COMPEON
Das Finanzportal für den Mittelstand

Seminar Statistische Lernverfahren

Klassifikation von Rezensionstypen

Till Gräfenberg, Alexander Kohlscheen, Michael Lau, Tanja Niklas,
Matthias Häußler, Jonathan Schmitz

12. Dezember 2019

Inhaltsverzeichnis

1. Problemstellung
2. Erstellen von Prädikatoren
3. Analysemethoden
   - 3.1 Naive Bayes
   - 3.2 Entscheidungsbaum
   - 3.3 Random Forest
   - 3.4 weitere Anpassungen und Modelle

Problemstellung

- Ziel: Klassifizierung von Reviews in folgende Typen

Texttyp	introvertiert	extrovertiert
emotional	stetig	initiativ
rational	gewissenhaft	dominant

- Gegeben: 439 bereits klassifizierte Reviews

Schwierigkeiten

- Keine eindeutige Klassifikation
  - Auch für Menschen nicht eindeutig
  - Teilweise sehr geringe Unterschiede zwischen den Typen
- Geringe Zahl an Trainingsdaten
- Unbalanciertes Studiendesign
- Representativität
  - Introvertierte Kunden schreiben weniger häufig Reviews
  - Nur positive Bewertungen lagen vor

Erstellen von Prädiktoren

- Klassifikation sollte durch verwendete Wörter geschehen
- Zurückführung auf Grundwörter notwendig
- Benutzung verschiedener Packages in `R` bzw. `Python` ermöglichte verschiedene Verfahren.

Erstellen von Prädiktoren

Stemming

- Durch Abschneiden von Prä-/In- und Suffixen und Ersetzen von Umlauten, Diphtongen etc. erzeugen von Wortstämmen.
- Eigene Implementierung nach Vorgabe von COMPEON in `R`
- Für Englische Sprache bereits vorgefertigte Tools z.B.
  - `porterstemmer` von `nltk` in Python
  - `snowballstemmer` von `nltk` in Python

Probleme:

- Unregelmäßigkeit von Verben im Deutschen
- Komposita

Erstellen von Prädiktoren

Lemmatisierung

- Alternative: Zurückführung auf grammatikalische Grundformen
- Erfordert vorgefertigte Packages z.B.
  - `SpaCy` in Python
  - `nltk` in Python
- Diese lieferten zusätzlich Informationen über die Wortart
- Auch hier für Englische Sprache ausgereifter als die deutsche Alternative

Erstellen von Prädiktoren

Filterung der Prädikatoren, weitere

- Nach Erstellung der Grundwörter konnte gefiltert werden, welche Wörter häufig auftraten
- Denkbare Filtermethoden:
  - Nur Wörter, die mind. n Mal aufgetaucht sind
  - Nur Wörter, die in mind. $p\%$ der Reviews verwendet wurden
- Anschließend Erstellung einer binären Document-Term-Matrix, die kodiert, welche Grundwörter in welchen Reviews auftauchten
- Alternative: PCA um aussagekräftige „Wörterachsen“ zu bestimmen. Kein sichtbarer Erfolg.

Erstellen von Prädiktoren

PCA - Principal Component Analysis

Ziel: Dimensionsreduktion

Idee: Suche die Datenachsen, auf denen die Varianz am größten ist

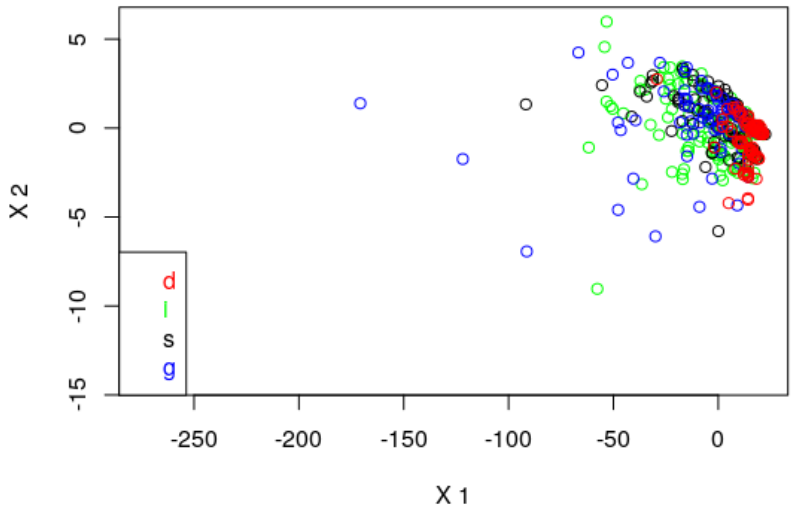
Verfahren:

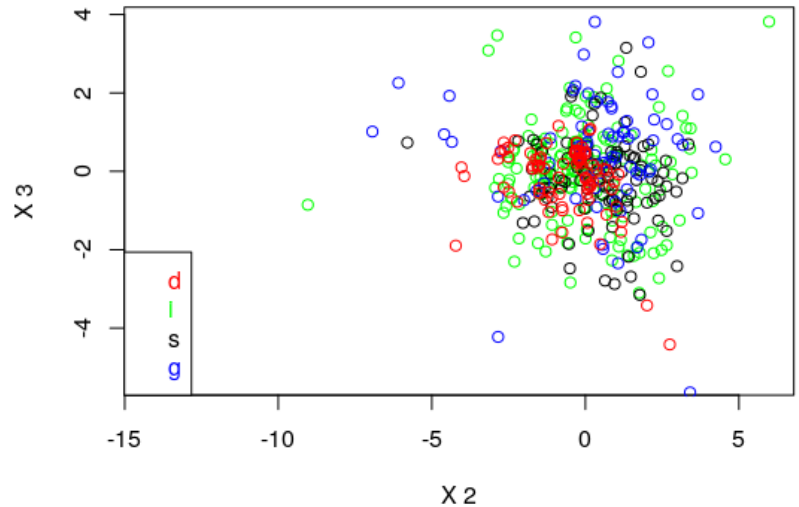
- Sei X die DT-Matrix (Spaltenmittelwerte $= 0$)
- Bestimme die Kovarianzmatrix $Cov = X^T X$
- Bestimme die Eigenwerte λ_i und Eigenvektoren v_i von Cov
- Sei $V = (v_1|v_2|...)$
- Transformiere die Daten zu $\hat{X} = XV$

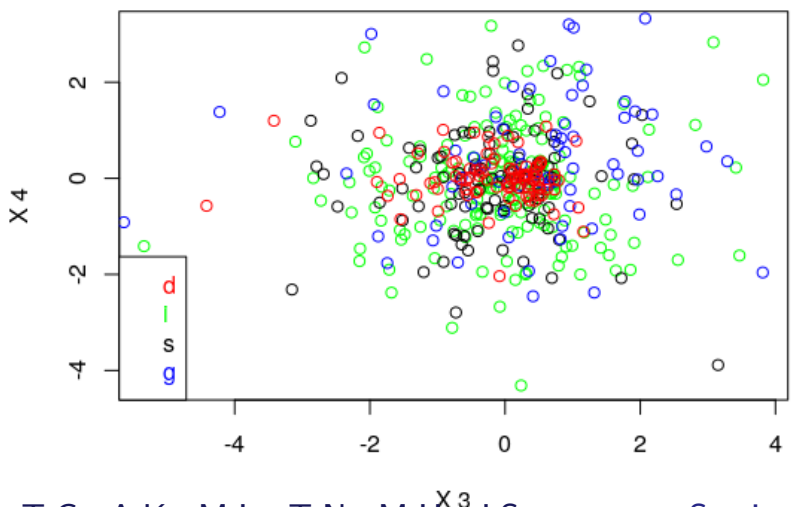
Problem: Die Resultate verlieren an Interpretierbarkeit

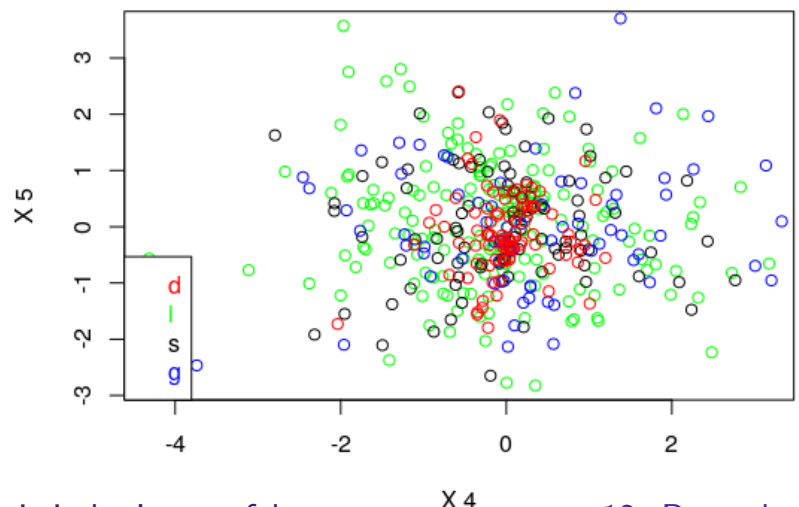
Erstellen von Prädiktoren

PCA - Principal Component Analysis









Erstellen von Prädiktoren

PCA - Principal Component Analysis

Fazit:

- Die Dominanten Reviews haben eine geringere Varianz
- keine erkennbaren Gruppen
- Mittelwerte der Gruppen sind ähnlich

Das Verfahren liefert keine besseren Ergebnisse.

Analysemethoden

Naive Bayes

- bl

Resultate Naive Bayes, R, mind. 20 mal Wörter

	Dominant	Gewissenhaft	Initiativ	Stetig
Dominant	14	2	8	1
Gewissenhaft	0	0	0	0
Initiativ	4	11	28	17
Stetig	0	1	0	0

Resultate Naive Bayes, Python, mind. in 1% der Texte, englisch

	Dominant	Gewissenhaft	Initiativ	Stetig
Dominant	13	0	4	1
Gewissenhaft	4	3	5	2
Initiativ	12	3	16	5
Stetig	4	0	11	3

Resultate Naive Bayes, Python, mind. in 1% der Texte, deutsch

	Dominant	Gewissenhaft	Initiativ	Stetig
Dominant	16	0	2	0
Gewissenhaft	2	5	5	2
Initiativ	13	4	16	3
Stetig	2	1	13	2

Analysemathoden

weitere Anpassungen und Modelle

Mit Naive Bayes und den Wortarten als Prädiktoren lässt sich zuverlässig voraussagen, ob eine Person extrovertiert ist:

	extrovertiert	introvertiert
extrovertiert	30	5
introvertiert	24	27

Idee:
Nutze die Vorhersage dieses Modells um ein neues Modell anzupassen.

Analysemathoden

weitere Anpassungen und Modelle

Random Forest

	D	G	I	S
D	14	2	8	1
G	1	4	1	0
I	3	7	25	15
S	0	1	2	2

Random Forest mit Naive Bayes

	D	G	I	S
D	15	2	9	1
G	0	4	1	1
I	3	8	24	13
S	0	0	2	3

Das modifizierte Verfahren liefert im Schnitt keine besseren Ergebnisse