Quantitative Textanalyse 2: Maschinelles Lernen

Vorlesung *Einführung in die Digital Humanities*
MSc Digital Humanities | Wintersemester 2018/19

Prof. Dr. Christof Schöch

Universität Trier

Einstieg

Semesterüberblick

- 23.10.: Digital Humanities im Überblick
- 30.10.: Digitalisierung: Text und Bild
- 06.11.: Grundbegriffe des Programmierens
- 13.11.: Datenmodellierung 1: Modellierung
- 20.11.: Datenmodellierung 2: Datenbanken
- 27.11.: Datenmodellierung 3: Text, Markup, XML
- 04.12.: Digitale Edition
- 11.12.: Geschichte der Digital Humanities
- 18.12.: Informationsvisualisierung
- 22.12.-2.1.: *Weihnachtspause*
- 08.01.: Computerlinguistik / Natural Language Processing
- 15.01.: Quantitative Analyse 1: Stilometrie
- 17.01.: **Quantitative Analyse 2: Superv. Machine Learning**
- 29.01.: Open Humanities
- 05.02.: Klausurtermin

Sitzungsüberblick

1. Machine Learning (ML)
2. Unüberwachtes ML: Verfahren
3. Überwachtes ML: Anwendungsbeispiel
4. Überwachtes ML: verschiedene "Classifier"

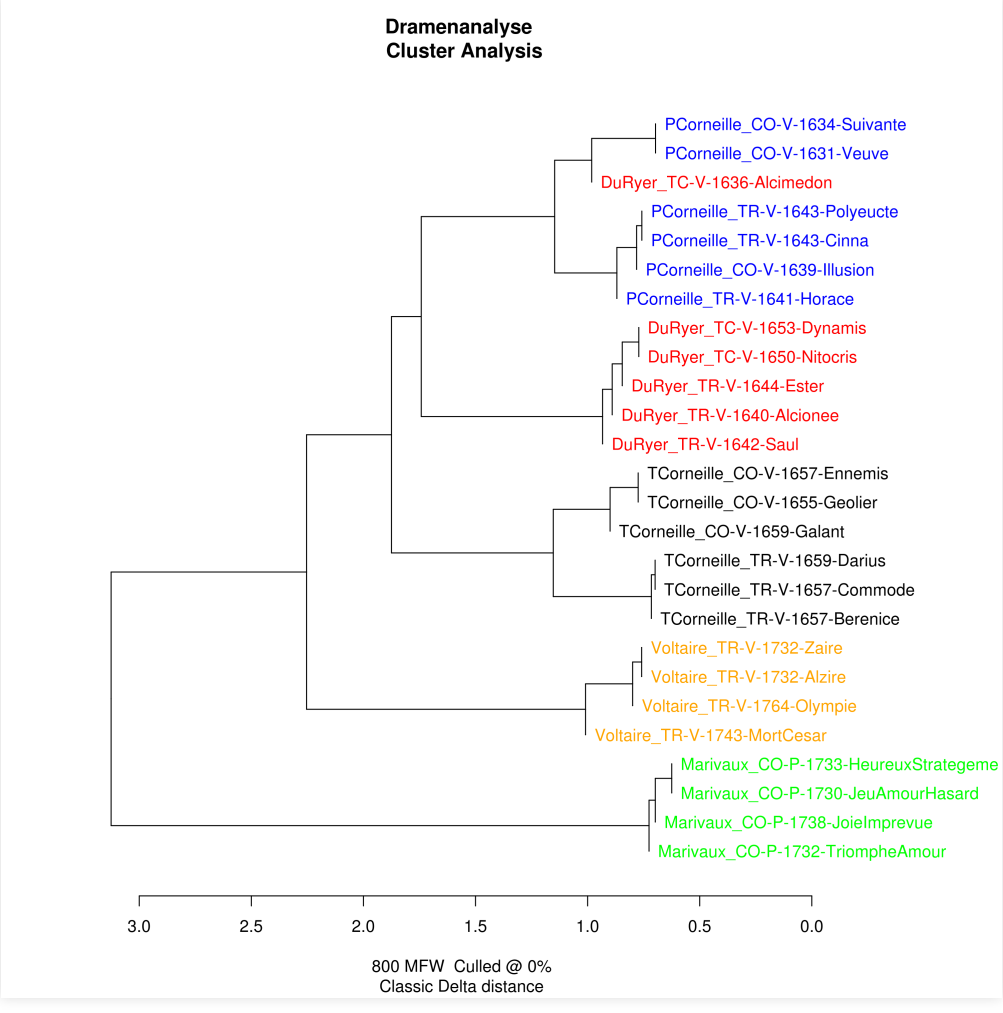
1. Machine Learning

Zwei Typen von ML

unüberwacht / unsupervised	**überwacht / supervised**
Clustering	Klassifikation
Bilden von Gruppen	Zuordnung zu Klassen
keine Klassen	vorher bekannte Klassen
ein Datensatz	Training/Test/Anwendung
eher explorativ	hypothesengeleitet
Evaluation möglich	Evaluation leicht
Topic Modeling PCA, CA	Annotation OCR, NER

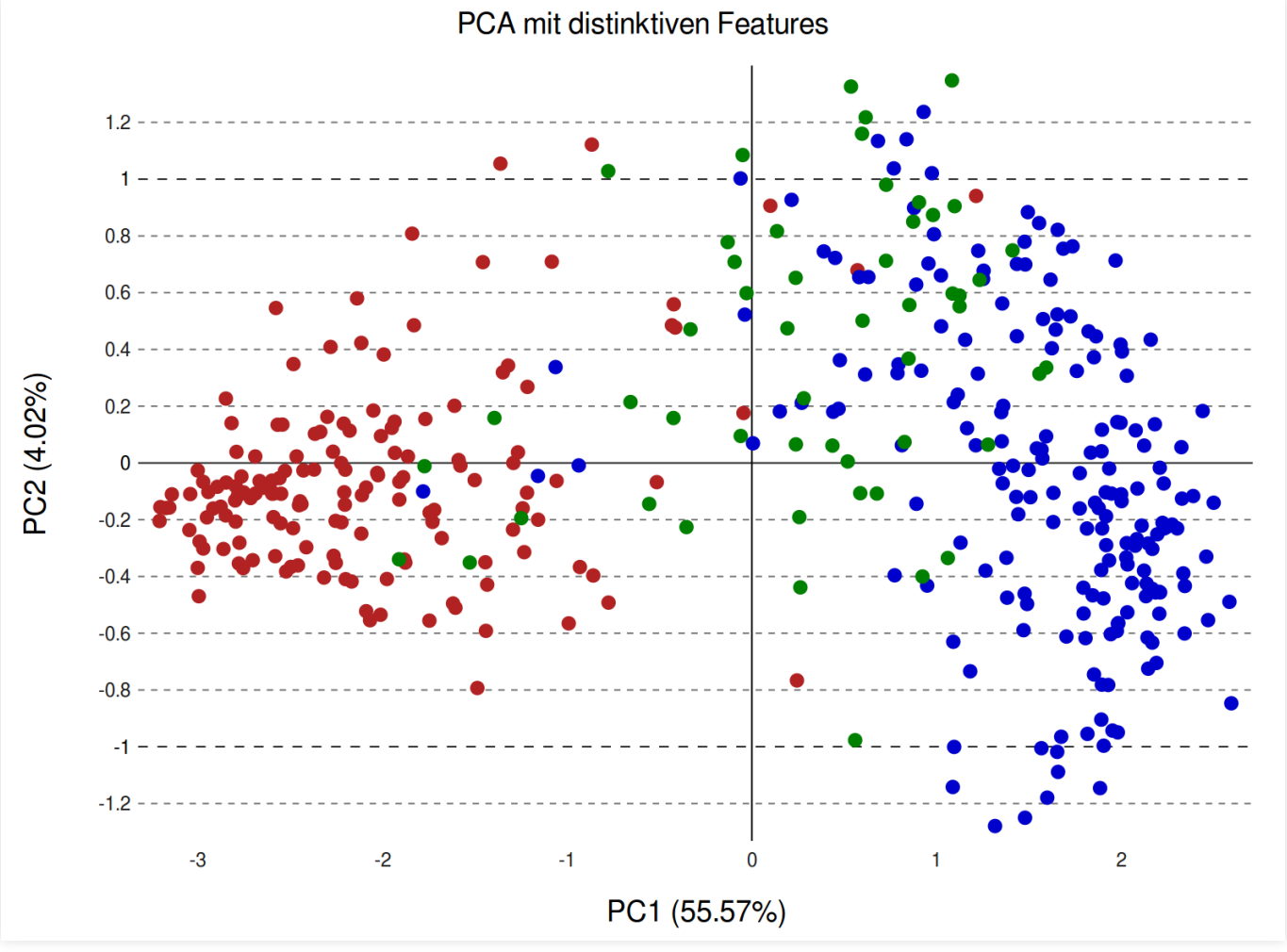
2. Unüberwachtes ML: Verfahren

Cluster-Analyse



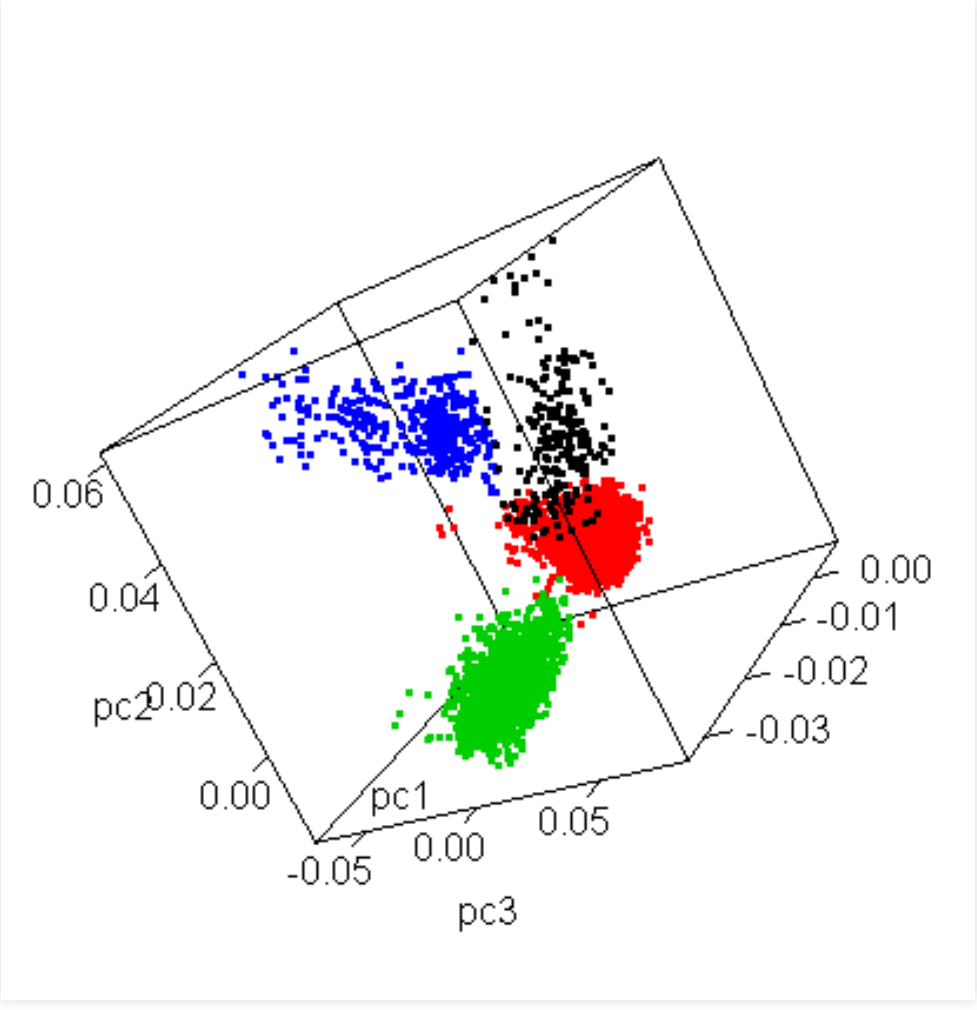
(Stilometrie)

Principal Components Analysis



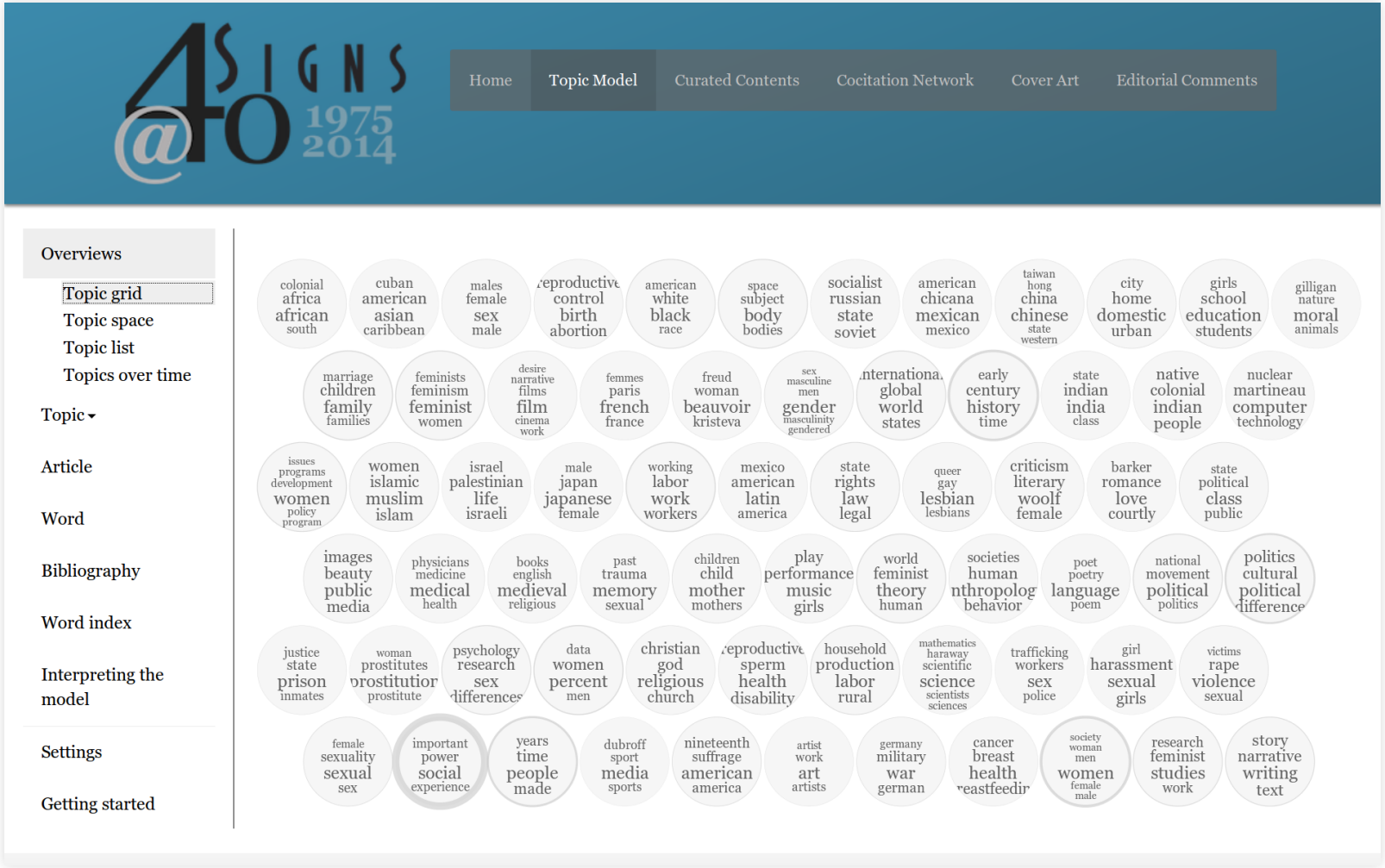
(Dimensionalitätsreduktion)

Principal Components Analysis



(Dimensionalitätsreduktion;
Quelle: http://www.gettinggeneticsdone.com/2011/10/new-dimension-to-principal-components_27.html

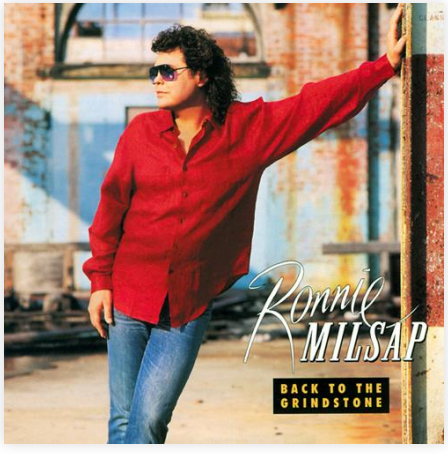
Topic Modeling



Signs at 40

3. Überwachtes ML (Klassifikation): Anwendungsbeispiel

Projektseminar: Albencover



Klassifikation: Rock, Pop, Hip-Hop, Country, Electronic.
Quelle: https://musicbrainz.org/

Prototypischer Ablauf

1. Vorbereitung (Gegenstand, Fragestellung)
2. Datensammlung erstellen
3. Annotieren nach Klassen (Teil)
4. Merkmale generieren
5. Trainingsphase
6. Evaluationsphase
7. Anwendungsphase (Datensätze ohne Klasse)
8. Interpretation der Ergebnisse

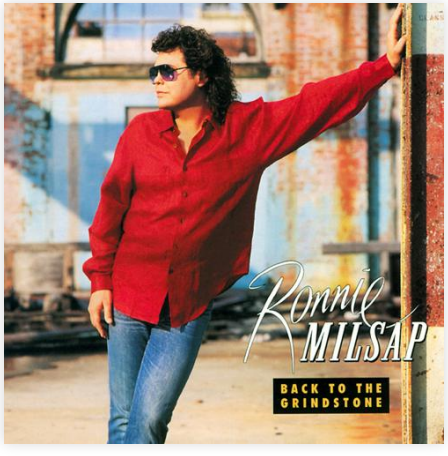
(1) Vorbereitung

- Annahme: Musiker sind Künstler, denen auch die künstlerische Gestaltung ihrer Albumcovers wichtig ist
- Hypothese: Es gibt einen Zusammenhang zwischen Musikrichtung und Cover Art
- Aufgabe: Albencover nach Musikrichtung klassifizieren
- Nur auf Grundlage der visuellen Information
- Bei fünf Genres: Zufallsbaseline 20%, Human Baseline: knapp 50%

(2) Datensammlung erstellen

- Datenquelle: musicbrainz.org, Abruf über API
- Struktur: Fünf Genres
  Rock, Pop, Electronic, Hip-Hop, Country
- Umfang: 5 x 3.000 = 15.000 Albumcover
- Daten: Bilddatei und Metadaten
  (Jahr, Titel, Band, Genre)

Beispiele für Cover



Quelle: https://musicbrainz.org/

(3) Annotieren nach Klassen

- Jedes Album wird einer Musikrichtung zugeordnet
- Wir übernehmen die Zuordnung von Musicbrainz

(4) Merkmale generieren

- Einfach
  - Dominante Farben (Histogramm des HSV-Farbraums)
  - Sättigung und Helligkeit (HSV-Farbraum)
- Komplex
  - Anzahl der Gesichter (OpenCV)
  - Welche Objekte sind sichtbar (ClarifAI API)
- Daten in einer Merkmals-Matrix zusammengefasst
- Optional: Merkmalsskalierung (z-scores)

(5) Merkmals-Matrix

	A	B	C	D	E	F	G
1	**hash**	**genre**	**people**	**faces**	**max_R**	**max_G**	**max_B**
2	1193941c-de68-3299-95b9-87b0922e27b0	hip-hop	0	2	33	11	15
3	28d79e63-2e5a-404d-b95f-f1b383babcfe	country	0	2	0	0	0
4	3950d762-f987-46ac-8044-86d90e40dd5d	country	0	0	73	138	217
5	3f6ba929-ff4a-4af8-a896-2803ca3112c0	rock	0	0	200	9	233
6	491ded48-ec70-42cd-bba3-7fc4b0115f56	rock	0	0	0	0	0
7	498f132b-c8a3-4d51-a906-367704f692ae	country	1	0	211	247	241
8	55267b7f-24b1-3180-b988-4990e07f88c8	electronic	0	0	10	10	10
9	67ebbc43-0415-4a07-90b9-3f8f8be296eb	hip-hop	0	2	0	0	0
10	83219409-a39c-3c3c-9928-06f0eac04423	electronic	0	1	37	16	16
11	912a8e16-3730-46e4-847d-dfba119df13a	pop	0	0	231	230	240
12	d7bd5278-6a18-4542-a1be-34ab59d050b7	country	0	0	34	107	255
13	d8e9e0f9-7ced-4176-a935-cfac17409d10	rock	0	0	17	17	0
14	f4afbe11-984e-4865-9677-b244dce5c8ed	country	2	4	0	0	0
15	052c18c4-2fd7-4be8-bf9e-c538d131a039	electronic	1	0	241	245	246
16	1174e529-a321-4a65-8c02-6dcbd573a383	country	1	0	20	24	186

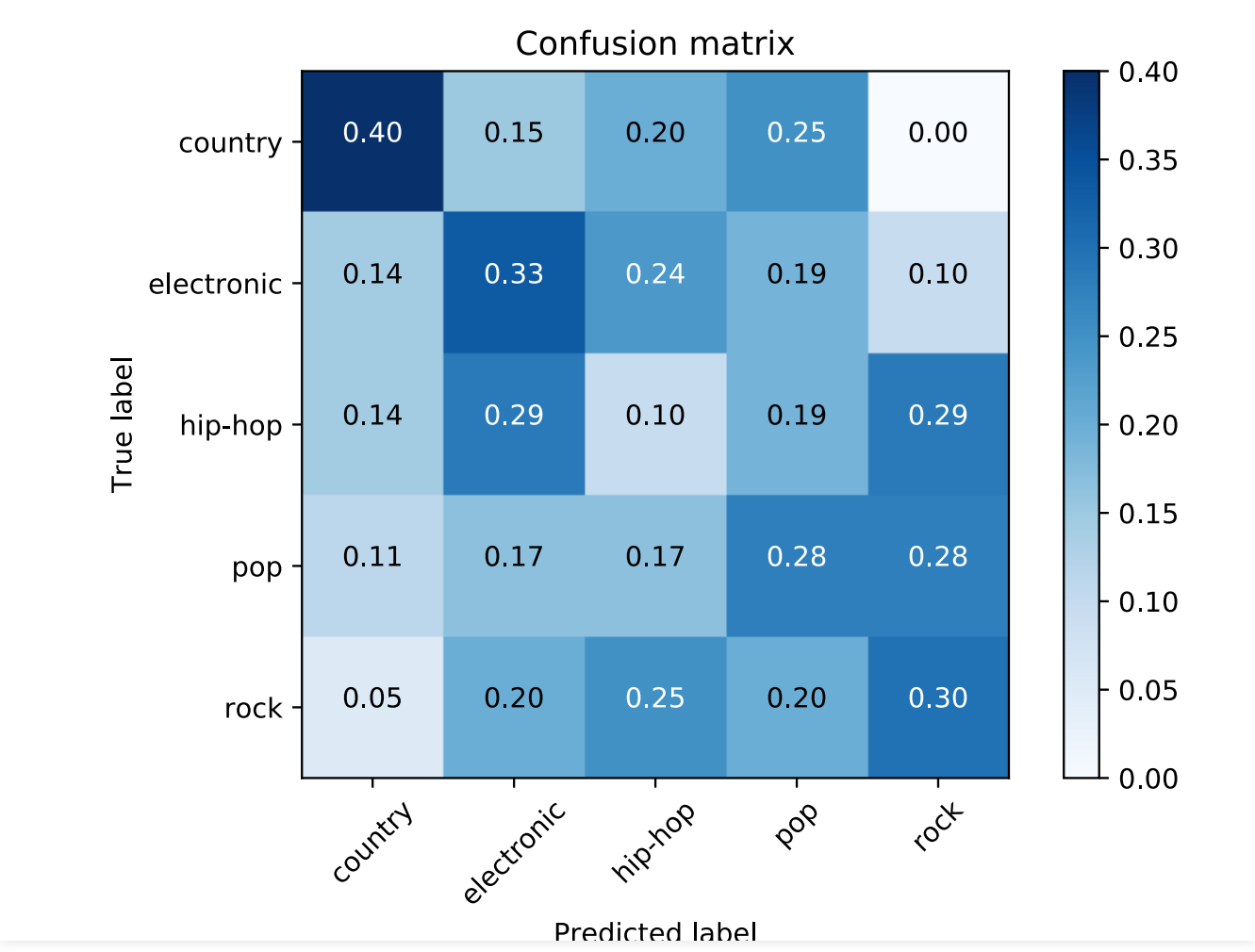
(5) Trainingsphase

- Ein Teil der gelabelten Daten (bspw. 90%) zum "Trainieren"
- Algorithmus "lernt" einen Zusammenhang zwischen Merkmalen und Klassen
- Verschiedene "Classifier" mit ihren Parametern
- Bspw. "k-nearest neighbor"

(6) Evaluationsphase

- Rest der Daten (10%) zur Evaluation
- Vergleich der tatsächlichen Klasse mit der vom Algorithmus ernmittelten Klasse
- F-Score
  - Precision: welcher Anteil der als "Pop" erkannten Alben sind tatsächlich "Pop"
  - Recall: welcher Anteil der Pop-Alben wurden als solche erkannt?
  - F-Score: 2 x (precision x recall) / (precision + recall)
- Confusion Matrix

(6) Confusion Matrix



7. Anwendungsphase

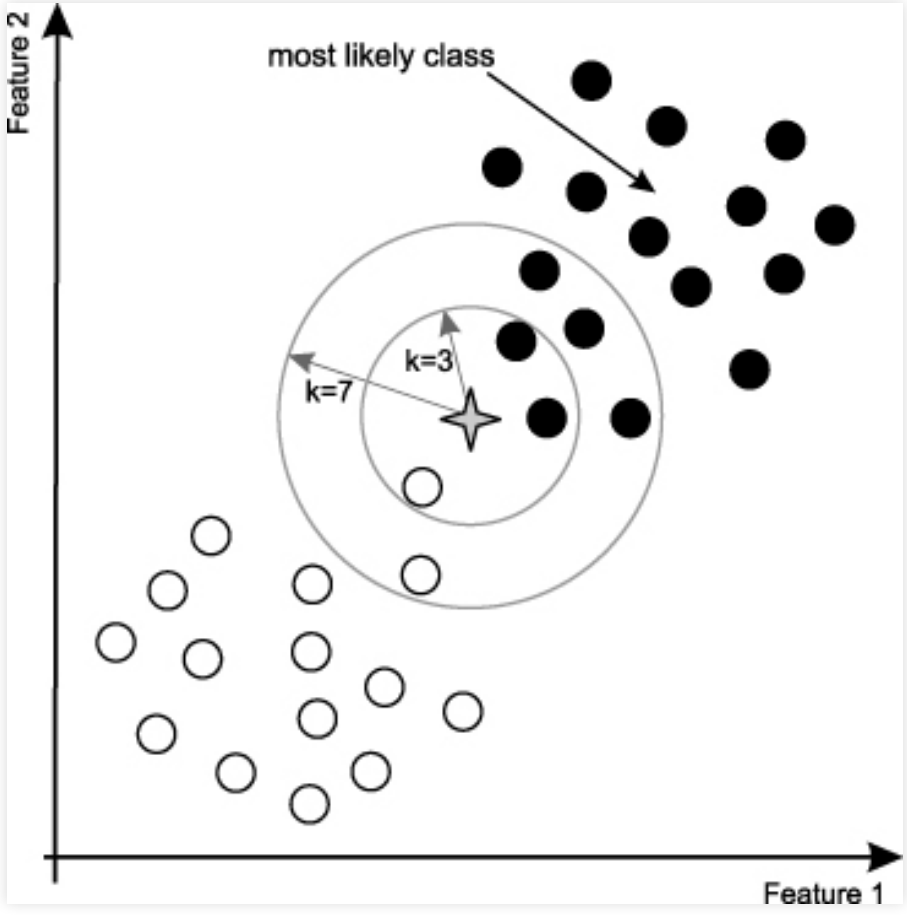
- Entfällt in diesem Beispiel, weil alle Daten gelabelt sind
- Man könnte jetzt aber für weitere Alben Genrelabels vergeben

8. Interpretation der Ergebnisse

- Wie stark ist der angenommene Zusammenhang Cover / Genre?
- Gibt es Unterschiede zwischen den Genres?
- Sind die Klassen wirklich disjunkt?
- Welche Merkmale sind entscheidend?

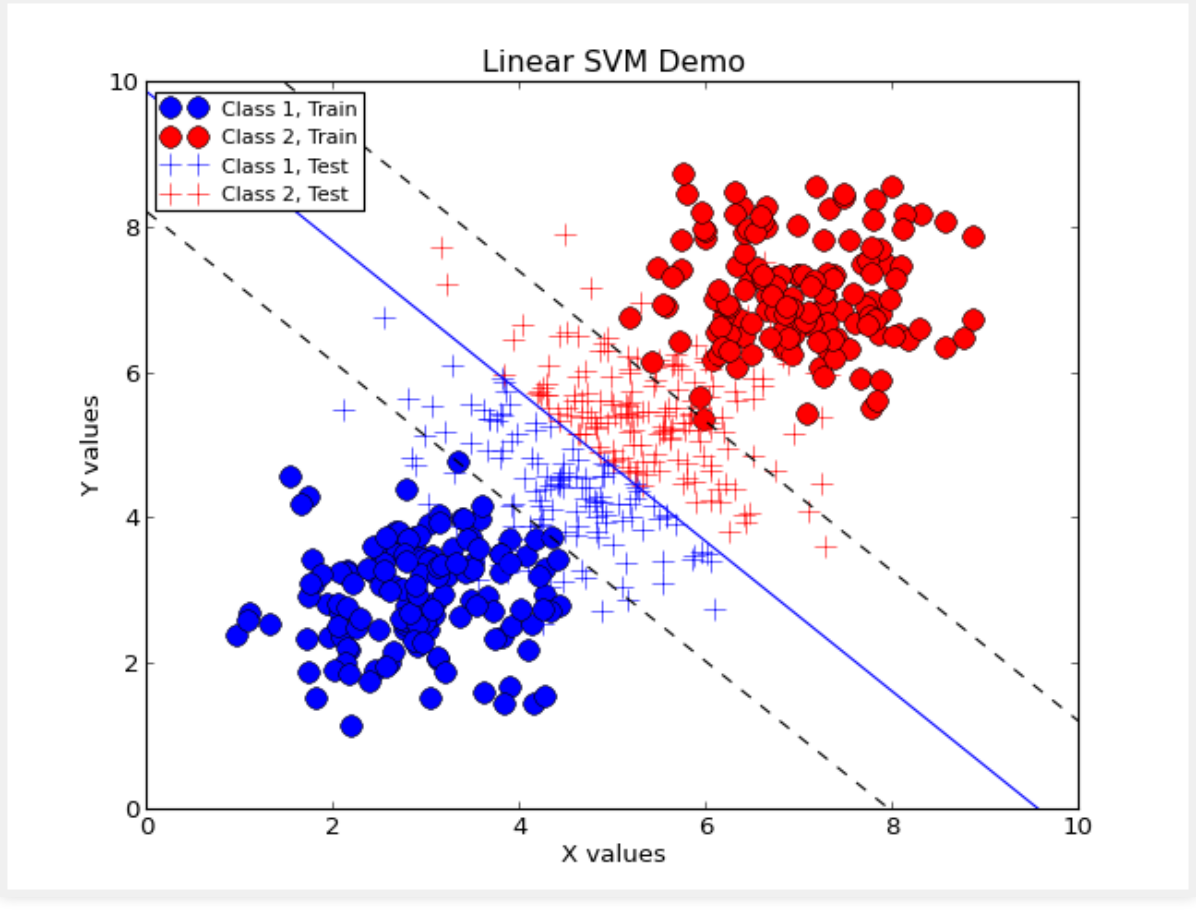
3. Verschiedene ”Classifier”

Classifier: k-nn



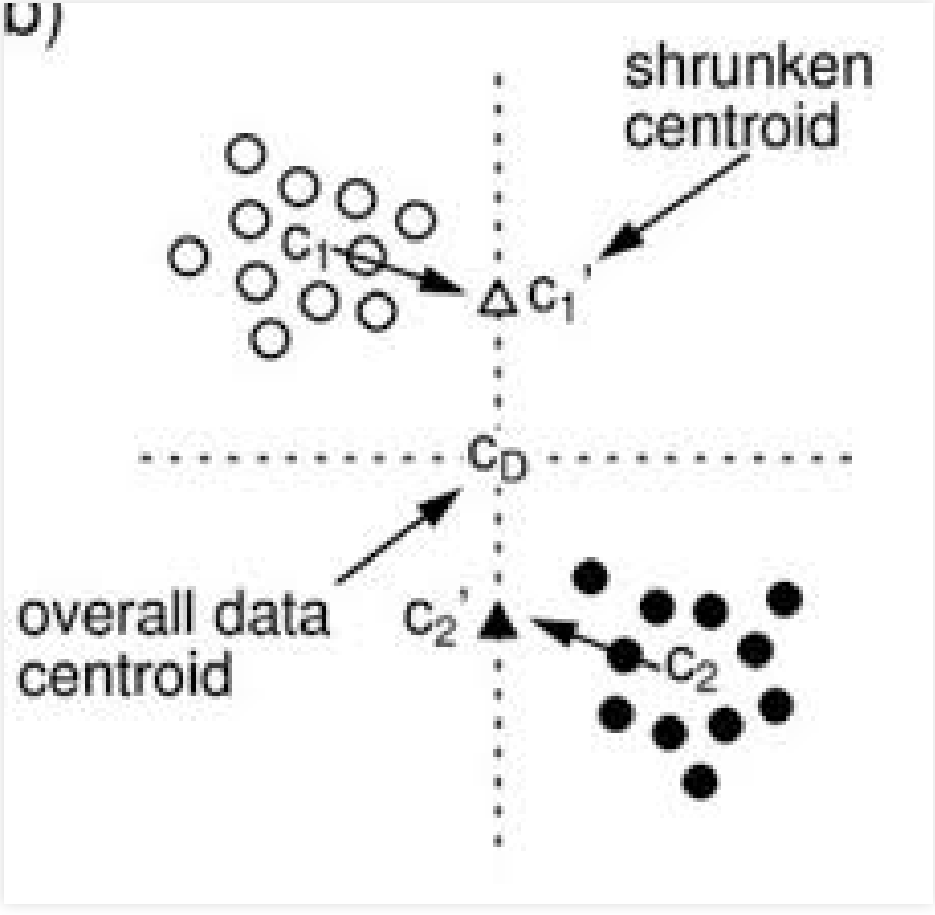
Quelle: Struyf, Jan; Dobrin, Seth; Page, David: "Combining gene expression, demographic and clinical data in modeling disease: A case study of bipolar disorder and schizophrenia", https://www.researchgate.net/figure/Illustration-of-the-a-support-vector-machines-b-nearest-shrunken-centroids-c_fig1_23459323, Lizenz CC-BY

Classifier: SVM



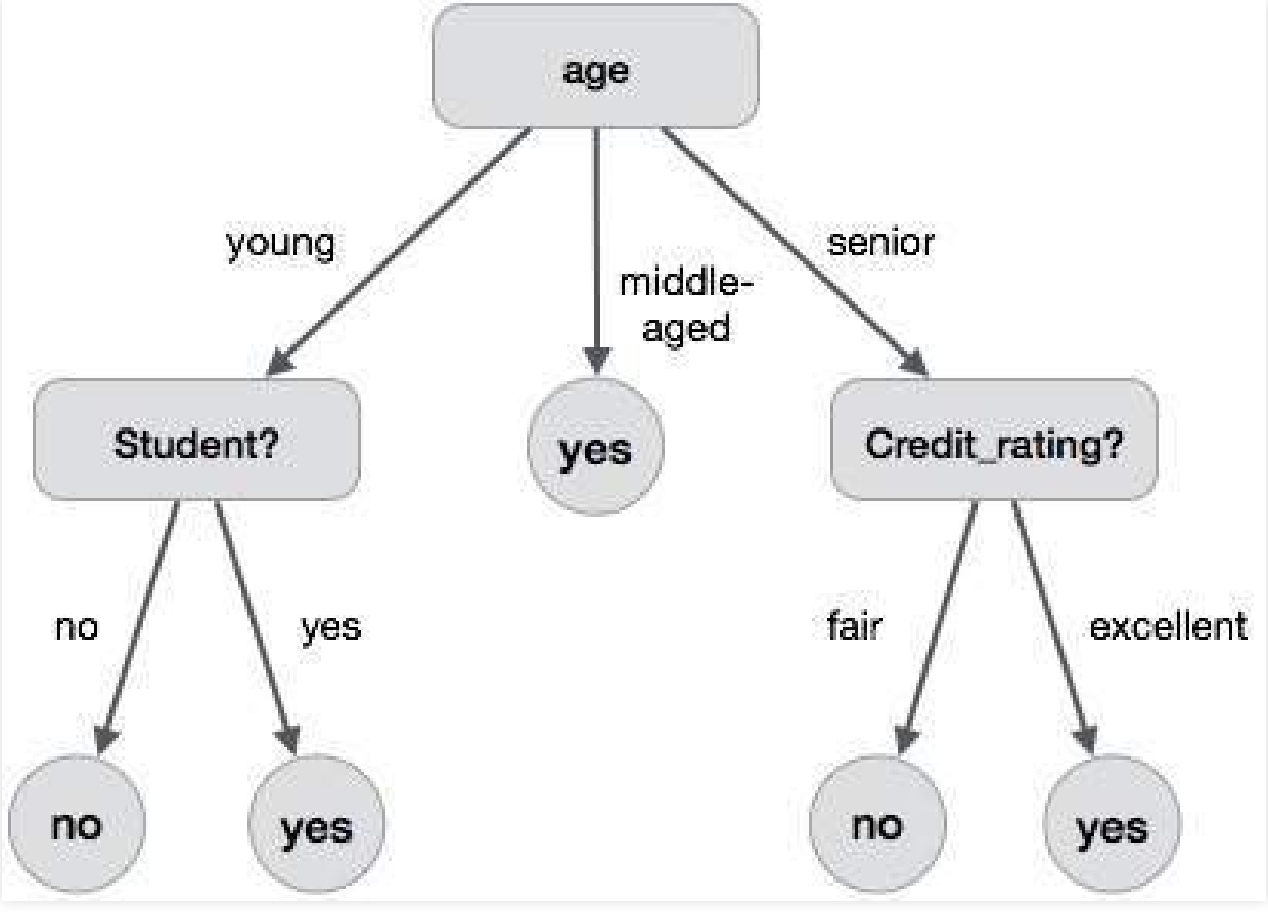
Bildquelle: "A Linear Support Vector Machine", 2014: https://randomforests.wordpress.com/2014/01/29/a-linear-support-vector-machine/

Classifier: NSC



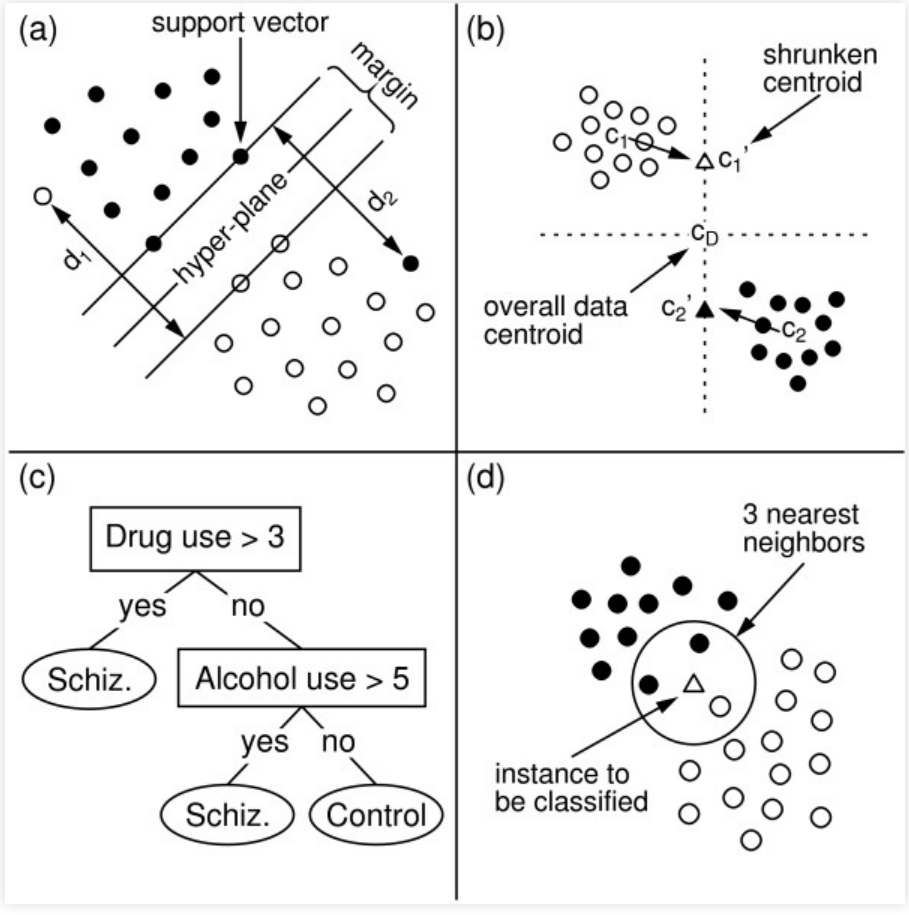
Quelle: Struyf, Jan; Dobrin, Seth; Page, David: "Combining gene expression, demographic and clinical data in modeling disease: A case study of bipolar disorder and schizophrenia", https://www.researchgate.net/figure/Illustration-of-the-a-support-vector-machines-b-nearest-shrunken-centroids-c_fig1_23459323, Lizenz CC-BY

Classifier: Decision Tree



Quelle: Struyf, Jan; Dobrin, Seth; Page, David: "Combining gene expression, demographic and clinical data in modeling disease: A case study of bipolar disorder and schizophrenia", https://www.researchgate.net/figure/Illustration-of-the-a-support-vector-machines-b-nearest-shrunken-centroids-c_fig1_23459323, Lizenz CC-BY

Vier Classifier



Quelle: Struyf, Jan; Dobrin, Seth; Page, David: "Combining gene expression, demographic and clinical data in modeling disease: A case study of bipolar disorder and schizophrenia", https://www.researchgate.net/figure/Illustration-of-the-a-support-vector-machines-b-nearest-shrunken-centroids-c_fig1_23459323, Lizenz CC-BY

Abschluss

Fragen?

Lektürehinweise

- Christof Schöch, "Quantitative Analyse", in: *Digital Humanities: Eine Einführung*. Hrsg. von Fotis Jannidis, Hubertus Kohle, Malte Rehbein. Stuttgart: Metzler.

Weitere Empfehlungen

- Blei, D. M. (2012). "Probabilistic topic models". In: *Communications of the ACM*, 55(4): 77–84. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf

Darüber hinaus

- Alpaydin, E. (2010). *Introduction to Machine Learning*. 2nd ed. Cambridge, Mass: MIT Press.
- Ramsay, Stephen (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana Ill.: University of Illinois Press.

Nächste Sitzung

- 1.2.2018 (letzte Sitzung vor der Klausur)
- Mein Themenvorschlag: "Open Humanities"
- Vorbereitung: "Informationen zu Open Access" (mit den Unterpunkten im Menü): http://open-access.net/informationen-zu-open-access/

Christof Schöch, 2018
http://www.christof-schoech.de
