ECL: Textklassifizierung

Katja Markert

Institut für Computerlinguistik
Universität Heidelberg
markert@cl.uni-heidelberg.de

November 27, 2019

Heute

1. Bis jetzt: Zählen und Verarbeitung von *n*-grams
2. Jetzt: Bag of Words in Textklassifikation mit überwachtem maschinellen Lernen
3. Problemdefinition
4. Naive Bayes für Topic Classification
    1. Multinomiales Modell
    2. Binomiales Modell
5. Evaluation, Textklassifikation in der Praxis

Heute

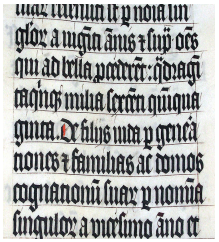
1. Bis jetzt: Zählen und Verarbeitung von *n*-grams
2. Jetzt: Bag of Words in Textklassifikation mit überwachtem maschinellen Lernen
3. Problemdefinition
4. Naive Bayes für Topic Classification
    1. Multinomiales Modell
    2. Binomiales Modell
5. Evaluation, Textklassifikation in der Praxis

Was ist eine Klassifikationsaufgabe?

Klassifikation

Gegeben eine Instanz oder eine Menge von Instanzen, weise der Instanz ein Konzept/Klasse/Kategorie aus einer fixen, diskreten Menge von Konzepten zu. Tue dies auf der Basis von Merkmalen (features/explanatory variables).

- Foul
- Elfmeter
- Ecke ✓



- Latein ✓
- Deutsch
- Englisch

Merkmale:

	e	ch	ing	L/D/E
t1	50	10	0	?
t2	10	0	15	?
t3	1	1	2	?

Terminologie

- **Instanz:** einzelnes Beispiel im Datensatz
- **Attribute/Feature/explanatory variable/Merkmal:** ein Aspekt einer Instanz. Beispiel: *ing*,*bank*, Länge, . . .
- **Wert (Value):** Wert eines Merkmals. Beispiel: ja, nein für das Merkmal *ing*, oder 0,1,2,3,4 . . . 100 . . . für das Merkmal *ing*
- **Konzept/Klasse/Response Variable/Antwortvariable:** was man lernen muss. Beispiel: Latein, Deutsch, Englisch . . .

Überwachtes Lernen

Grundidee

Ein Algorithmus, der einem Input einen Output zuweist, kann schwer zu entwerfen sein. Bei manchen Aufgaben ist es aber einfach Beispielinputs mit bekannter Klasse zu generieren oder zu erhalten. Man will dann den Algorithmus lernen.

Supervised Learning/Überwachtes Lernen

- Generiere Trainingsset, wo man die Klassen kennt.
- Wähle Merkmale und einen Modeltyp
- Lerne ein Vorhersagemodell (classifier) aus dem Trainingsset = schätze Modell aus Daten.
- Verifiziere das Model mit Testdaten (auf dem die Klassen mithilfe des gelernten Modells “geraten” werden sollen)

Textklassifikation Definition

Gegeben:

- (Repräsentation) eines Dokumentes d
- Fixe Menge an Klassen $C = \{c_1, c_2, \ldots, c_j\}$

Bestimme: Kategorie von $d : \gamma(d) \in C$, wobei γ eine Klassifikationsfunktion ist, die Dokumente auf Klassen abbildet

Häufigste Textklassifikation: Topic Classification

MedLine Artikel

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Mesh Subject Categories

- Blood Supply
- Chemistry
- Drug Therapy
- Epidemiology
- Embryology
- ...

Manuelle Klassifikation: Library of Congress, PubMed, Yahoo directory (damals) . . . . Vorteile und Nachteile?

Überwachtes ML für Textklassifikation

Gegeben:

- Ein (Test)dokument d
- Eine fixe Menge an Klassen $C = \{c_1, c_2, \ldots, c_j\}$
- Ein Trainingsset D von Dokumenten je mit einem Label in C $(d_1, c_1), \ldots (d_m, c_m)$

Betimme:

- Eine Lernmethode, die einen Klassifizierer γ lernt
- Für Testdokument d, weisen wir Klasse $\gamma(d) \in C$ zu

“Bag of words” für Textklassifikation: Intuition

About hotels, restaurants or movies?

A good budget hotel’ .... Price includes breakfast with really nice food. Rooms are modern and of a reasonable size. The centre of Leeds is about a 15 min walk at the most. Hotel has bar area.

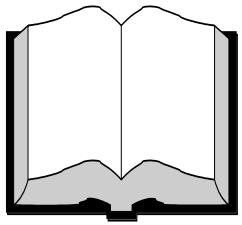
budget	1
hotel	2
price	1
rooms	1
breakfast	1
food	1
...	...

BOW in Multinomial Naive Bayes

Test Doc

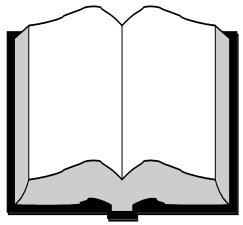
budget 1
hotel 2
price 1
breakfast 1
food 1

Hotels



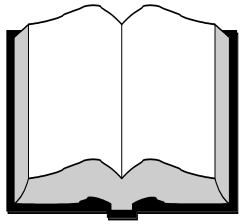
hotel 55
cleaner 10
reception 20
eggs 1
food 8

Restaurants



food 40
price 10
hotel 4
waiter 7

Movies



price 7
character 14
food 2
screen 7
plot 33

Naive Bayes: Multinomiales Model

Intuition

Benutze BoW Modell mit Worten als Merkmalne, Worthäufigkeiten als Merkmalswerten und Bayes Regel

Für Dokument d, bestimme wahrscheinlichste Klasse $c \in C$.

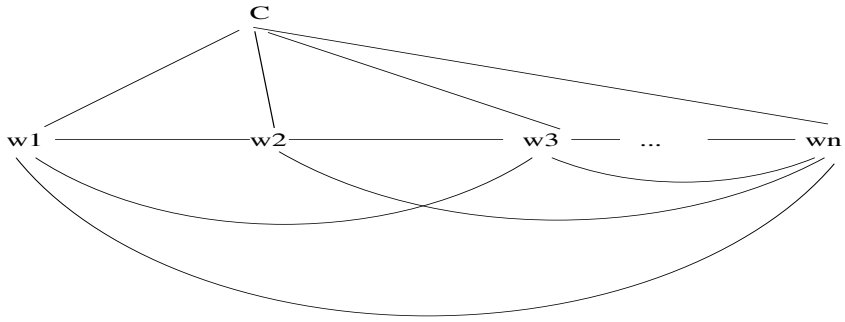
$$
\begin{aligned}
C_{MAP} &= \operatorname*{argmax}_{c \in C} P(c|d) && (1)\\
&= \operatorname*{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} && (2)\\
&= \operatorname*{argmax}_{c \in C} P(d|c)P(c) && (3)\\
&= \operatorname*{argmax}_{c \in C} P(w_1, w_2, \ldots w_{n_d}|c)P(c) && (4)\\
&= \operatorname*{argmax}_{c \in C} P(X_1 = w_1, X_2 = w_2, \ldots, X_{n_d} = w_{n_d}|c)P(c) && (5)
\end{aligned}
$$

MAP = maximum a posteriori; w_1 Wort in Vokabular in Position 1 im Dokument

Eigenschaften, die Klassifikation haben sollte

- Generalisierung zu Daten nicht im Trainingsset (ungesehene Daten)
- Fähigkeit zur Diskriminierung (verwechsle ähnliche Inputs mit verschiedenen Outputs nicht)
- Das volle Modell auf der letzten Folie entspricht einem table look up und kann nicht generalisieren

Ganzes Modell als Netzwerk



Wie könnte man das Modell simplifizieren?

Relaxing dependencies: Extremfall

Man könnte annehmen, dass die Antwortvariable überhaupt nicht von den Merkmalen abhängt!

$$c_{\mathrm{MFC}} = \operatorname*{argmax}_{c \in C} P(c)$$

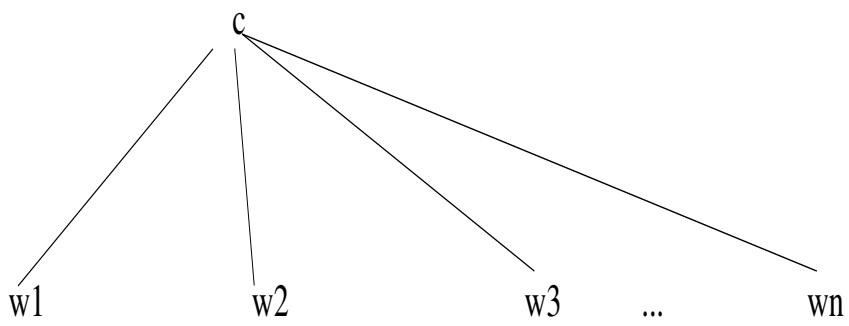
Most frequent class

Dies entspricht einem Algorithmus, der immer die häufigste Klasse jeder Testinstanz zuweist (maximum a posteriori). Oft **Baseline**, die der richtige Klassifizierer schlagen sollte

1. Wann ist dieser Algorithmus schwer zu schlagen?
2. Wann sollte man ihn vielleicht verwenden?

Relaxing dependencies: Naive Bayes

Berücksichtige nur die Dependenzen zwischen den Merkmalen (Wörtern) und der Klasse, aber nicht zwischen den verschiedenen Merkmalen?



Naive Bayes: Unabhängigkeitsannahmen

$$P(X_1 = w_1, X_2 = w_2, \ldots X_{n_d} = w_{n_d}|c)$$

- Conditional Independence: Nimm an, dass die Worte voneinander unabhängig sind, wenn die Klasse c gegeben ist.

$$\begin{aligned} P(X_1 = w_1, \ldots, X_{n_d} = w_{n_d}|c) &= P(X_1 = w_1|c) \cdot P(X_2 = w_2|c) \\ &\cdot \ldots \cdot P(X_{n_d} = w_{n_d}|c) \end{aligned}$$

- Bag of Words Annahme: Nimm an, dass Wortposition egal ist

$$P(X_k = w|c) = P(X_l = w|c)$$

für alle Klassen c, alle Positionen l, k and alle Worttypen w.

Wie lernt man den multinomialen NB Klassifizierer

1. Extrahiere Vokabular aus Trainingskorpus
2. Berechne $P(c)$ für alle c aus dem Trainingskorpus:

$$P(c) = \frac{\#docs\ of\ class\ c}{total\ \#\ of\ docs\ in\ training}$$

3. Berechne $P(w_i|c)$ für alle w_i im Vokabular und allen c:
   1. Konkateniere alle Trainingsdokumente der Klasse c zu einem langen Dokument
   2. $$P(w_i|c) = \frac{n_i^c}{n^c}$$

      wobei n_i^c Frequenz von w_i in langem Dokument und n^c Länge des langen Dokuments

Wie lernt man den multinomialen NB Klassifizierer

1. Extrahiere Vokabular aus Trainingskorpus
2. Berechne $P(c)$ für alle c aus dem Trainingskorpus:

$$P(c) = \frac{\#docs\ of\ class\ c}{total\ \#\ of\ docs\ in\ training}$$

3. Berechne $P(w_i|c)$ für alle w_i im Vokabular und allen c:
   1. Konkateniere alle Trainingsdokumente der Klasse c zu einem langen Dokument
   2. $$P(w_i|c) = \frac{n_i^c}{n^c}$$

      wobei n_i^c Frequenz von w_i in langem Dokument und n^c Länge des langen Dokuments

Wie lernt man den multinomialen NB Klassifizierer

1. Extrahiere Vokabular aus Trainingskorpus
2. Berechne $P(c)$ für alle c aus dem Trainingskorpus:

$$P(c) = \frac{\#docs\ of\ class\ c}{total\ \#\ of\ docs\ in\ training}$$

3. Berechne $P(w_i|c)$ für alle w_i im Vokabular und allen c:
   1. Konkateniere alle Trainingsdokumente der Klasse c zu einem langen Dokument
   2. $$P(w_i|c) = \frac{n_i^c}{n^c}$$

      wobei n_i^c Frequenz von w_i in langem Dokument und n^c Länge des langen Dokuments

Ein Problem: smoothing

Vermeide Nullen:

$$P(w_i|c) = \frac{n_i^c + 1}{n^c + |V|}$$

wobei $|V|$ Anzahl des Vokabulars

Testen des multinomialen NB

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in positions} P(w_i|c)$$

wobei wir über alle Positionen im Testdokument iterieren

Multinomialer Naive Bayes: Beispiel

	docID	words in doc	in c=China?
Training set	1	Chinese Bejing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

Lernphase: extrahiere Vokabular, schätze $P(c)$ und die bedingten Wahrscheinlichkeiten $p(w|c)$ im Trainingsset

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{5+1}{8+6} = \frac{3}{7}$
$p(Tokyo|China = yes) = \frac{0+1}{8+6} = \frac{1}{14}$
$p(Japan|China = yes) = \frac{0+1}{8+6} = \frac{1}{14}$
$p(Chinese|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$
$p(Tokyo|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$
$p(Japan|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$

Multinomialer Naive Bayes: Beispiel

Testphase für Testdokument 5 *Chinese Chinese Chinese Tokyo Japan*.

$$C_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i \in positions} P(w_i|c) = \operatorname*{argmax}_{c \in C} [\log P(c) + \sum_{i \in positions} \log P(w_i|c)]$$

$$P(China = yes|d) \propto \frac{3}{4} \cdot (\frac{3}{7})^3 \cdot \frac{1}{14} \cdot \frac{1}{14} = 0.0003$$
$$P(China = no|d) \propto \frac{1}{4} \cdot (\frac{2}{9})^3 \cdot \frac{2}{9} \cdot \frac{2}{9} = 0.0001$$

Zusammenfassung Multinomialer Naive Bayes

- Benutzt Unabhängigkeitsannahmen
- BoW: ignoriert Wortpositionen beim Schätzen
- Training: sieht alle Trainingsdokumente einer Klasse als ein langes Dokument
- In Training und Testing: Wortfrequenzen wichtig!
- Ignoriert beim Testen Vokabular, das nicht im Testdokument vorkommt

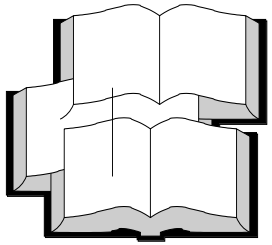
Binomialer Naive Bayes

Immer noch BoW und Unabhängigkeitsannahmen, aber nun nur Wortvorkommen ohne Häufigkeit

Test Doc

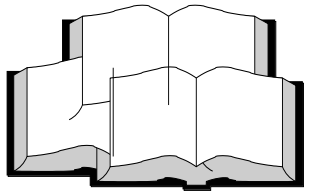
budget 1
hotel 1
price 1
breakfast 1
food 1
plot 0
screen 0
character 0
eggs 0
waiter 0
reception 0
cleaner 0

Hotels (50 doc)



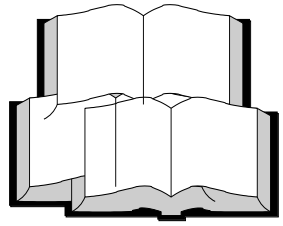
hotel 42
cleaner 6
reception 18
eggs 1
food 6

Restaurants (50 docs)



food 39
price 7
hotel 4
waiter 6

Movies (50 docs)



price 4
character 12
food 1
screen 5
plot 31

Binomialer Naive Bayes: Die Herleitung

$$
\begin{aligned}
C_{NB_{Bi}} &= \operatorname*{argmax}_{c \in C} P(c|d) && (6)\\
&= \operatorname*{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} && (7)\\
&= \operatorname*{argmax}_{c \in C} P(d|c)P(c) && (8)\\
&= \operatorname*{argmax}_{c \in C} P(e_1, \ldots, e_{|v|}|c)P(c) && (9)\\
&= \operatorname*{argmax}_{c \in C} P(c) \prod_{w_i \in V} P(e_i|c) && (10)
\end{aligned}
$$

wobei e_1 = yes/no, je nachdem ob w_1 im Dokument vorkommt

Binomialer Naive Bayes: Beispiel

	docID	words in doc	in c=China?
Training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{3+1}{3+2} = \frac{4}{5}$

$p(Japan|China = yes) = p(Tokyo|China = yes) = \frac{0+1}{3+2} = \frac{1}{5}$

$p(Bejing|China = yes) = p(Macao|China = yes) =$

$p(Shanghai|China = yes) = \frac{1+1}{3+2} = \frac{2}{5}$

$p(Chinese|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$

$p(Japan|China = no) = p(Tokyo|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$

$p(Bejing|China = no) = p(Macao|China = no) =$

$p(Shanghai|China = no) = \frac{0+1}{1+2} = \frac{1}{3}$

Binomialer Naive Bayes: Beispiel

	docID	words in doc	in c=China?
Training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{3+1}{3+2} = \frac{4}{5}$

$p(Japan|China = yes) = p(Tokyo|China = yes) = \frac{0+1}{3+2} = \frac{1}{5}$

$p(Bejing|China = yes) = p(Macao|China = yes) =$

$p(Shanghai|China = yes) = \frac{1+1}{3+2} = \frac{2}{5}$

$p(Chinese|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$

$p(Japan|China = no) = p(Tokyo|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$

$p(Bejing|China = no) = p(Macao|China = no) =$

$p(Shanghai|China = no) = \frac{0+1}{1+2} = \frac{1}{3}$

Binomialer Naive Bayes: Beispiel

	docID	words in doc	in c=China?
Training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{3+1}{3+2} = \frac{4}{5}$
$p(Japan|China = yes) = p(Tokyo|China = yes) = \frac{0+1}{3+2} = \frac{1}{5}$
$p(Bejing|China = yes) = p(Macao|China = yes) =$
$p(Shanghai|China = yes) = \frac{1+1}{3+2} = \frac{2}{5}$
$p(Chinese|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$
$p(Japan|China = no) = p(Tokyo|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$
$p(Bejing|China = no) = p(Macao|China = no) =$
$p(Shanghai|China = no) = \frac{0+1}{1+2} = \frac{1}{3}$

Binomialer Naive Bayes: Beispiel

Testphase für Testdokument 5 *Chinese Chinese Chinese Tokyo Japan*.

$$C_{NB_{Bi}} = \operatorname*{argmax}_{c \in C} P(c) \prod_{w_i \in V} P(e_i|c) = \operatorname*{argmax}_{c \in C} [\log P(c) + \sum_{w_i \in V} \log P(e_i|c)]$$

$$\begin{aligned} P(China = yes|d) &\propto P(c) \cdot P(Chinese|China = yes) \cdot P(Japan|China = yes) \\ &\cdot P(Tokyo|China = yes) \\ &\cdot (1 - P(Bejing|China = yes)) \cdot (1 - P(Shanghai|China = ye \\ &\cdot (1 - P(Macao|China = yes)) \\ &= \frac{3}{4} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot (1 - \frac{2}{5}) \cdot (1 - \frac{2}{5}) \cdot (1 - \frac{2}{5}) \\ &= 0.005 \end{aligned}$$

Auf gleiche Weise $P(China = no|d) = 0.022$

Binomial vs Multinomial

	multinomial	binomial
Zufallsvariable	$X = w$ wenn w an pos X	$e_w = 1$ wenn w in Doc
doc rep	Sequenz von Worthäuf.	Sequenz von 0, 1
Mehrfachvork..	ja	ignoriert
Doklänge	gut für länger	nur für kurze
$\|V\|$	auch für groß	besser klein
Schätzung f. *the*	$p(X = the\|c) = 0.05$	$p(e_{the} = 1\|c) = 1.0$

Zusammenfassung NB

- Sehr schnell, braucht wenig Speicher
- Multinomial: gut beim Ignorieren irrelevanter Merkmale
- Gut, wenn Merkmale gleich wichtig
- gute Baseline für Textklassifikation
- Wie müssen sich die Merkmale ändern, wenn man nicht nach Themen klassifiziert?

Evaluation von Klassifikation: contingency table

Wenn man eine Klasse C betrachtet:

	wirklich C	nicht C
klassifiziert als C	tp	fp
nicht klassifiziert als C	fn	tn

- Accuracy: $\frac{tp+tn}{tp+fp+tn+fn}$
- Fehlerrate = 1 - Accuracy
- Precision: $\frac{tp}{tp+fp}$
- Recall: $\frac{tp}{tp+fn}$
- F-measure: $\frac{2 \cdot Pr \cdot Rec}{Pr+Rec}$

Mehr als zwei Klassen: Menge von binären Klassifizierern

- Bei any-of Klassifizierern
- Für jede Klasse $c \in C$: baue Klassifizierer γ_c um c von allen anderen Klassen zu unterscheiden
- Geg. Testdoc d, evaluiere die Klassenzugehörigkeit mit jedem γ_c. Dann gehört d zu jeder Klasse c für die γ_c "ja" zurückgibt.

Mehr als zwei Klassen: Mengen binärer Klassifizierer

- One-of Klassifizierer (Klassen schliessen sich gegenseitig aus)
- Für jede Klasse $c \in C$: baue Klassifizierer γ_c wie auf vorheriger Folie
- Geg. Testdok. d, evaluiere die Klassenzugehörigkeit mit jedem γ_c. Dann gehört d zu der Klasse mit dem maximalem Score.

Beispielsevaluation: Reuters-21578 dataset

- 21,578 docs
- ModApte split: 9,603 training, 3,299 test docs
- 118 Kategorien: Artikel kann in mehr als einer Kategorie sein
- Nur um die 10 Kategorien sind groß

class	#train	#test	class	#train	#test
earn	2877	1087	trade	369	119
acquisitions	1650	179	interest	347	131
money-fx	538	179	ship	197	89
grain	433	149	wheat	212	71
crude	389	189	corn	182	56

- das durschnittliche Dokument gehört zu 1.24 Klassen

Ein typisches Reuters Dokument

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-
NEWID="798">
<DATE> 2-MAR-1987 16:51:43.42</DATE>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American
Pork Congress kicks off tomorrow,
March 3, in Indianapolis with 160 of the nations pork produc
from 44 member states determining industry positions
on a number of issues, according to the National
Pork Producers Council, NPPC.
Delegates to the three day Congress will be considering
26 resolutions concerning various issues, including the futu
direction of farm policy and the tax law
 as it applies to the agriculture sector. The delegates will
endorse concepts of a national PRV (pseudorabies virus)
```

Micro vs. Macro-averaging

Man kann Precision/Recall per Klasse berechnen, aber wie kombiniert man die vielen Maße?

Macroaveraging

Berechne per Klasse und dann bilde den Durchschnitt

Microaveraging

Sammle Entscheidungen für alle Instanzen für alle Klasse, berechne contingency table, evaluiere

Micro vs. Macroaveraging

Class1

	Truth: y	Truth: n
classifier=y	10	10
classifier =n	10	970

Class2

	Truth: y	Truth: n
classifier=y	90	10
classifier =n	10	890

Micro-averaged table

	Truth: y	Truth: n
classifier=y	100	20
classifier =n	20	1860

- Macroaveraged precision: (0.5+0.9)/2=0.7
- Microaveraged precision: 100/120 = 0.83
- Microaveraged precision wird von häufigen Klassen dominiert

Andere Evaluationsthemen

- Training vs. Test set
- Grösse Trainings-, Testset
- Learning curves (Lernkurven)
- Cross-validation

Crossvalidation/Kreuzvalidierung

Cross-Validation

- Teile Daten zufällig in k Teilmengen (folds) von gleicher Größe (z.B. $k = 10$)
- Trainiere Modell auf $k - 1$ folds, nimm ein fold fürs Testing
- Wiederhole k mal
- Berechne durschnittliche Performanz auf k Testsets.

Crossvalidation

Beispiel: 100 Instanzen, 5-fold crossvalidation

test set	training set

$i_1, i_2, i_3, i_4, \ldots i_{20}$
$i_{21}, i_{22}, i_{23}, i_{24} \ldots i_{40}$
$i_{41}, i_{42}, i_{43}, i_{44}, \ldots i_{60}$
$i_{61}, i_{62}, i_{63}, i_{64} \ldots i_{80}$
$i_{81}, i_{82}, i_{83}, i_{84} \ldots i_{100}$

Fehlerrate: 20%

Crossvalidation

Beispiel: 100 Instanzen, 5-fold crossvalidation

test set	training set

$i_1, i_2, i_3, i_4, \ldots i_{20}$
$i_{21}, i_{22}, i_{23}, i_{24} \ldots i_{40}$
$i_{41}, i_{42}, i_{43}, i_{44}, \ldots i_{60}$
$i_{61}, i_{62}, i_{63}, i_{64} \ldots i_{80}$
$i_{81}, i_{82}, i_{83}, i_{84} \ldots i_{100}$

error rate: 25%

Crossvalidation

Beispiel: 100 Instanzen, 5-fold crossvalidation

test set	training set

$i_1, i_2, i_3, i_4, \ldots i_{20}$
$i_{21}, i_{22}, i_{23}, i_{24} \ldots i_{40}$
$i_{41}, i_{42}, i_{43}, i_{44}, \ldots i_{60}$
$i_{61}, i_{62}, i_{63}, i_{64} \ldots i_{80}$
$i_{81}, i_{82}, i_{83}, i_{84} \ldots i_{100}$

Fehlerrate: 15%

Crossvalidation

Beispiel: 100 Instanzen, 5-fold crossvalidation

test set	training set

$i_1, i_2, i_3, i_4, \ldots i_{20}$
$i_{21}, i_{22}, i_{23}, i_{24} \ldots i_{40}$
$i_{41}, i_{42}, i_{43}, i_{44}, \ldots i_{60}$
$i_{61}, i_{62}, i_{63}, i_{64} \ldots i_{80}$
$i_{81}, i_{82}, i_{83}, i_{84} \ldots i_{100}$

Fehlerrate: 10%

Crossvalidation

Beispiel: 100 Instanzen, 5-fold crossvalidation

test set	training set

$i_1, i_2, i_3, i_4, \ldots i_{20}$
$i_{21}, i_{22}, i_{23}, i_{24} \ldots i_{40}$
$i_{41}, i_{42}, i_{43}, i_{44}, \ldots i_{60}$
$i_{61}, i_{62}, i_{63}, i_{64} \ldots i_{80}$
$i_{81}, i_{82}, i_{83}, i_{84} \ldots i_{100}$

Fehlerrate: 30%

durchschnittliche Fehlerrate: 20%

Kreuzvalidierung

***k*-fache Kreuzvalidierung** maximiert den Nutzen der Daten und kann durch Varianz auch akkuratere Schätzungen der Performanz bieten.

`http://www.cs.bham.ac.uk/~pxt/NC/ASSIGNMENT/MICHAEL/crossValid.html`

Fazit Evaluation

Wir brauchen immer ein unabhängiges Testset.

Performanz auf dem Trainingsset ist kein guter Indikator für die Performanz auf ungesehenen Daten.

Das Testset kann ein **unabhängiges Sample** sein oder **mittels Holdout Techniken** (Kreuzvalidierung).

Merkmale

BoW:

- Nur Worte
- bis jetzt alle Wörter im Text ODER wir haben offen gelassen, wie man das Vokabular auswählt
- Warum könnte es besser sein, nicht alle Wörter zu benutzen?

Warum Feature Selection?

- Textkollektionen haben sehr große Merkmalsanzahl
- Einige Klassifiziere können mit großer Merkmalsanzahl nicht umgehen: NB aber schon
- Reduziert Trainingszeit
- Kann Generalisierung verbessern
  - eliminiert Rauschen
  - vermeidet overfitting
- Binomialer NB besonders anfällig für Rauschen
- Eine Option: eliminiere alle Funktionswörter (stop words)

Grundidee

Feature selection/Merkmalsselektion

Berechne ein Maß $A(w, c)$ für jedes Wort und jede Klasse c und selektiere die k Terme mit den höchsten Werten von $A(t, c)$

Wir konzentrieren uns auf drei Maße und zeigen diese für zwei Klassen.

Häufigkeit als Merkmalsselektion

- Benutze einfach die *k* häufigsten Wörter
- Exkludiert informative Worte wie *super-entertaining*
- Warum ist es in der Praxis oft trotzdem ein gutes Verfahren?

Mutual Information zur Merkmalsselektion

Wieviel trägt ein Term zu korrekter Klassifizierung bei?

- Idee: vergleiche beobachtete Wahrscheinlichkeiten mit erwarteten Wahrscheinlichkeiten, wenn Term und Klasse unabhängig werden
- Formel:

$$I(T;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(T = e_t, C = e_c) \log \frac{P(T = e_t, C = e_c)}{P(T = e_t) P(C = e_c)}$$

Ein Assoziationsmaß: Mutual Information

$$I(T;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(T = e_t, C = e_c) \log \frac{P(T = e_t, C = e_c)}{P(T = e_t)P(C = e_c)}$$

Beispiel mit der Klasse `poultry` und dem Term *export* im reuters-Korpus

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$	Marginals
$e_t = e_{export} = 1$	49	27,652	27, 701
$e_t = e_{export} = 0$	141	774,106	774, 247
	190	801,758	801,948

Einige Beispielberechnungen

- $P(e_c = 1, e_t = 1) = \frac{49}{801948}$
- $P(e_c = 1) = \frac{190}{801948}$
- $P(e_t = 0) = \frac{774{,}347}{801{,}948}$

Ein Assoziationsmaß: Mutual Information

Zusammen:

$$
\begin{aligned}
I(T;C) &= \frac{49}{801,948} \log \frac{801,948 \cdot 49}{(49+27,652)(49+141)} \\
&+ \frac{141}{801,948} \log \frac{801,948 \cdot 141}{(141+774,106)(49+141)} \\
&+ \frac{27,652}{801,948} \log \frac{801,948 \cdot 27,652}{(49+27,652)(27,652+774,106)} \\
&+ \frac{774,107}{801,948} \log \frac{801,948 \cdot 774,106}{(141+774,106)(27,652+774,106)} \\
&= 0.0001105
\end{aligned}
$$

Mutual information für drei Reuters-Klassen

UK class	sports class	poultry class
london 0.192	soccer 0.0682	poultry 0.0013
uk 0.0755	cup 0.0515	meat 0.0008
british 0.0596	match 0.0441	chicken 0.0006
stg 0.0555	matches 0.0408	agriculture 0.0005
britain 0.0469	played 0.0388	avian 0.00004
plc 0.0357	league 0.0386	broiler 0.0003
england 0.0238	beat 0.0301	

Daten aus Manning et al: Introduction to IR

χ^2 Merkmalsselelektion (Optional)

- Bestimme Unabhängigkeit der beiden Ereignisse: Vorkommen des terms (Merkmals) und Vorkommen der Klasse
- Vergleiche wirklich beobachtetes Vorkommen mit erwartetem Vorkommen, wenn die beiden Ereignisse unabhängig wären
- Beruht ebenfalls auf der 2x2 Matrix

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$	Marginals
$e_t = e_{export} = 1$	49	27,652	27, 701
$e_t = e_{export} = 0$	141	774,106	774, 247
	190	801,758	801,948

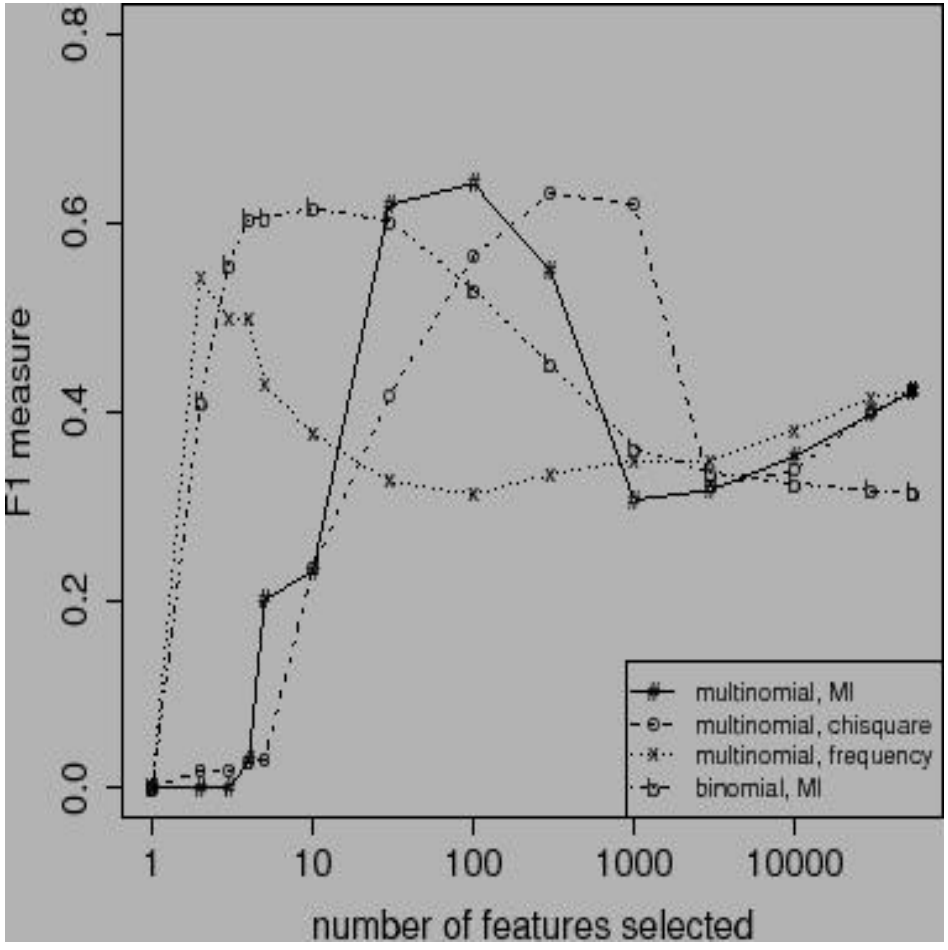
bzw

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$	Marginals
$e_t = e_{export} = 1$	N11	N10	N1.
$e_t = e_{export} = 0$	N01	N00	N0.
	N.1	N.0	N

Hierzu müssen Sie Kapitel 13.5.2 in Manning et al, Introduction to IR lesen. Online auch unter `https://nlp.stanford.edu/IR-book/`

Effekt auf Performanz

Für 5 Klassen yes/no (F-measure über 5 Klassen) aus Manning, Raghavan, Schuetze: Introduction to Information retrieval
100K Docs im Training und 100K im Testing



Sentimentklassifikation: Daumen hoch?

Original text

A good budget hotel' .... Price includes breakfast. Rooms are modern and of a reasonable size. The centre of Leeds is about a 15 min walk at the most. Hotel has bar area.

Würden Sie andere Merkmale als für Themenklassifikation benutzen?

Textklassifikation: spam?

Subject: Conference on the Government???s communications strategy
From: Edward Rees
To: Katja Markert

Dear Dr Markert

I hope you won???t mind this final reminder about the above semina
taking place in Central London on Tuesday, 29th October 2013, but
you don???t currently appear to be represented. Please note there
a charge for most delegates, although concessionary and
complimentary places are available (subject to terms
and conditions - see below).

The focus:

Merkmale in SpamAssassin

- Mentions Generic Viagra
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase:impress...girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area
- Claims you can be removed from the list

`http://spamassassin.apache.org/`

Autorenidentifikation

- Berühmtes Problem: Federalist papers 1787-8.
- 12 Briefe schwer zuortbar
- 1963: Mosteller and Wallace mit Bayes-Methoden

Geschlecht/Alterklassifikation

1. The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. . . . The methods proposed are intended to enable students to obtain insights into aspects of cohesion
2. My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding . . . In this paper I follow Sperber and Wilson's suggestion that . . .

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. Gender, Genre, and Writing Style in Formal Written Texts, Text, volume 23, number 3, pp. 321– 346

See also: K. Filippova. User demographics and language in an implicit social network. EMNLP 12. Jeju, Korea, July 12-14, 2012.

Demo: `http://www.hackerfactor.com/GenderGuesser.php`

Fazit

In der Praxis

- braucht man Merkmalsselektion sowie gutes Smoothing
- muss man den Algorithmus auf mehrere Klassen anpassen
- vorsichtig evaluieren
- nciht nur Wortmerkmale benutzen

References

- Jurafsky and Martin: 3rd online edition, Chapter 4
- *Manning, Raghavan and Schuetze: Introduction to Information Retrieval. Chapter 13
- Aufgabenblatt 5/6

Eine Warnung: Jurafsky und Martin besprechen multinomial, aber nicht binomial, sondern statt dem letzteren eine Mischform aus multinomial und binomial, die multinomial binary genannt wird.