Wahrscheinlichkeitstheorie und Naive Bayes

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Sommersemester 2011

12.05.2011

Elementare Wahrscheinlichkeitstheorie

Ziel

Wie wahrscheinlich ist es, dass ein Ereignis, E, eintreten wird?

Beispiel

Wie wahrscheinlich ist es, dass

Ziel

Wie wahrscheinlich ist es, dass ein Ereignis, E, eintreten wird?

Beispiel

Wie wahrscheinlich ist es, dass

- ein unbekanntes Wort ein Tippfehler ist?
- zwei Wörter x und x koreferent sind?
- Wort x auf Wort y folgt?
- Wort x im Deutschen mit y ins Englische übersetzt wird?
- etc.

Ziel

Wie wahrscheinlich ist es, dass ein Ereignis, E, eintreten wird?

Beispiel

Wie wahrscheinlich ist es, dass

- ein unbekanntes Wort ein Tippfehler ist?
- zwei Wörter x und x koreferent sind?
- Wort x auf Wort y folgt?
- Wort x im Deutschen mit y ins Englische übersetzt wird?
- etc.

$\Rightarrow$ **statistische Inferenz** basierend auf gemachten Beobachtungen

Beispiel: Münzwurf

Drei Münzen werden geworfen. Wie hoch ist die Wahrscheinlichkeit, dass zweimal 'Zahl' kommt?

Wenn die Wahrscheinlichkeitsdistribution nicht bekannt ist, kann man die Wahrscheinlichkeit annähernd durch Experimente und Beobachtungen abschätzen.

- ein dreifacher Münzwurf = ein Experiment (engl. trial)
- ein mögliches Ergebnis (engl. basic outcome) z.B. ZKK
- der Sample Space, Ω, ist die Gesamtmenge aller möglichen Ergebnisse:
  Ω={KKK,KKZ,KZZ,ZKK,ZKZ,ZZK,KZK,ZZZ}
- der Sample Space kann diskret oder kontinuierlich (=unendlich viele mögliche Outcomes) sein
- ein Ereignis (engl. event) ist eine Teilmenge von Ω, z.B. 'einmal Kopf, zweimal Zahl' (={KZZ,ZKZ,ZZK})
- der Ereignisraum, $\mathcal{F}$, ist die Potenzmenge von Ω ($=2^{\Omega}$)

Wahrscheinlichkeiten

Eine Wahrscheinlichkeitsfunktion (auch Wahrscheinlichkeitsdistribution) bildet Ereignisse auf Zahlen zwischen 0 und 1 ab:
$P : \mathcal{F} \rightarrow [0, 1]$

Dabei:
$P(\Omega) = 1$

Zurück zum Beispiel Münzwurf (1)

Drei Münzen werden geworfen. Wie hoch ist die Wahrscheinlichkeit, dass zweimal 'Zahl' kommt?

Wenn die Münzen nicht getürkt sind, ist jedes Outcome gleich wahrscheinlich (uniforme Distribution, engl. uniform distribution):
$P(KKK) = P(KKZ) = P(KZZ)... = \frac{1}{8}$

Das Ereignis, A, für das wir uns interessieren ist:
$A = \{ZZK, KZZ, ZKZ\}$

$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$

Zurück zum Beispiel Münzwurf (2)

Was, wenn wir nicht wissen, ob die Münzen getürkt sind?

Zurück zum Beispiel Münzwurf (2)

Was, wenn wir nicht wissen, ob die Münzen getürkt sind?

Wir führen eine Anzahl, T, von Experimenten (=eine Sequenz von dreifachen Münzwürfen) durch und notieren die Outcomes.

- Ereignis A={KZZ,ZKZ,ZZK}
- $P(A) \approx \frac{|A|}{|T|}$ (Maximum Likelihood Estimation, MLE)
- $\hat{P}(A)$ als Notation für die empirische Approximation von $P(A)$
- Achtung: je größer T (im Verhältnis zur Größe von Ω), desto genauer ist die Approximation von $P(A)$

Beispiel: einfacher Münzwurf

Wie hoch ist die Wahrscheinlichkeit, dass bei einem Münzwurf 'Zahl' kommt?

- $\Omega = \{K, Z\}$
- $A = Z$

Ihr werft die Münze 10-mal und bekommt 8-mal Zahl. Würdet ihr daraus folgern, dass die Münze getürkt ist und $P(A) = \frac{8}{10}$?

Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit (engl. conditional probability) gibt an, wie wahrscheinlich Ereignis A ist unter der Voraussetzung, dass Ereignis B wahr ist:

$P(A|B)$ (="A unter der Bedingung, dass B", engl. "A given B")

$\Rightarrow$ dadurch können **Abhängigkeiten zwischen Ereignissen** modelliert werden

$P(B)$ nennt man A-Priori-Wahrscheinlichkeit (engl. prior probability)
$P(A|B)$ nennt man A-Posteriori-Wahrscheinlichkeit (engl. posterior probability)

Bedingte Wahrscheinlichkeit: Beispiele

Beispiele?

Bedingte Wahrscheinlichkeit: Beispiele

Beispiele?

- $P(\text{8-mal Zahl}|\text{Münze nicht getürckt})$
- $P(Lungenkrebs|Raucher)$
- $P(w_i = Holmes|w_{i-1} = Sherlock)$

Abhängigkeiten zwischen Wahrscheinlichkeiten

Joint Probability

$P(A, B)$ ist die Verbundwahrscheinlichkeit (engl. joint probability), das A und B gemeinsam auftreten (auch geschrieben als $P(A \cap B)$).

Abhängikeit und Unabhängigkeit

Was ist die Wahrscheinlichkeit, dass zwei Ereignisse, A und B, zusammen vorkommen?

Abhängigkeiten zwischen Wahrscheinlichkeiten

Joint Probability

$P(A, B)$ ist die Verbundwahrscheinlichkeit (engl. joint probability), das A und B gemeinsam auftreten (auch geschrieben als $P(A \cap B)$).

Abhängikeit und Unabhängigkeit

Was ist die Wahrscheinlichkeit, dass zwei Ereignisse, A und B, zusammen vorkommen?

Hängt davon ab, welche **Abhängigkeiten** zwischen A und B bestehen:

- wenn A von B **abhängig** ist: $P(A, B) = P(B) \times P(A|B)$
- wenn A und B **unabhängig** sind: $P(A, B) = P(A) \times P(B)$ (da dann gilt: $P(A|B) = P(A)$)

Abhängigkeiten zwischen Wahrscheinlichkeiten

Joint Probability

$P(A, B)$ ist die Verbundwahrscheinlichkeit (engl. joint probability), das A und B gemeinsam auftreten (auch geschrieben als $P(A \cap B)$).

Abhängikeit und Unabhängigkeit

Was ist die Wahrscheinlichkeit, dass zwei Ereignisse, A und B, zusammen vorkommen?

Hängt davon ab, welche **Abhängigkeiten** zwischen A und B bestehen:

- wenn A von B **abhängig** ist: $P(A, B) = P(B) \times P(A|B)$
- wenn A und B **unabhängig** sind: $P(A, B) = P(A) \times P(B)$ (da dann gilt: $P(A|B) = P(A)$)

Konditionale Unabhängigkeit

A und B sind konditional unabhängig voneinander gegeben C genau dann wenn: $P(A|B, C) = P(A|C)$

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (1)

A:Münze-1 zeigt Zahl
B:Münze-2 zeigt Zahl

$P(A, B)=$

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (1)

A:Münze-1 zeigt Zahl
B:Münze-2 zeigt Zahl

$P(A, B)$=$0.5 \times 0.5 = 0.25$

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (1)

A:Münze-1 zeigt Zahl
B:Münze-2 zeigt Zahl

$P(A, B)$=$0.5 \times 0.5 = 0.25$

Oder anders hergeleitet:

- wir definieren ein neues Ereignis C ($\equiv A \cap B$): $\{ZZK, ZZZ\}$
- $P(C) = \frac{2}{8} = \frac{1}{4} = 0.25$

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (2)

A:Wort$_i$=Sherlock
B:Wort$_{i+1}$=Holmes

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (2)

A:Wort$_i$=Sherlock
B:Wort$_{i+1}$=Holmes

Wir zählen Worthäufigkeiten in einer Sherlock Holmes Geschichte und benutzen MLE um die folgende Wahrscheinlichkeiten zu bekommen:

$P(B) = 0.1$
$P(A|B) = 0.5$

Dann ist:

$P(A, B) =$

Beispiel: Abhängigkeiten zwischen Wahrscheinlichkeiten (2)

A:Wort$_i$=Sherlock
B:Wort$_{i+1}$=Holmes

Wir zählen Worthäufigkeiten in einer Sherlock Holmes Geschichte und benutzen MLE um die folgende Wahrscheinlichkeiten zu bekommen:

$P(B) = 0.1$
$P(A|B) = 0.5$

Dann ist:

$P(A, B) = 0.1 \times 0.5 = 0.05$

Regeln (1)

Bayes' Theorem

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

Regeln (1)

Bayes' Theorem

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

Herleitung:

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

Regeln (1)

Bayes' Theorem

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

Herleitung:

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

Außerdem gilt:

$$\begin{aligned} P(A) &= P(A, B) + P(A, \bar{B}) \qquad \text{(Additivität)} \\ &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \end{aligned}$$

Regeln (2)

Multiplikationsregel

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

Regeln (3)

Kettenregel

$$P(A_1 \cap A_2 \cap .... \cap A_n) =$$
$$P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n| \cap_{i=1}^{n-1} A_i)$$

Naive Bayes Klassifikator

Beispiel: Wortdisambiguierung (1)

Bank=Kreditinstitut oder Bank=Sitzgelegenheit?

In den letzten dreißig Jahren fanden an den weltweiten Finanzmärkten starke Positionsverschiebungen unter den **Banken** statt. Trotz einer stark gestiegenen Rendite der meisten deutschen Institute und teils erheblich gesteigerten Kernkapitals und ebenso gewachsenen Bilanzsumme fielen deutsche Institute weiter ab.

Beispiel: Wortdisambiguierung (2)

Welche Bedeutung ist am wahrscheinlichsten?

Annahme: Bedeutung hängt vom Kontext ab.

- $s_1 ... s_k ... s_K$: die Bedeutungen des Zielworts w
- $c_1 ... c_i ... c_I$: die Kontexte, in denen w vorkommt
- $\underset{s_k}{\text{argmax}}\, P(s_k|c)$: die wahrscheinlichste Bedeutung
  $\Rightarrow$ Maximum a posteriori (MAP) Wahrscheinlichkeit

Modellierung des Kontexts

- die (Inhalts)Wörter in einem Kontextfenster um das Zielwort
- z.B.
  c=(dreißig,Jahre,finden,weltweit,Finanzmarkt,Positionsverschiebung,..
- v_j: ein Kontextwort

Beispiel: Wortdisambiguierung (3)

Berechnung der Wahrscheinlichkeiten

- Wenn wir unendlich viele disambiguierte Beispiele hätten, könnten wir $P(s_k|c)$ mit MLE direkt aus den Daten berechnen.
- **Aber:** um $P(s_k|c)$ einigermaßen akkurat abschätzen zu können, müßte jede Bedeutung sehr häufig mit jedem einzelnen Kontext vorkommen.
  $\Rightarrow$ selbst wenn wir alle Texte der Welt disambiguiert hätten, würden die Daten lange noch nicht ausreichen!
  $\Rightarrow$ wir haben ein Data Sparseness Problem
- je größer der in Betracht gezogene Kontext (=der Attribut-Wert-Vector), desto größer muß das Trainingsset sein

Beispiel: Wortdisambiguierung (4)

Der Naive Bayes Trick

$$P(s_k|c) = \frac{P(c|s_k)}{P(c)} P(s_k)$$

- $P(c)$ ist konstant für alle s_k
  $\Rightarrow P(s_k|c) \propto P(c|s_k)P(s_k)$
- $P(s_k)$ kann mit MLE aus den Trainingsdaten approximiert werden
- $P(c|s_k)$ ist immer noch sparse

Beispiel: Wortdisambiguierung (4)

Der Naive Bayes Trick

$$P(s_k|c) = \frac{P(c|s_k)}{P(c)} P(s_k)$$

- $P(c)$ ist konstant für alle s_k
  $\Rightarrow P(s_k|c) \propto P(c|s_k)P(s_k)$
- $P(s_k)$ kann mit MLE aus den Trainingsdaten approximiert werden
- $P(c|s_k)$ ist immer noch sparse

Naive Bayes Annahme

die einzelnen Kontextwörter, v_j, sind **konditional unabhängig** gegeben s_k:

$$P(c|s_k) = P(\{v_j|v_j \ in \ c\}|s_k) = \Pi_{v_j \ in \ c} P(v_j|s_k)$$

$\Rightarrow P(v_j|s_k)$ kann mit deutlich weniger Daten relativ genau berechnet werden.

Warum 'Naive' Bayes?

Warum 'Naive' Bayes?

- In natürlicher Sprache gibt es kaum (konditionale) Unabhängigkeiten.
- $P(\{v_j | v_j \ in \ c\} | s_k) = \Pi_{v_j \ in \ c} P(v_j | s_k)$ ist eindeutig falsch!
- Trotzdem funktioniert Naive Bayes erstaunlich gut!

Induktiver Bias

Was ist der induktive Bias von Naive Bayes?

Induktiver Bias

Was ist der induktive Bias von Naive Bayes?

konditional Unabhängigkeit der Attribute

Smoothing (1)

Wenn die Daten trotzdem nicht reichen

Wenn $\hat{P}(A) = 0$ bedeutet das, dass A nicht vorkommen kann?
Besser ungesehenen Ereignissen eine kleine, artifizielle Wahrscheinlichkeit zuzuweisen.
$\Rightarrow$ Smoothing

Laplace Smoothing

Laplace

$$P_{Lap}(X) = \frac{|X|+1}{|N|+|\Omega|}$$

Problem

- Es gibt meist sehr, sehr viele ungesehene Ereignis (cf. Zipf).
- Laplace Smoothing vergibt zu viel der gesamten Wahrscheinlichkeitsmasse an diese ungesehenen Ereignisse.
- Beispiel (Manning & Schütze, S. 2003): 99.97% an ungesehene Bigrams in einem Bigram-Modell

Lidstone Smoothing

Lidstone

$$P_{Lid}(X) = \frac{|X|+\lambda}{|N|+\lambda|\Omega|}$$

für $\lambda << 1$

Vorteile

- λ kann flexibel gesetzt werden
- je kleiner λ desto weniger Wahrscheinlichkeitsmasse wir an ungesehene Ereignisse verteilt

Zusammenfassung

Was ihr gelernt haben solltet

Wahrscheinlichkeitstheorie

- Terminologie (Experiment, Ereignis, Outcome, Sample Space etc.)
- Wie berechnet man Wahrscheinlichkeiten (z.B. MLE)
- Abhängigkeiten zwischen Wahrscheinlichkeiten (a priori, a posteriori, Joint, (konditionale) Unabhängigkeit, etc.)
- Bayes Theorem, Kettenregel, etc.

Naive Bayes Klassifikator

- Wie funktioniert er?
- Data Sparseness (und wie NB damit umgeht)
- Smoothing