrscheinlichkeiten berechnen.

A-priori-Wahrscheinlichkeit

pagina Publikations technologien

Zum Beispiel die zehn häufigsten Warengruppen in meinem Trainingskorpus:

WG-Code	Warengruppentext	Anteil im Korpus	Anteil im KNV-Katalog
112	erzählende Gegenwartsliteratur (ab 1945)	17,90%	10,12% (mit WG 110)
121	Krimis, Thriller, Spionage	8,30%	3,41%
260	Jugendbücher ab 12 Jahre	8,00%	0,47%
250	Kinderbücher bis 11 Jahre	4,06%	0,59%
132	Fantasy	3,89%	1,01%
481	Ratgeber Lebensführung, persönliche Entwicklung	3,72%	0,73%
973	Sachbuch: Gesellschaft	3,11%	0,42%
182	Manga	3,03%	0,03%
185	Humor	2,08%	0,23%
693	Fachbuch Medizin: klinische Fächer	1,60%	1,26%

Vorkommenswahrscheinlichkeiten von Attributen pro Klasse

(Wird in Trainingsphase berechnet)

Attribut	Wahrscheinlichkeit des Auftretens in Warengruppe...					
	112	**113**	**121**	**481**	**973**	**...**
angenehm	0,1944	0,3438	0,1328	0,3953	0,035	...
ankommen	0,4662	0,2188	0,4792	0,4128	0,2098	...
anmutig	0,0519	0,1094	0,0104	0,0174	0,007	...

Attribut	Mittelwert in Warengruppe...					
	112	**113**	**121**	**481**	**973**	**...**
Tiefe des IHVZ	1,4324	1,5	1,3958	2,0465	1,7692	...
Wortlänge	5,0921	5,1434	5,14	5,4322	5,7724	...
Anz. Personalpron.	0,0893	0,0854	0,0869	0,0758	0,0462	...

A-posteriori-Wahrscheinlichkeit

(wird in Evaluations- und Anwendungsphase berechnet)

Eingabe:

E-Book	Wortvektor				sonstige Eigenschaften			WG
	entspringend	erhebend	festfahren	...	% Adjektive	Anz. Abb.	...	
#4	nein	nein	ja	...	8,96	12	...	?

Ausgabe:

E-Book	Wahrscheinlichkeit für Warengruppe ...					
	112	113	121	481	973	...
#4	0,25	0,23	0,11	0,05	0,01	...

Wieso naiv?

- *Satz von Bayes* gilt nur, wenn Attribute (Dokumenteigenschaften) statistisch unabhängig voneinander sind
- Die Wahrscheinlichkeit eines Attributs (eines Wortes) darf nur von der Dokumentklasse (Warengruppe) abhängig sein
- Beispiel: Die Wahrscheinlichkeit des Vorkommens des Wortes *Meer* in Romanen (WG 112) sei 0,2. Die Wahrscheinlichkeit des Wortes *Insel* in Romanen sei 0,18. Diese Wahrscheinlichkeiten müssen auch dann gelten, wenn man weiß, dass das jeweils andere Wort ebenfalls im Text vorkommt: Wenn ich weiß, dass *Meer* in einem Roman vorkommt, muss die Wahrscheinlichkeit für das Vorkommen von *Insel* immer noch 0,18 betragen und nicht etwa 0,34.
- Diese Annahme ist offensichtlich naiv: *Meer* und *Insel* korrelieren sehr wahrscheinlich miteinander und hängen nicht allein von der Textklasse ab.
- Das Tolle: Naïve Bayes funktioniert in der Praxis trotzdem gut, selbst wenn die theoretische Grundbedingung missachtet wird.

pagina Publikations technologien

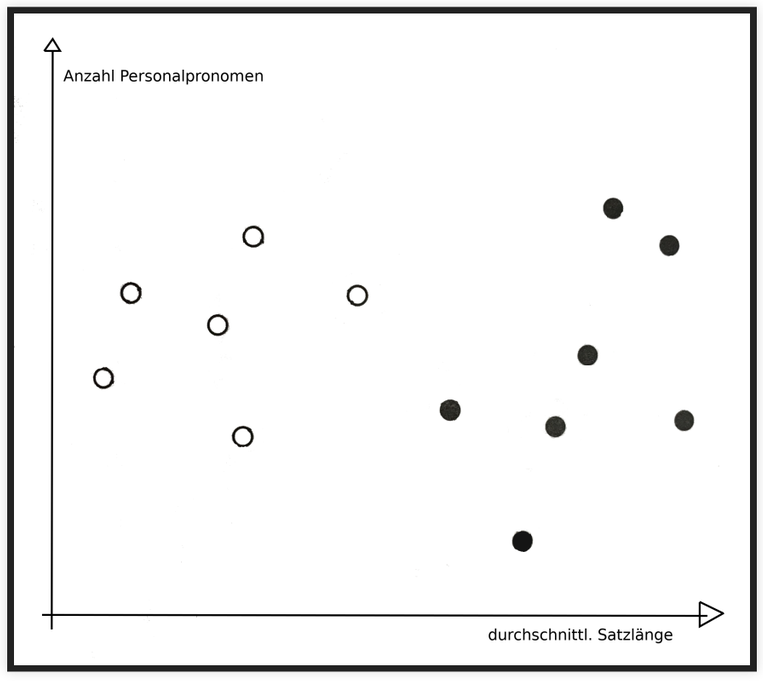
Textklassifikation mit Supportvektormaschine (SVM)

- als Algorithmus zur Mustererkennung von Vladimir Vapnik und Aleksej Červonenkis seit den 1970er Jahren entwickelt
- als Verfahren zur Textklassifikation seit den 1990ern intensiv erforscht und erfolgreich angewendet
- die Mathematik hinter SVMs ist kompliziert (strukturelle Risikominimierung, Lagrange-Multiplikatoren, Hilberträume)
- es gibt aber eine anschauliche, leicht verständliche geometrische Interpretation von SVMs

Anordnung der Trainingsinstanzen in einer (Hyper-)Ebene

Der Anschaulichkeit halber wird hier ein vereinfachtes Beispiel im zweidimensionalen Raum gegeben.

In der Praxis hat der verwendete Raum so viele Dimensionen, wie die E-Books Attribute haben (= Spalten im Attributvektor, siehe oben), also für unsere E-Books ca. 25.000

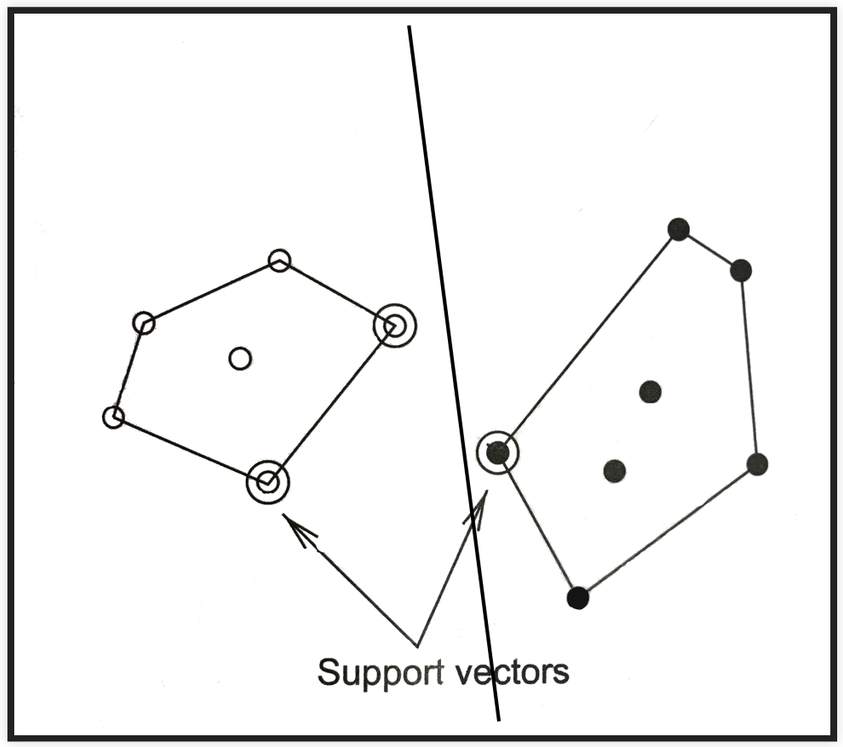


weiße Punkte: Warengruppe A, schwarze Punkte: Warengruppe B

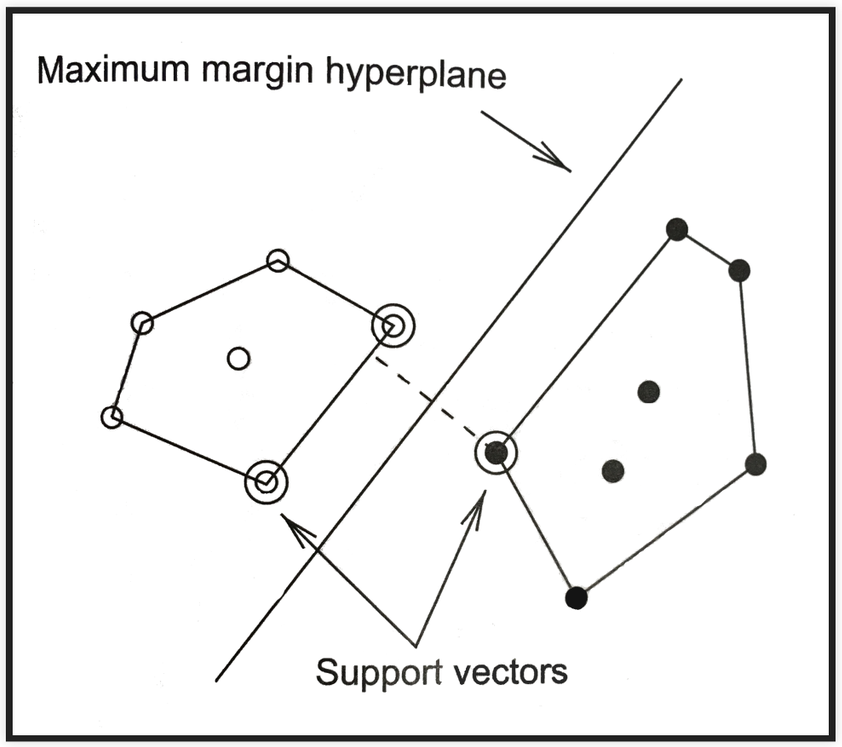
Trainingsphase: Trennebene berechnen

pagina Publikations technologien

Zum Trainieren einer SVM wird nun die „Maximum Margin Hyperplane“ berechnet. Das ist die Hyperebene, welche die Elemente einer Warengruppe von den Elementen der anderen Warengruppe trennt und dabei den maximalen Abstand zu den Supportvektoren (= „Elemente am Rand“) hat.



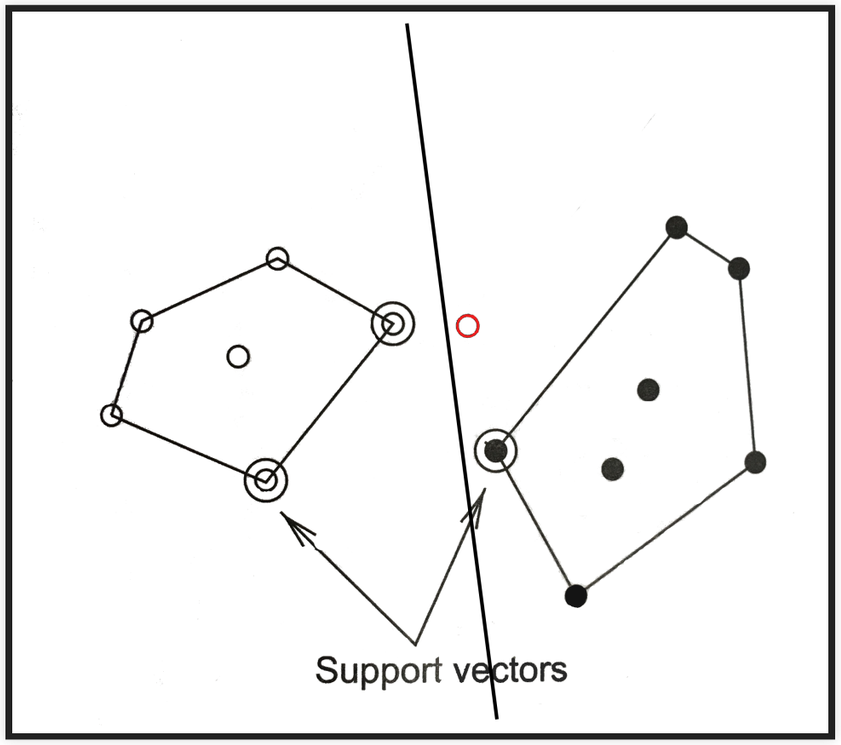
Beliebige Trennlinie zwischen den Warengruppen



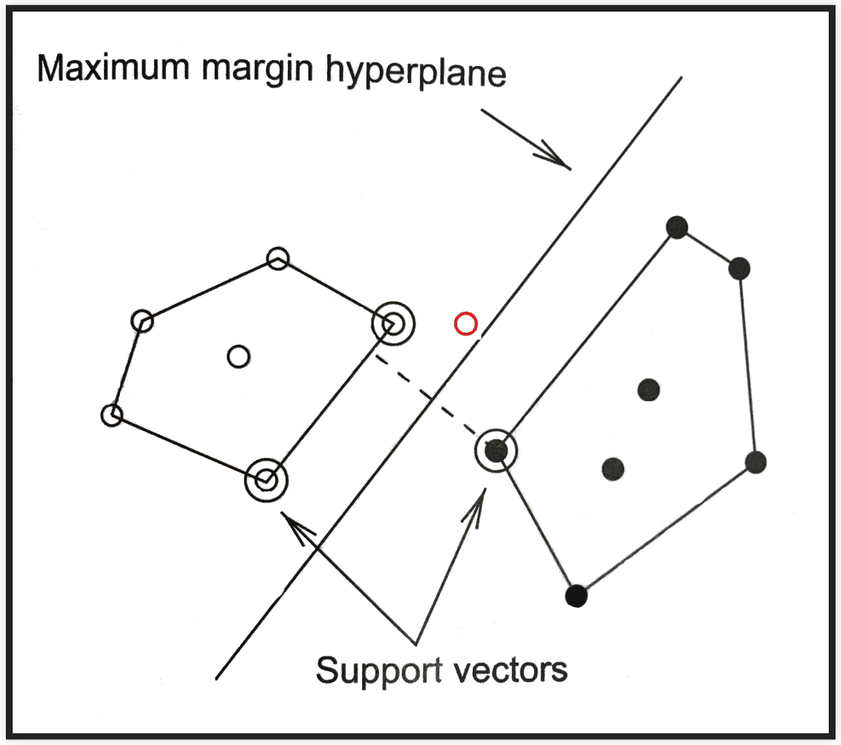
SVM mit Maximum Margin Hyperplane; Abb. aus Witten et al., Data Mining, 4. Aufl.

Anwendungsphase

Ein zu klassifizierendes Element wird im Hyperraum lokalisiert und berechnet, auf welcher Seite der Trennebene es sich befindet. Daraus ergibt sich dann die Warengruppe.



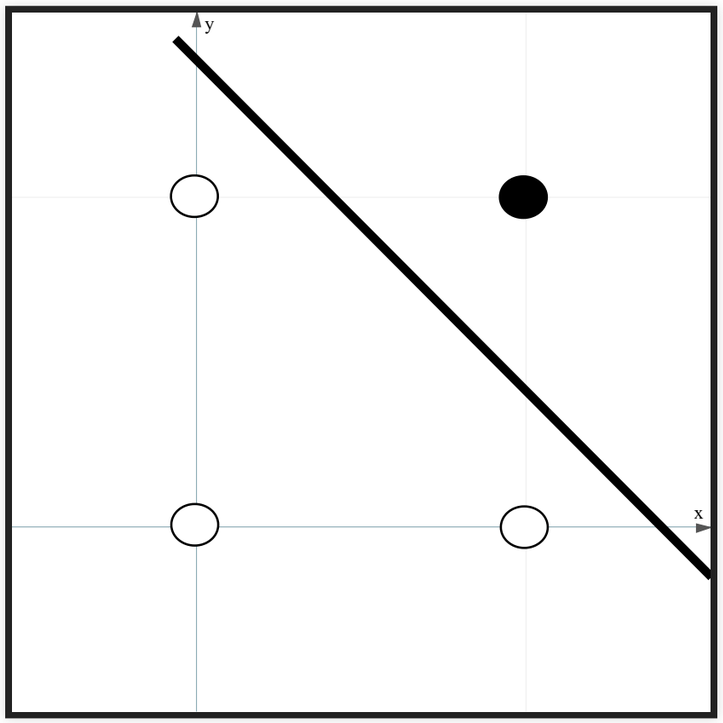
Beliebige Trennlinie: neue Instanz wird falsch klassifiziert



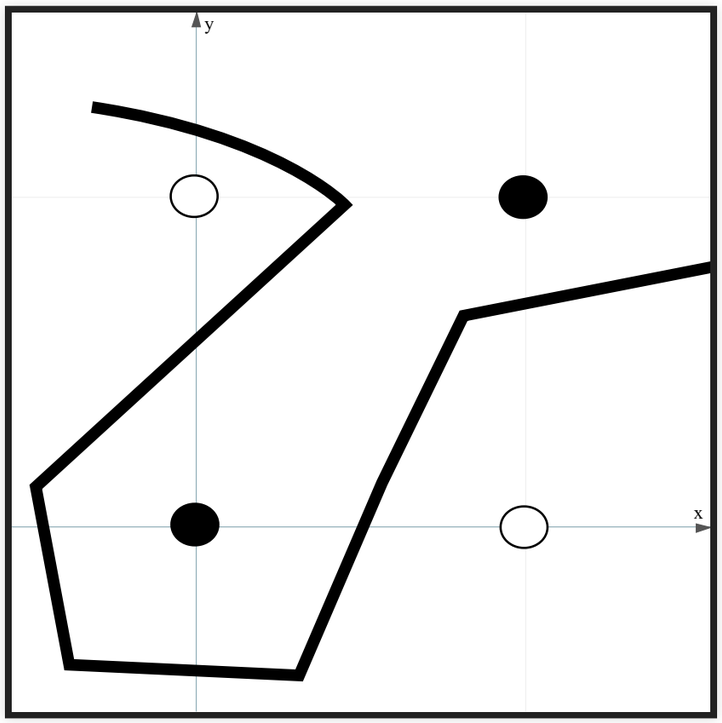
SVM mit Maximum Margin Hyperplane: neue Instanz wird korrekt klassifiziert

Nicht linear trennbare Probleme

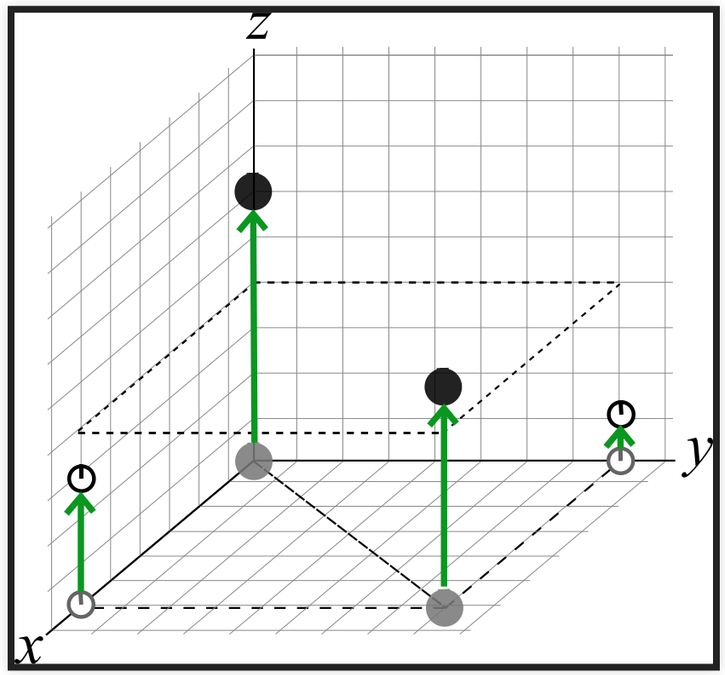
Nicht alle Klassifikationsaufgaben sind linear trennbar. Diese Problem lässt sich im SVM-Kontext lösen, indem die Datenpunkte mit einer bestimmten Funktion („Kernelfunktion“) in eine höhere Dimension projeziert werden.



Linear trennbares Problem



Nicht linear trennbares Problem



Nach Transformation: linear trennbar

Evaluation des E-Book-Klassifikators für VLB-Warengruppen

Verfahren	Trainingsinstanzen	Attribute	Klassen	Erfolgsquote
SMO mit Hyperonymen	4615	25912	149	67,80%
SMO mit kontrolliertem Vokabular ohne Duplikate	4615	16798	149	66,93%
SMO	4615	16270	149	66,62%
Naïve Bayes mit AdaBoost.M1	4615	16270	149	60,91%
Naïve Bayes	4615	16270	149	59,42%

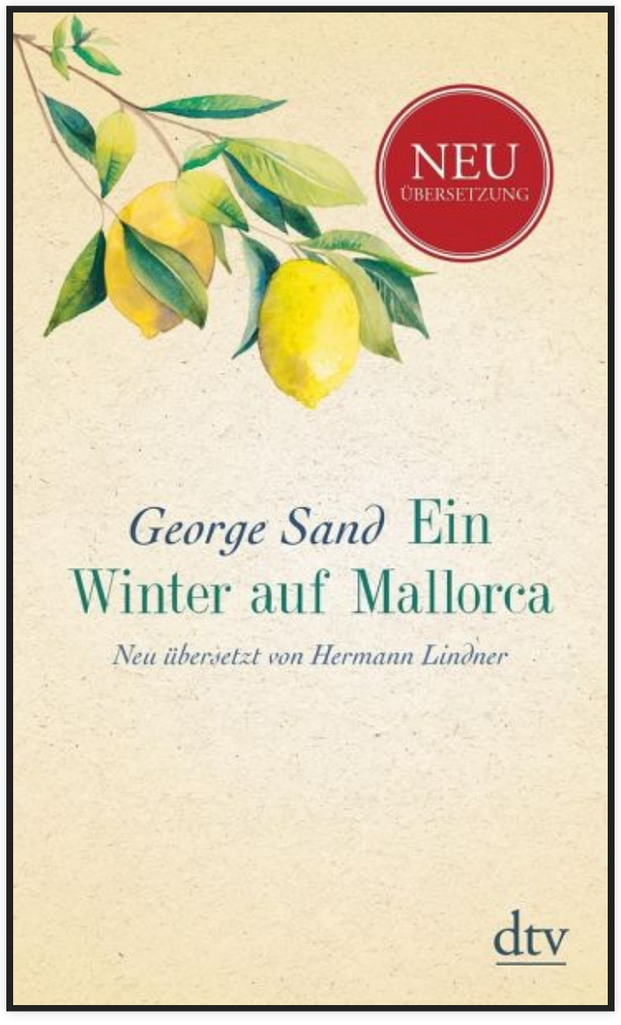
Nur 67,80% korrekt klassifiziert? Klingt nicht so gut…

Ja, stimmt, das geht noch besser, z. B. mit mehr Trainingsdaten.

Aber: Textklassifikation ist inhärent subjektiv!

Die Erfolgsquoten drücken nur aus, ob die trainierten Maschinen die E-Books genauso klassifizierten wie die Verlage. Nicht alle der übrigen 32,20% sind damit zwangsläufig falsch klassifiziert.

Wie würden Sie dieses Buch klassifzieren?



Der Verlag sagt: Warengruppe 111, d. h.: Erzählende Literatur von Autor/innen, deren Hauptwerk vor 1945 verfasst wurde.

Naïve Bayes sagt: Warengruppe 362, d. h.: Reiseerzählungen und Reiseromane.

Das passt beides, oder?

Vielen Dank für Ihre Aufmerksamkeit!