

Einführung in die Computerlinguistik

Statistische Grundlagen

Laura Kallmeyer
 Heinrich-Heine-Universität Düsseldorf
 Wintersemester 2011/2012

Statistik 1 12. Dezember 2011

Kallmeyer CL-Einführung

Überblick

1. Diskrete Wahrscheinlichkeitsräume
2. Laplace-Räume
3. Bedingte Wahrscheinlichkeiten
4. Unabhängigkeit von Ereignissen
5. Die Formel von Bayes

[Carstensen et al., 2010], Abschnitt 2.4.

[Manning and Schütze, 1999], Abschnitt 2.1.

Statistik 2 12. Dezember 2011

Diskrete Wahrscheinlichkeitsräume (1)

- In vielen Bereichen der CL kommt Wahrscheinlichkeitstheorie zur Anwendung, da es oft unmöglich ist, mit rein symbolischen Ansätzen ein vollständiges Bild aller möglichen Strukturen einschließlich Präferenzen bei Ambiguitäten zu gewinnen.
- Wir haben es meist mit einer endlichen oder abzählbar unendlichen Menge von sogenannten Ergebnissen zu tun, deren Wahrscheinlichkeit irgendwie abgeschätzt werden muss.

Bsp.:

- Wahrscheinlichkeit dafür, dass VP \rightarrow VP PP verwendet wird, vorausgesetzt, man möchte eine VP generieren.
- Wahrscheinlichkeit dafür, dass *chair* eine Nomen ist.

Statistik 3 12. Dezember 2011

Kallmeyer CL-Einführung

Diskrete Wahrscheinlichkeitsräume (2)

Wir unterscheiden einzelne Ergebnisse und Ereignisse, die Mengen von Ergebnissen sind.

Bsp.: Werfen eines Würfels.

- Ergebnismenge $\{1, 2, 3, 4, 5, 6\}$
- Mögliche Ereignisse:
 - Werfen einer 1: Ereignis $\{1\}$.
Wahrscheinlichkeit $\frac{1}{6}$.
 - Werfen einer geraden Zahl: $\{2, 4, 6\}$
Wahrscheinlichkeit $\frac{1}{2}$.

Wahrscheinlichkeit eines Ereignisses ist die Wahrscheinlichkeit dafür, dass eines der Ergebnisse aus dem Ereignis eintritt.

Statistik 4 12. Dezember 2011

Diskrete Wahrscheinlichkeitsräume (3)

Beziehungen zwischen Ereignissen A und B :

- entweder A oder $B \Rightarrow A \cup B$
- sowohl A als auch $B \Rightarrow A \cap B$
- A aber nicht $B \Rightarrow A \setminus B$
- nicht $A \Rightarrow \bar{A}$

Die leere Menge beschreibt das unmögliche Ereignis und die Gesamtmenge aller Ergebnisse das sichere Ereignis.

Diskrete Wahrscheinlichkeitsräume (4)

Ein **diskreter Wahrscheinlichkeitsraum** ist ein Paar $\langle \Omega, P \rangle$, bestehend aus

1. einer nicht leeren, abzählbaren Menge Ω von **Ergebnissen** und
2. einem Wahrscheinlichkeitsmaß $P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$, so dass
 - (a) $P(A) \geq 0$ für alle $A \in \mathcal{P}(\Omega)$;
 - (b) $P(\Omega) = 1$;
 - (c) für paarweise disjunkte Mengen $A_n \in \mathcal{P}(\Omega)$, $n \in \mathbb{N}$ gilt

$$P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$$

Diskrete Wahrscheinlichkeitsräume (5)

Es ergeben sich folgende Eigenschaften für Wahrscheinlichkeitsmaße:

1. $P(\emptyset) = 0$
2. Für Ereignisse A, B mit $A \cap B = \emptyset$ gilt
 $P(A \cup B) = P(A) + P(B)$.
3. $P(A) + P(\bar{A}) = 1$ für alle $A \subseteq \Omega$ (Tertium non datur)
4. Impliziert $A \subseteq B$, das heißt $A \subseteq B$, dann sollte
 $P(B \setminus A) = P(B) - P(A)$ gelten.
5. Kein Ereignis kann eine Wahrscheinlichkeit über 1 haben.

Diskrete Wahrscheinlichkeitsräume (6)

Bsp.: $\Omega = \{\text{is-noun, has-plural-s, is-adjective, is-verb}\}$.

Frage: Kann die Funktion f mit

$$\begin{aligned} f(\text{is-noun}) &= 0.45 \\ f(\text{has-plural-s}) &= 0.2 \\ f(\text{is-adjective}) &= 0.25 \\ f(\text{is-verb}) &= 0.3 \end{aligned}$$

zu einem Wahrscheinlichkeitsmaß $f : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ ergänzt werden?

Nein, da dann $f(\Omega) = 0.45 + 0.2 + 0.25 + 0.3 = 1.2 > 1$ wäre.

Diskrete Wahrscheinlichkeitsräume (7)

Besser:

$\Omega = \{\text{is-noun-with-plural-s, is-noun-without-plural-s, is-adjective, is-verb}\}$.

$$\begin{aligned} f(\text{is-noun-with-plural-s}) &= 0.09 \\ f(\text{is-noun-without-plural-s}) &= 0.36 \\ f(\text{is-adjective}) &= 0.25 \\ f(\text{is-verb}) &= 0.3 \end{aligned}$$

Laplace-Räume (1)

Laplace-Räume sind diskrete Wahrscheinlichkeitsräume, in denen alle Ergebnisse gleich wahrscheinlich sind.

Bsp.: Würfelexperiment. $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Jedes Ergebnis hat die Wahrscheinlichkeit $\frac{1}{|\Omega|} = \frac{1}{6}$.

In Laplace-Räumen gilt also

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

Laplace-Räume (2)

Bsp.: Wahrscheinlichkeit dafür, dass in einer Gruppe von 30 Personen mindestens zwei am gleichen Tag Geburtstag haben.

Vereinfachung: Wir ignorieren Schaltjahre und saisonale Variationen.

D.h., Wahrscheinlichkeit dafür, an einem bestimmten Tag Geburtstag zu haben, ist $\frac{1}{365}$.

Wahrscheinlichkeitsraum:

- $\Omega = \{1, \dots, 365\}^{30}$, also alle Folgen von 30 Zahlen aus $\{1, \dots, 365\}$.
- $|\Omega| = 365^{30}$. Alle Folgen sind gleichwahrscheinlich.

Laplace-Räume (3)

Ziel: Wahrscheinlichkeit dafür, dass eine Folge eintritt, in der sich mindestens ein Element wiederholt.

Einfacher: Wahrscheinlichkeitsermittlung über das Komplement.

Wieviel Folgen gibt es, in denen sich kein Element wiederholt?

$$365 \times 364 \times \dots \times (365 - 29) = \frac{365!}{(365 - 30)!}$$

\Rightarrow Wahrscheinlichkeit dafür, dass zwei am gleichen Tag Geburtstag haben ist

$$1 - \frac{365!}{365^{30}(365 - 30)!} \approx 1 - 0.29 = 0.71$$

Bedingte Wahrscheinlichkeiten (1)

Bsp.:

- Wahrscheinlichkeit für eine Produktion $VP \rightarrow NP V$ für die Generierung einer VP, gegeben, dass es sich um das Verb *kisses* handelt.
- Wahrscheinlichkeit dafür, dass *chairs* ein Nomen ist, gegeben die Tatsache, dass das vorangehende Wort ein Artikel ist.
- Wahrscheinlichkeit dafür, dass *chairs* ein Nomen ist, gegeben die Tatsache, dass das nachfolgende Wort ein Artikel ist.

Bedingte Wahrscheinlichkeiten (2)

In einem diskreten Wahrscheinlichkeitsraum $\langle \Omega, P \rangle$, gegeben ein Ereignis $A \subseteq \Omega$ mit $P(A) > 0$, ist durch

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

das **durch A bedingte Wahrscheinlichkeitsmaß** $P(\cdot|A) : \Omega \rightarrow \mathbb{R}$ auf $\mathcal{P}(\Omega)$ definiert.

$\langle \mathcal{P}(\Omega), P(\cdot|A) \rangle$ ist ein diskreter Wahrscheinlichkeitsraum.

Unabhängigkeit von Ereignissen

Zwei Ereignisse A und B heißen **unabhängig**, falls $P(A \cap B) = P(A)P(B)$. Das heißt $P(A|B) = P(A)$.

Bsp. Würfelexperiment.

- Die Ereignisse, dass (A) eine gerade Zahl gewürfelt wird und (B) eine Zahl ≤ 2 sind unabhängig:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2\})}{P(\{1, 2\})} = 0.5 = P(A)$$

- Die Ereignisse A wie oben und B , dass genau die 2 gewürfelt wird, sind nicht unabhängig:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2\})}{P(\{2\})} = 1 \neq P(A)$$

Die Formel von Bayes (1)

Ziel: $P(A|B)$ berechnen auf der Grundlage von $P(B|A)$, $P(A)$ und $P(B)$.

Laut Definition gilt

$$P(A \cap B) = P(A|B)P(B) \text{ und } P(B \cap A) = P(B|A)P(A)$$

Daraus ergibt sich

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

Die Formel von Bayes (2)

Man kann das Theorem von Bayes noch verallgemeinern:

Angenommen, es gibt eine endliche oder abzählbar unendliche Folge $(A_i)_{i \in \mathbb{N}}$ von paarweise disjunkten Ereignissen mit $A_i \subseteq \Omega$ und $P(A_i) > 0$ für alle $i \in \mathbb{N}$, die eine Zerlegung von Ω bilden, dann gilt für jedes Ereignis $B \subseteq \Omega$: $(B \cap A_i)_{i \in \mathbb{N}}$ bildet eine disjunkte Zerlegung von B , und daher

$$P(B) = \sum_{i \in \mathbb{N}} P(B|A_i)P(A_i)$$

Spezialfall: Zerlegung in A und \bar{A} :

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \\ P(B) &= \frac{P(B \cap A)}{P(A)}P(A) + \frac{P(B \cap \bar{A})}{P(\bar{A})}P(\bar{A}) \\ P(B) &= P(B \cap A) + P(B \cap \bar{A}) \end{aligned}$$

Die Formel von Bayes (3)

Aus

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

und

$$P(B) = \sum_{i \in \mathbb{N}} P(B|A_i)P(A_i)$$

ergibt sich dann für die Folge $(A_i)_{i \in \mathbb{N}}$ und das Ereignis B wie oben die verallgemeinerte Formel von Bayes:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i \in \mathbb{N}} P(B|A_i)P(A_i)}$$

Die Formel von Bayes (4)

Bsp.: Angenommen, wir interessieren uns für eine relativ seltene Konstruktion, z.B. *Parasitic Gaps*, die ungefähr alle 100.000 Sätze einmal vorkommt.^a Joe Linguist hat einen Pattern-Matching Algorithmus zur Erkennung von Parasitic Gaps implementiert, der, falls ein Satz ein Parasitic Gap enthält, dies mit einer Wahrscheinlichkeit von 0.95 auch erkennt. Enthält ein Satz kein Parasitic Gap, liefert der Algorithmus mit einer Wahrscheinlichkeit von 0.005 das falsche Ergebnis, dass ein Parasitic Gap in dem Satz vorhanden ist.

Frage: Angenommen, der Test meldet ein Parasitic Gap in einem Satz. Wie wahrscheinlich ist es, dass es sich wirklich um eines handelt?

^aZ.B. *which book did she review _ without reading _?*

Die Formel von Bayes (5)

$\Omega = \{gt, \bar{g}t, g\bar{t}, \bar{g}\bar{t}\}$, wobei g für ein parasitic gap steht, t für einen positiven Test.

Sei $G = \{gt, g\bar{t}\}$ das Ereignis eines parasitic gaps, $T = \{gt, \bar{g}t\}$ das eines positiven Tests.

Wir wollen $P(G|T)$ berechnen. Wir partitionieren Ω in G und $\bar{G} = \{\bar{g}t, \bar{g}\bar{t}\}$.

$$\begin{aligned} P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\bar{G})P(\bar{G})} \\ &= \frac{0.95 \times 0.00001}{0.95 \times 0.00001 + 0.005 \times 0.99999} \\ &\approx 0.002 \end{aligned}$$

References

- [Carstensen et al., 2010] Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Langer, H., and Klabunde, R., editors (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum Akademischer Verlag. 3. überarbeitete und erweiterte Auflage.
- [Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, London, England.