

Statistische Sprachverarbeitung

Sebastian Pado

27.01.2004

Übersicht

- Warum Statistik?
- Statistik
 - Einige grundlegende Konzepte
- Statistische Sprachverarbeitung
 - In 5 Schritten zum erfolgreichen Statistiker
- Konkretes Beispiel: PP Attachment
- (Nächstes Mal: Probabilistische kontextfreie Grammatiken)

Formalisierung von Wissen: Symbolische Modelle

Modelle

■ Formale Sprachen

- Reguläre Ausdrücke
- Kontextfreie Grammatiken
- Komplexere Grammatiken

■ Logiken

- Sammlung von Aussagen, Regeln, ...
- Aussagenlogik, Prädikatenlogik, Baumlogiken, Temporallogiken...
- Gewichtete Regeln:
 - Wenn A, dann B (Gewicht 1),
Wenn A, dann C (Gewicht 3)

Verarbeitung

- Endliche Automaten
- Kellerautomaten / Parser
- Komplexere
Verarbeitungsmodelle
- Verschiedene Methoden
- Je komplexer die Logik, desto
komplexer die Verarbeitung

Probleme von symbolischen Modellen

- Im Allgemeinen binär (ja oder nein)
 - Syntax: Grammatisch / ungrammatisch
 - Ich kenne ihn, weil wir uns letztes Jahr getroffen haben.
 - (?)Ich kenne ihn, weil wir haben uns letztes Jahr getroffen.
 - Spracherwerb und –entwicklung sind kontinuierliche Prozesse
- Vollständigkeit
 - Jede Regel hat eine Ausnahme
 - Handkodierte Regeln haben nie 100% Abdeckung auf einem Korpus
- Aufwand
 - Eine große Grammatik zu schreiben ist sehr, sehr, sehr aufwändig
 - Schwer zu überblicken (Interaktionen)
 - Von Hand gewichtete Regeln: schwer zu kontrollieren

Attraktives Konzept: lernen

Lernen von symbolischen Modellen

■ Schwer

- „Echtes“ Lernen von kontextfreien Grammatiken benötigt negative Beispiele
 - Noch schwerer, „linguistisch sinnvolle“ Grammatiken zu erhalten

■ Allgemeines, mathematisch einfaches Framework nötig

- Statistische Sprachverarbeitung
 - „Gewichtete Regeln“ mit automatisch gelernten Gewichten

2. Statistik – die grundlegende Idee

- Statistik beschreibt Ergebnisse vergangener („gesehener“) Experimente und sagt damit Ergebnisse neuer („ungesehener“) Experimente voraus
- Menge der möglichen Elementarereignisse Ω
 - Werfen eines Würfels: $\Omega = \{1,2,3,4,5,6\}$
- Jedes Elementarereignis $\omega \in \Omega$ hat eine Wahrscheinlichkeit $P(\omega)$:

$$P(\omega) = \frac{f(\omega)}{N}$$

 - Ermittlung z.B. durch
 - Beobachtung von Frequenzen bisheriger Experimente
 - Annahmen über Experiment
 - Werfen eines Würfels: $P(\omega) = 1/6$ für alle ω .

Vorteil dieser Modellierung

- Abgestufte Regeln
 - Experiment X mit möglichen Ausgängen Z_1 (Wahrscheinlichkeit p_1), Z_2 (Wahrscheinlichkeit p_2) entspricht Regelschema
 - $X \rightarrow Z_1 (p_1)$
 - $X \rightarrow Z_2 (p_2)$
- Keine binäre Entscheidung mehr
 - Beliebige viele Ausgänge möglich
- Aus den Daten lernbar
 - Beobachtungen bisheriger Experimente

Beispiel:

Wortart des ersten Wort von Sätzen

- Alte Experimente:
 - Extrahiere 10 Sätze aus einem Korpus
 - Beobachtete Frequenzen für Wortarten der ersten Worte:
Det 5; NE 2; Konj 2; NN 1
 - Neues Experiment: Ziehe „erste Wortart“ des 11. Satzes
 - Ereignisraum $\Omega = \{\text{Det}, \text{NE}, \text{Konj}, \text{NN}\}$
 - Elementarereignisse und Wahrscheinlichkeiten:
 - $P(\text{Det}) = f(\text{Det})/10 = 5/10 = 1/2$
 - $P(\text{NE}) = f(\text{NE})/10 = 2/10 = 1/5$
 - $P(\text{Konj}) = f(\text{Konj})/10 = 2/10 = 1/5$
 - $P(\text{NN}) = f(\text{NN})/10 = 1/10$
-

Wahrscheinlichkeiten

- Ereignisse E sind Mengen von Elementarereignissen
 - Wahrscheinlichkeiten für Ereignisse:
 - Frequenzen $P(\text{Ereignis } E) = \frac{f(E)}{N}$
 - Aus den anderen Ereignissen (siehe unten)
- Axiome für Wahrscheinlichkeiten von Ereignissen:
 - $P(E) \geq 0$ für alle A
 - $P(\Omega) = 1$
 - Wenn E_1, \dots, E_n disjunkt, dann $P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i)$
 - Disjunkt: keine gemeinsamen Elementarereignisse

Beispiel

- $P(\text{Det}) = 1/2$
 - $P(\text{Konj}) = 1/5$
 - $P(\text{NE}) = 1/5$
 - $P(\text{NN}) = 1/10$
- Elementarereignisse
-
- $P(A) = P(\text{Det oder NE}) = (5+2)/10 = 7/10$ Ereignis 1
 - $P(A) = P(\text{Det}) + P(\text{NE}) = 1/2 + 1/5 = 7/10$
-
- $P(B) = P(\text{Konj oder NN}) = (2+1)/10 = 3/10$ Ereignis 2
 - $P(C) = P(\text{Konj oder NE}) = (2+2)/10 = 4/10$ Ereignis 3

Beispiel

- $P(\text{Det}) = 1/2$
- $P(\text{Konj}) = 1/5$
- $P(\text{NE}) = 1/5$
- $P(\text{NN}) = 1/10$
- $P(A) = P(\text{Det oder NE}) = 7/10$
- $P(B) = P(\text{Konj oder NN}) = 3/10$
- $P(C) = P(\text{Konj oder NE}) = 4/10$
- $P(A \text{ oder } B) = (5+2+2+1)/10 = 7/10 + 3/10 = 1$ $\begin{matrix} A \text{ oder } B \\ = \Omega \end{matrix}$
- $P(A \text{ oder } C) = (5+2+2)/10 = 9/10$
- $P(A \text{ oder } C) = 7/10 + 4/10 = 11/10$ $\begin{matrix} A \text{ und } C \text{ sind} \\ \text{nicht disjunkt!} \end{matrix}$

Wahrscheinlichkeitsverteilungen

- Eine Wahrscheinlichkeitsverteilung ordnet jedem Ereignis E eine Wahrscheinlichkeit zu

- Beliebte Verteilungen:
 - Gleichverteilung: Jedes Ereignis ist gleich wahrscheinlich
 - Augenzahl beim Würfeln mit einem Würfel
 - Normalverteilung: Gaußsche Glockenkurve
 - Wahrscheinlichkeit für Kopf beim Werfen einer Münze (unendlich viele Würfe)

Gemeinsame Wahrscheinlichkeit

- $P(A,B)$ ist die Wahrscheinlichkeit, daß sowohl Ereignis A als auch Ereignis B eintreten
 - Kann sowohl gleichzeitig als auch sequentiell sein
 - $P(\text{Regen, Sonntag})$
 - $P(\text{Det, NN, VI})$
 - Kann oft im Experiment beobachtet werden

Bedingte Wahrscheinlichkeit

- Bedingte Wahrscheinlichkeiten setzen zwei Ereignisse in Beziehung
 - Typischerweise: was wir gesehen haben und was wir sehen werden
- $P(A | B)$: die Wahrscheinlichkeit, A zu sehen, wenn wir schon B gesehen haben („A gegeben B“)
- B heißt das konditionierende Ereignis (Konditionierung)

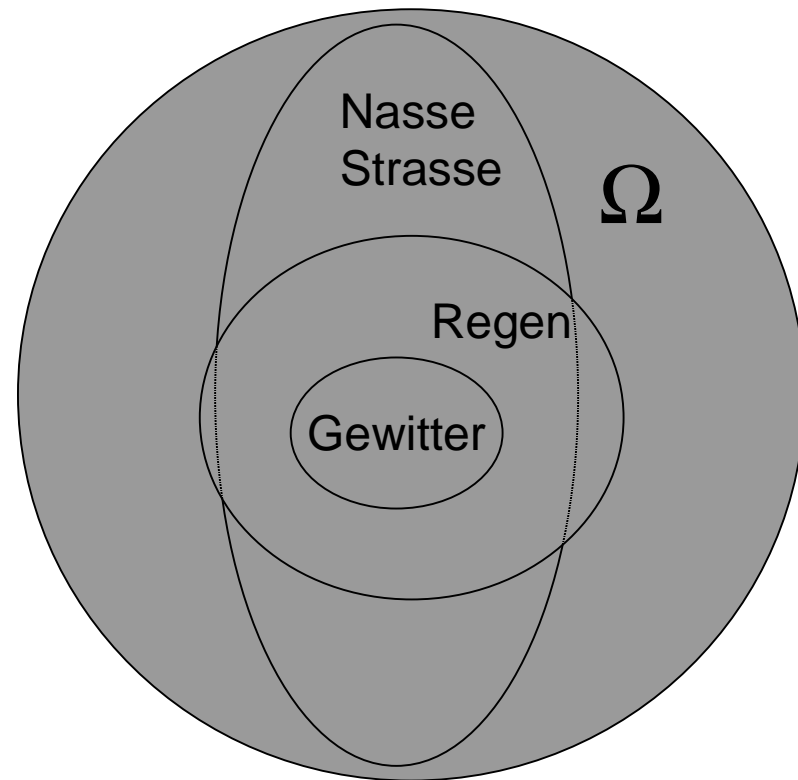
$$P(A|B) = \frac{P(A,B)}{P(B)}$$

Gemeinsame
Wahrsch.

Bedingte Wahrscheinlichkeit

- Deutung als Experiment: $P(A|B)$ ist die Wahrscheinlichkeit von A, wenn man die Welt kleiner macht (von Ω auf B einschränkt)
- Geometrische Deutung: $P(A|B)$ ist Anteil der Schnittfläche von A und B an der Fläche von B
- $P(\text{Nasse Straße} | \text{Regen}) = 0,9$
 - „Wenn Regen, dann mit hoher Wahrscheinlichkeit Nasse Straße“
- $P(\text{Regen} | \text{Nasse Straße}) = 0,5$
 - „Wenn Nasse Straße, dann in 50% der Fälle auch Regen“

- $P(\text{Gewitter} | \text{Regen}) = 0,2$
- $P(\text{Regen} | \text{Gewitter}) = 1$



Umrechnen zwischen bedingten Wahrscheinlichkeiten: Satz von Bayes

- Wie verhalten sich $P(A | B)$ und $P(B | A)$?

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A, B) = P(A|B)P(B)$$

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

$$P(A, B) = P(B|A)P(A)$$

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$$

Unabhängigkeit

- Zwei Ereignisse sind voneinander unabhängig, wenn sie sich gegenseitig nicht beeinflussen
 - genau dann wenn
$$P(A \mid B) = P(A) \text{ und } P(B \mid A) = P(B)$$
 - Intuition: „ob B eintritt oder nicht, ändert nichts an $P(A)$ “
- Für unabhängige Ereignisse ist die gemeinsame Wahrscheinlichkeit einfach zu berechnen:
$$P(A, B) = P(A) * P(B)$$
 - Erklärung: Bei Unabhängigkeit gilt
$$P(A, B) \stackrel{\text{df}}{=} P(A \mid B) * P(B) = P(A) * P(B)$$

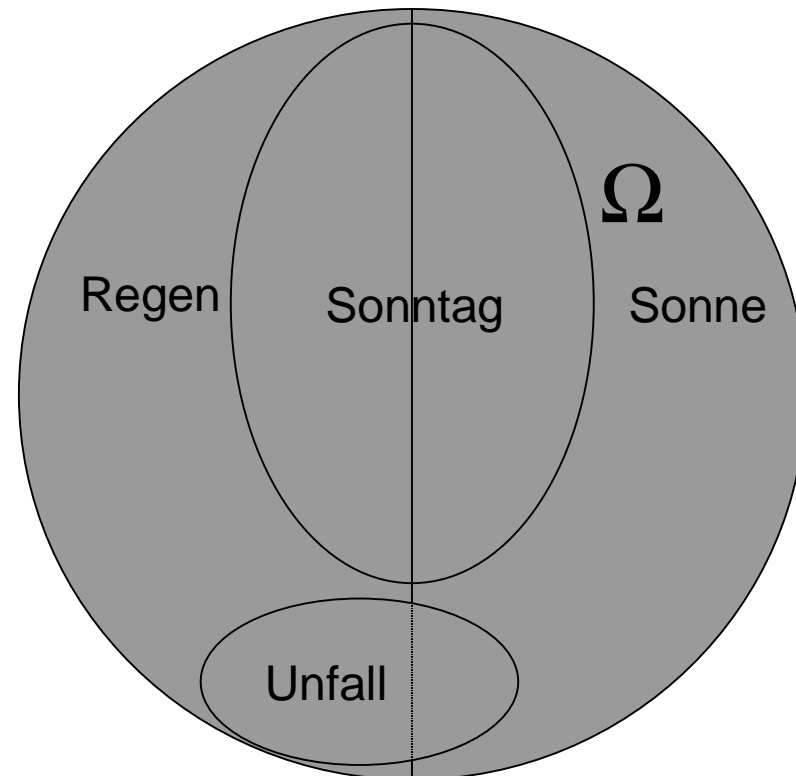
Unabhängigkeit

■ Beispiel:

- $P(\text{Sonne scheint} \mid \text{Sonntag}) = P(\text{Sonne scheint})$
- $P(\text{Sonntag} \mid \text{Sonne scheint}) = P(\text{Sonntag})$
-) Unabhängigkeit

■ Gegenbeispiel:

- $P(\text{Unfall} \mid \text{Regen}) > P(\text{Unfall})$
-) Nicht unabhängig



Gemeinsame Wahrscheinlichkeit und Unabhängigkeit

■ Beispiel:

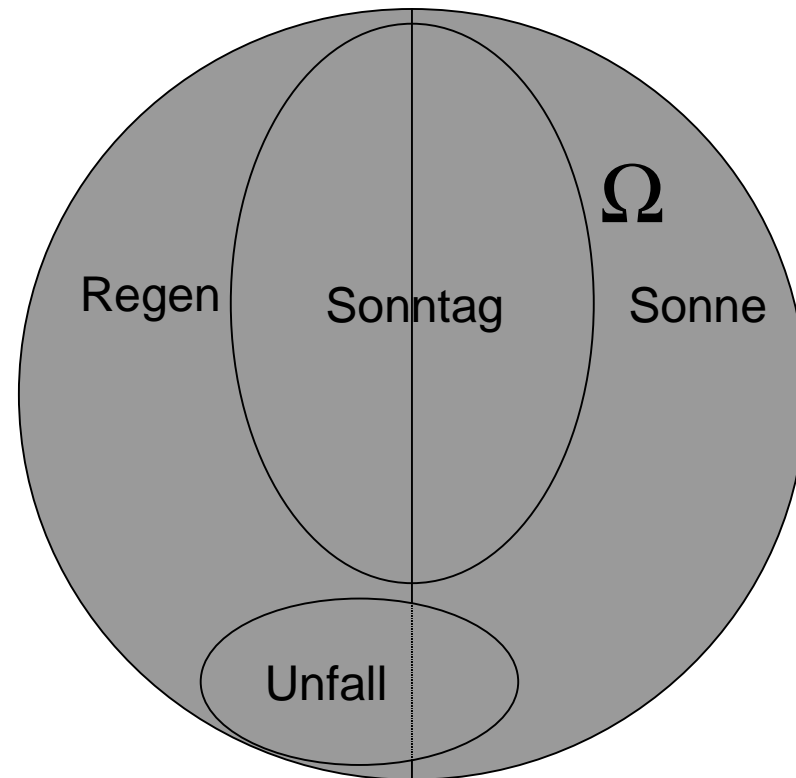
- $P(\text{Sonne, Sonntag}) = P(\text{Sonne}) * P(\text{Sonntag}) = 1/2 * 1/7 = 1/14$

■ Gegenbeispiel:

- $P(\text{Unfall, Regen}) = 1/10$

?

- $P(\text{Unfall}) * P(\text{Regen}) = 1/8 * 1/2 = 1/16$



Umrechnen zwischen gemeinsamen und bedingten Wahrscheinlichkeiten

- Wie verhalten sich $P(A,B)$ und $P(A | B)$?
 - Definition der bedingten Wahrscheinlichkeit:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

- Mehr als zwei Ereignisse: Kettenregel:

$$P(A_1, A_2, A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)$$

- Wahrscheinlichkeit von A_1 : $P(A_1)$
 - Angenommen, A_1 tritt ein, Wahrscheinlichkeit von A_2 : $P(A_2 | A_1)$
 - Angenommen, A_2 tritt ein, Wahrscheinlichkeit von A_3 : $P(A_3 | A_1, A_2)$
- Vereinfachung durch Unabhängigkeit

3. Statistische Modellierung von Sprache

- Man kann oft gemeinsame Wahrscheinlichkeiten beobachten
 - Z.B. Worte + Wortarten $P(W, POS)$ aus Korpus
- Interessant sind meistens bedingte Wahrscheinlichkeiten
 - Für jedes Wort W_i : wie wahrscheinlich ist $P(POS_1 | W_i), P(POS_2 | W_i), P(POS_3 | W_i), \dots?$
- Worte $W_1 \dots W_n$ heißen Instanzen
- $POS_1 \dots POS_n$ heißen Klassen

Klassifikation

- Konditionaler statistischer Klassifikator:

$$P(\text{Klasse } K \mid \text{Instanz } I) = \\ P(\text{Klasse } K \mid \text{Features } \underline{f}_I)$$

- Die Entscheidung kann normalerweise nicht allein aufgrund der Instanz getroffen werden
 - Mehr Information: Repräsentation durch Features repräsentiert
- Entscheidungsalgorithmus: Liefere die Klasse K , die am wahrscheinlichsten für die Instanz I (mit den Features \underline{f}_I) ist:
 $K = \operatorname{argmax}_K P(K|\underline{f}_I)$
- Interpretation als Experiment:
 - Es gibt so viele Töpfe, wie es Instanzen (Kombinationen von Features) gibt
 - Beim Ziehen aus dem Topf \underline{f}_I erscheint die Klasse K mit $P(K|\underline{f}_I)$
 - $\operatorname{argmax}_K P(K|I)$ ist die wahrscheinlichste Klasse beim Ziehen aus \underline{f}_I

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden (haben wir schon!)
- Schritt 5: Modell evaluieren

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden
- Schritt 5: Modell evaluieren

Schritt 2: Features

- Das konditionierende Ereignis muß näher beschrieben werden, z.B.

$P(\text{Wortart} \mid \text{Wort, letztes Wort, vorletztes Wort})$

- Features repräsentieren das „gesehenen“ Ereignis, auf dessen Grundlage die Wahrscheinlichkeit der Klasse berechnet wird
 - Wahl von Features ist sehr wichtig
 - Welche Information hilft, die Entscheidung zu treffen?
 - Manche statistischen Modelle funktionieren nur mit wenigen Features gut

Features für Wortartenzuordnung

- Der Mann geht schnell.
 - Features:
 - Wort
 - letztes Wort
 - Wortart des letzten Wortes
 - Klassifikator: $\text{POS}(w) = \text{argmax}_p P(p \mid w, w_h, p_h)$
 - Verschiedene Features helfen an verschiedene Stellen, um von gesehenen Beispielen auf ungesehene Beispiele zu schließen:
 - $P(K \mid \text{der}, x, x)$ „der“ ist nur als Det aufgetaucht) $K=\text{Det}$
 - $P(K \mid \text{Mann}, \text{der}, \text{Det})$ Großschreibung; letztes POS Det) $K=\text{NN}$
 - $P(K \mid \text{geht}, \text{Mann}, \text{NN})$ Personalendung; letztes POS NN) $K=\text{V}$
 - $P(K \mid \text{schnell}, \text{geht}, \text{V})$ letztes Wort ist „geht“ (Verb)) $K=\text{Adv}$

Erinnerung: Statistik als Regeln mit automatisch gelernten Gewichten

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden
- Schritt 5: Modell evaluieren

Schritt 3: Training von Modellen

- Aus den Daten muß die Wahrscheinlichkeit $P(c|f_1, \dots, f_n)$ für jede Klasse c und jede Kombination von Features f_1, \dots, f_n abgeschätzt werden (Estimation)

- Beispiele am Anfang der Vorlesung: Über Frequenz
 - Maximum Likelihood Estimation (MLE):

$$P(c|f_1, \dots, f_n) = \frac{P(c, f_1, \dots, f_n)}{P(f_1, \dots, f_n)} = \frac{f(c, f_1, \dots, f_n)}{f(f_1, \dots, f_n)}$$

- Was ist, wenn Kombination (f_1, \dots, f_n) nie gesehen wurde?

Sparse Data

- Features haben große Wertebereiche
 - Worte: jedes Feature hat 100 000 (?) Werte
 - Kombination von Features (Featureraum): Multiplikation
 - Wortartenklassifikation: 3 Features, (W,W_h,POS_h) hat $100\,000 \times 100\,000 \times 30 = 300$ Milliarden Werte
 - Kann nie alle Kombinationen aller Features sehen!
 - $P(\text{ungesehene Featurekombination}) = \text{nicht definiert}$
- Kompromiß bei MLE: Kombination mehrerer Modelle
 - Viele Features: Genaues Modell, schlechte Abdeckung
 - Wenige Features: Schlechteres Modell, bessere Abdeckung
 - Kombination mit Back-off
- Smoothing (Verteilen von Wahrscheinlichkeitsmasse auf ungesehene Ereignisse)

Andere statistische Modelle vermeiden direkte Abschätzung

Beispiel: Semantic Role Assignment

Distribution	Coverage %	Accuracy %	Performance
$P(r \mid t)$	100	40,9	40,9
$P(r \mid pt, t)$	92,5	60,1	55,6
$P(r \mid pt, g, v)$	92,0	66,6	61,3
$P(r \mid pt, p, v)$	98,8	57,1	56,4
$P(r \mid pt, p, v, t)$	90,8	70,1	63,7
$P(r \mid h)$	80,3	73,6	59,1
$P(r \mid h, t)$	56,0	86,6	48,5
$P(r \mid h, pt, t)$	50,1	87,4	43,8

Quelle: Diss D. Gildea

Andere Modelle

- Maximum Entropy Models
 - Jedes Feature hat ein Gewicht
 - Optimierte Gewichte
 - Wahrscheinlichkeiten: Summe über Gewicht*Feature
- Support Vector Machines
 - Hochdimensionaler Raum (Dimensionen = Features)
 - Finde Hyperebenen, die die Klassen trennen
 - Wahrscheinlichkeit für Klasse = Abstand von Ebene
- Hidden Markov Models: Klassifikation von Sequenzen
 - Andere Modelle modellieren nur einzelne Klassifikationen
 - Kann Abfolge modellieren

Anwendung von Modellen

■ Beispiele für Klassifikatoren

- Parsing: $P(\text{syntaktische Analyse} \mid \text{Satz})$ – Maxent
- Tagging: $P(\text{Wortart} \mid \text{Wort})$ – klassisch für HMMs
- Aussprache: $P(\text{Phonem} \mid \text{Buchstabe})$ – HMMs

- Siehe Folien von der letzten Vorlesung (Korpora)

■ Implementierte Systeme für alle Methoden im Internet frei verfügbar: es geht nur noch ums Anwenden auf Daten

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden
- Schritt 5: Modell evaluieren

Klassifikation: Entscheidungsalgorithmus

- Konditionaler statistischer Klassifikator

$$P(\text{Klasse } K \mid \text{Instanz } I) = \\ P(\text{Klasse } K \mid \text{Features } \underline{f}_I)$$

- Klassifiziere Instanz I mit Feature-Repräsentation f_I mit Klasse

$$K(I) = \operatorname{argmax}_K P(K \mid f_I)$$

Trainings- und Testdaten

- Typischerweise werden statistische Modelle auf einem bestimmten Korpus entwickelt (trainiert) und getestet
 - Man darf nicht auf genau denselben Daten trainieren und testen!
 - Overfitting: Modell kann nicht zwischen allgemeinen Regelmäßigkeiten (die wir wollen) und speziellen Eigenheiten des Trainingskorpus unterscheiden
 - Modell ist besser auf Trainingskorpus
 - Modell ist schlechter auf allen anderen Daten
- (Intuition: Training ist Hypothesen bilden – müssen an unabhängigen Daten verifiziert werden)

Trainings- und Testdaten

- **Korpus:**
 - Trainingsdaten (training set)
 - [Entwicklungsdaten (development set)] (optional)
 - Testdaten (test set)

 - **Auf Trainingsdaten trainieren**
 - Falls vorhanden, auf Entwicklungsdaten freie Parameter optimieren
 - **Auf Testdaten anwenden und evaluieren**
-

Cross Validation

- Aufteilung in Trainings- und Testdaten muß zufällig sein
 - Einmal Zufall ist nicht zufällig genug
 - Könnte besonders (un-)vorteilhafte Aufteilung sein
- Idee: Teile Korpus n Mal (zufällig) in Trainings- und Testdaten auf
 - Trainiere und evaluiere Modell unabhängig auf jeder Aufteilung
 - Berechne Durchschnitt und Standardabweichung

Ablauf statistischer Modellierung

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden (haben wir schon!)
- Schritt 5: Modell evaluieren

Evaluation

■ Wie gut ist eine Klassifikation?

- Accuracy (Akkuratheit):
Prozent richtiger Klassifikationen
- Error (Fehler):
Prozent Fehler

- Precision (Präzision, Genauigkeit)
- Recall (Vollständigkeit)

Detaillierter,
werden deshalb
zunehmend
benutzt

Konfusionsmatrix und Evaluationsmaße

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

- Welcher Anteil der als X klassifizierten Instanzen hat wirklich Klasse X? (Genauigkeit)
- Werte zwischen 0 und 1 (höher = besser)

Konfusionsmatrix und Evaluationsmaße II

	Echtes X	Echtes Nicht X
Als X klassifiziert	True positives	False positives
Als Nicht-X klass.	False negatives	True negatives

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

- Welcher Anteil der echten X wurde als X klassifiziert?
(Vollständigkeit)
- Werte zwischen 0 und 1 (höher = besser)

F-Score

- Nur gute Precision nutzt nichts
- Nur guter Recall nutzt auch nichts
 - Nichts als X klassifiziert: P=100%, R=0%
 - Alles als X klassifiziert: P=0%, R=100%
 - Man muß immer einen Kompromiß zwischen Precision und Recall finden
- F-Score: Kombination aus P und R:
 - Ein Maß für „Güte“ der Klassifikation
 - Werte zw. 0 und 1 (höher = besser)
 - Bevorzugt „true positives“

$$F = \frac{2PR}{P+R}$$

Nach der Evaluation...

- Schritt 1: Klassen festlegen
- Schritt 2: Informative Features identifizieren
- Schritt 3: Statistisches Modell wählen und trainieren
- Schritt 4: Modell anwenden (haben wir schon!)
- Schritt 5: Modell evaluieren



4. PP-Attachment: Das Phänomen

Friedrich sieht den Mann mit dem Fernrohr

- Friedrich hat das Fernrohr:
 - Friedrich [_{VP} sieht [_{NP} den Mann] [_{PP} mit dem Fernrohr]]

- Der Mann hat das Fernrohr:
 - Friedrich [_{VP} sieht [_{NP} den Mann [_{PP} mit dem Fernrohr]]]

PP-Attachment soll statistisch modelliert werden

Schritt 1: Klassen festlegen

- Klasse 1: PP modifiziert NP
 - Der Mann mit dem Fernrohr
- Klasse 2: PP modifiziert VP
 - Mit dem Fernrohr sehen

Schritt 2: Informative Features wählen

- Kopf der NP (n_1) Mann
- Kopf der VP (v) sieht
- Kopf der NP in der PP (n_2) Fernrohr
- Kopf der PP (Präposition) (p) mit

Schritt 3: Statistisches Modell wählen und trainieren

■ MLE?

- 95% der 4-Tupel (v, n_1, n_2, p) in den Testdaten sind nicht in den Trainingsdaten

■ Strategie 1: Maximum Entropy

- Viele einzelne binäre Features über Worte
 - $V = \text{„sieht“}$ und $N_2 = \text{„Teleskop“}$?
 - $V = \text{„sieht“}$ und $N_2 = \text{„Mantel“}$?
- Modell „erkennt“ hilfreiche Features
- Ergebnis: rund 200 Features

Strategie 2: Back-off

1. **If** $f(v, n1, p, n2) > 0$

Falls 4-Tupel gesehen

$$\hat{p}(1|v, n1, p, n2) = \frac{f(1, v, n1, p, n2)}{f(v, n1, p, n2)}$$

2. **Else if** $f(v, n1, p) + f(v, p, n2) + f(n1, p, n2) > 0$ Falls 3-Tupel gesehen

$$\hat{p}(1|v, n1, p, n2) = \frac{f(1, v, n1, p) + f(1, v, p, n2) + f(1, n1, p, n2)}{f(v, n1, p) + f(v, p, n2) + f(n1, p, n2)}$$

3. **Else if** $f(v, p) + f(n1, p) + f(p, n2) > 0$

Falls Paare gesehen

$$\hat{p}(1|v, n1, p, n2) = \frac{f(1, v, p) + f(1, n1, p) + f(1, p, n2)}{f(v, p) + f(n1, p) + f(p, n2)}$$

4. **Else if** $f(p) > 0$

Falls Präposition gesehen

$$\hat{p}(1|v, n1, p, n2) = \frac{f(1, p)}{f(p)}$$

5. **Else** $\hat{p}(1|v, n1, p, n2) = 1.0$ (default is noun attachment). Default

Schritt 4/5: Modell anwenden und evaluieren

- Daten: Computer-Handbücher von IBM
- Maxent-Approach:
 - 82% Accuracy für „normales“ Modell
 - 84% Accuracy für „getuntetes“ Modell
- Backing-off-Modell:
 - 84% Accuracy

Literatur

- Die Bibel der statistischen Sprachverarbeitung:
 - Manning & Schütze: Foundations of statistical language processing (MIT Press 1999)
- Einführung in die Verwendung statistischer Methoden:
 - Steven Abney: Statistical Methods and Linguistics
 - David Magerman: Everything you wanted to know about probability theory but were afraid to ask