Wahrscheinlichkeitstheorie 2

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Sommersemester 2011

19.05.2011

Wiederholung (1): Wie funktioniert Naive Bayes?

Wiederholung (1): Wie funktioniert Naive Bayes?

Ziel

$\underset{c_i}{\text{argmax}}\ P(c_i|a_1..a_n)$ für Ausgabeklasse c_i und Attribute $a_1...a_n$

Modellierung

$P(c|a_1...a_n) =$

Wiederholung (1): Wie funktioniert Naive Bayes?

Ziel

$\underset{c_i}{\text{argmax}}\ P(c_i|a_1..a_n)$ für Ausgabeklasse c_i und Attribute $a_1...a_n$

Modellierung

$P(c|a_1...a_n) = \frac{P(a_1...a_n|c)}{P(a_1...a_n)} P(c)$ (Bayes Theorem)

Wiederholung (1): Wie funktioniert Naive Bayes?

Ziel

$\underset{c_i}{\text{argmax}}\ P(c_i|a_1..a_n)$ für Ausgabeklasse c_i und Attribute $a_1...a_n$

Modellierung

$$P(c|a_1...a_n) = \frac{P(a_1...a_n|c)}{P(a_1...a_n)} P(c) \qquad \text{(Bayes Theorem)}$$

$$P(c|a_1...a_n) \propto P(a_1...a_n|c)P(c)$$

Wiederholung (1): Wie funktioniert Naive Bayes?

Ziel

$\underset{c_i}{\text{argmax}}\ P(c_i|a_1..a_n)$ für Ausgabeklasse c_i und Attribute $a_1...a_n$

Modellierung

$P(c|a_1...a_n) = \frac{P(a_1...a_n|c)}{P(a_1...a_n)} P(c)$ (Bayes Theorem)

$P(c|a_1...a_n) \propto P(a_1...a_n|c)P(c)$

Korrekt (Kettenregel):

$P(a_1...a_n|c) = P(a_1|c)P(a_2|a_1, c)P(a_3|a_1, a_2, c)...P(a_n|a_1, a_2....a_{n-1}, c)$

Wiederholung (1): Wie funktioniert Naive Bayes?

Ziel

$\underset{c_i}{\text{argmax}}\, P(c_i|a_1..a_n)$ für Ausgabeklasse c_i und Attribute $a_1...a_n$

Modellierung

$P(c|a_1...a_n) = \frac{P(a_1...a_n|c)}{P(a_1...a_n)} P(c)$ (Bayes Theorem)

$P(c|a_1...a_n) \propto P(a_1...a_n|c)P(c)$

Korrekt (Kettenregel):

$P(a_1...a_n|c) = P(a_1|c)P(a_2|a_1, c)P(a_3|a_1, a_2, c)...P(a_n|a_1, a_2....a_{n-1}, c)$

Naive Bayes:

$P(a_1...a_n|c) \approx P(a_1|c)P(a_2|c)P(a_3|c)....P(a_n|c)$

(Annahme: **konditionale Unabhängigkeit der Attribute gegeben** c)

Wiederholung (2): Wahrscheinlichkeitstheorie

- Maximum Likelihood Estimation
- konditionale und gemeinsame (joint) Wahrscheinlichkeiten
- a priori, a posteriori
- Regeln: Bayes Theorem, Kettenregel etc.
- Smoothing

Heute

- Zufallsvariablen
- Parametrische vs. Nicht-Parametrische Methoden
- Wahrscheinlichkeitsdistributionen

Zufallsvariablen

Wir gehen davon aus, dass alle Ergebnisse (basic outcomes) des Sample Space Ω eine bestimmte Wahrscheinlichkeit haben und $P(\Omega) = 1$.
Z.B. $P(Kopf) = 0.5$ und $P(Zahl) = 0.5$.

Wir stellen uns vor, dass es einen abstrakten Prozess gibt, der Ergebnisse auf Wahrscheinlichkeiten abbildet (stochastischer Prozess).
Ferner gibt es eine Funktion, X, die Ergebnisse auf reale Werte abbildet:
$X : \Omega \rightarrow \mathbb{R}$
Diese Funktion, X, nennen wir Zufallsvariable (engl. random variable).

Z.B.: $X(\omega) = \left\{ \begin{array}{l} 0, \ wenn\ \omega = Kopf, \\ 1, \ wenn\ \omega = Zahl. \end{array} \right\}$

Wahrscheinlichkeitsfunktion

Die Wahrscheinlichkeitsfunktion (engl. probability (mass) function, pmf) einer Zufallsvariable X gibt die Wahrscheinlichkeit an, dass X einen bestimmten Wert hat:

pmf: $p(x) = p(X = x) = P(A_x)$ wobei $A_x = \{\omega \in \Omega : X(\omega) = x\}$

$\Sigma_i p(x_i) = P(\Omega) = 1$

Die Wahrscheinlichkeitsverteilung (engl. probability distribution) gibt an, wie sich die Wahrscheinlichkeiten auf die möglichen Ergebnisse verteilen.

Beispiel: Anzahl von 'Zahl' bei Wurf von drei Münzen

Sample Space:

$\Omega = \{KKK, KKZ, KZZ, ZZZ, ZZK, ZKK, ZKZ, KZK\}$

Wahrscheinlichkeitsfunktion:

$P(X = x)$ wobei $x \in \{0, 1, 2, 3\}$, z.B. $P(X = 0) = 0.125$

Beispiel: Anzahl von 'Zahl' bei Wurf von drei Münzen

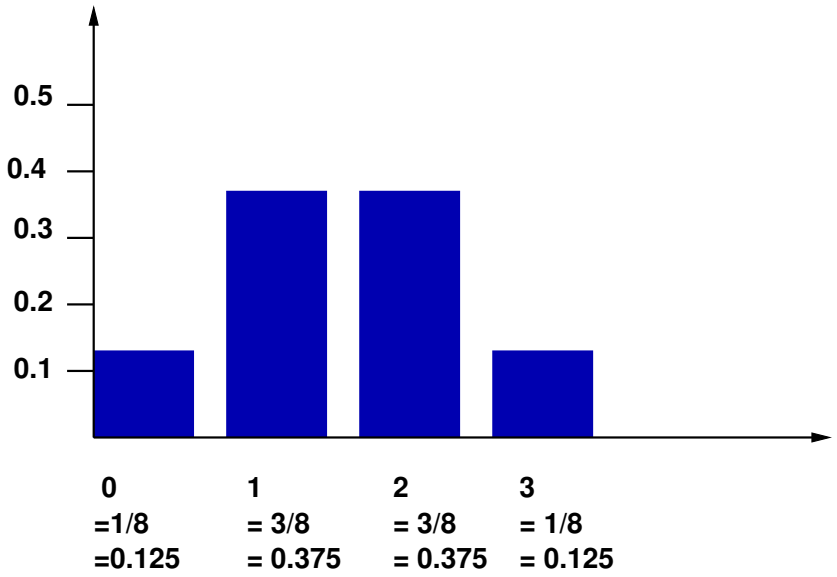
Sample Space:

$\Omega = \{KKK, KKZ, KZZ, ZZZ, ZZK, ZKK, ZKZ, KZK\}$

Wahrscheinlichkeitsfunktion:

$P(X = x)$ wobei $x \in \{0, 1, 2, 3\}$, z.B. $P(X = 0) = 0.125$

Wahrscheinlichkeitsverteilung:



Erwartungswert

Erwartungswert (auch Mittelwert, engl. mean oder expectated value) einer Zufallsvariablen, X:

$E(X) = \Sigma_x x p(x)$

$E(X)$ wird auch als μ geschrieben, wenn von einer bestimmten Wahrscheinlichkeitsfunktion die Rede ist.

Beispiel: Anzahl 'Zahl' bei 3-fachen Münzwurf

$E(X) = 0 \times 0,125 + 1 \times 0,375 + 2 \times 0,375 + 3 \times 0,125 = 1,5$

(Alternativ: $E(X) = \frac{0+1+1+1+2+2+2+3}{8} = 1,5$)

Varianz

Die Varianz (engl. variance) ist ein Maß dafür, wie weit eine Zufallsvariable im Allgemeinen von ihren Erwartungswert abweicht.

$$V(X) = E((X - E(X))^2) = E(X^2) - E^2(X) = \Sigma_x p(x)(x - E(X))^2$$

$V(X)$ wird auch $Var(X)$ oder σ^2 geschrieben.

σ ist die Standardabweichung (engl. standard deviation):
$\sigma_x = \sqrt{Var(X)}$

Beispiel: Anzahl 'Zahl' bei 3-fachen Münzwurf
$V(X) = 0,125(0 - 1,5)^2 + 0,375(1 - 1,5)^2 + 0,375(2 - 1,5)^2 + 0,125(3 - 1,5)^2 = 0,75$

Zentrale Momente

Erwartungswert und Varianz sind zwei zentralen Momente (engl. central moments) einer Zufallsvariablen X.

Wahrscheinlichkeitsverteilungen können durch Angabe der Art der Verteilung (mehr später) und der zentralen Momente, z.B. Erwartungswert und Varianz beschrieben werden.

Parametrische vs. Nicht-Parametrische Methoden

Bestimmung von Wahrscheinlichkeiten:

- empirisch durch auszählen (Maximum Likelihood Estimation) oder

- Man nimmt an, dass die gesuchte Wahrscheinlichkeitsverteilung von einem bestimmten Typ ist. Dann müssen nur noch die Parameter der Verteilung empirisch bestimmt werden (z.B. Erwartungswert und Varianz). Dazu werden weniger Daten benötigt.

Parametrische vs. Nicht-Parametrische Methoden

Bestimmung von Wahrscheinlichkeiten:

- empirisch durch auszählen (Maximum Likelihood Estimation) oder
  nicht-parametrische Methode
- Man nimmt an, dass die gesuchte Wahrscheinlichkeitsverteilung von einem bestimmten Typ ist. Dann müssen nur noch die Parameter der Verteilung empirisch bestimmt werden (z.B. Erwartungswert und Varianz). Dazu werden weniger Daten benötigt.
  parametrische Methode

Arten von Wahrscheinlichkeitsverteilungen

Diskrete Verteilungen

z.B.

- Bernoulli-Verteilung
- Binomialverteilung
- Multinomialverteilung
- Zipf-Verteilung

Kontinuierliche Verteilungen

z.B.

- Normalverteilung (auch Gauß-Verteilung)

Bernoulli-Verteilung

Eine Bernoulli-Verteilung liegt vor wenn nur ein Experiment durchgeführt wird und es nur zwei mögliche Ergebnisse gibt (0 oder 1). Ein solches Experiment wird auch Bernoulli-Versuch (engl. Bernoulli trial) genannt.

$P(X = 1) = p$ und $P(X = 0) = q = 1 - p$

$E(X) = p$ und $V(X) = p(1 - p) = pq$

Beispiele

Bernoulli-Verteilung

Eine Bernoulli-Verteilung liegt vor wenn nur ein Experiment durchgeführt wird und es nur zwei mögliche Ergebnisse gibt (0 oder 1). Ein solches Experiment wird auch Bernoulli-Versuch (engl. Bernoulli trial) genannt.

$P(X = 1) = p$ und $P(X = 0) = q = 1 - p$

$E(X) = p$ und $V(X) = p(1 - p) = pq$

Beispiele

- ein Münzwurf ('Kopf' oder 'Zahl')
- Lotteriegewinn ('ja' oder 'nein')
- Bestehen einer Klausur ('ja' oder 'nein')

Binomialverteilung (1)

Die Binomialverteilung (engl. binomial distribution) beschreibt wie of ein bestimmtes Ergebnis in einer Reihe von **unabhängigen Experimenten** vorkommt, wobei jedes Experiment ein Bernoulli-Experiment ist. Die Bernoulli-Verteilung ist daher ein Spezialfall der Binomialverteilung.

Für eine Anzahl von Erfolgen r (d.h. $X = 1$) und eine Anzahl von Experimenten n ist die Wahrscheinlichkeit eines Erfolgs in einem Experiment gegeben durch p:

$$b(r; n, p) = \binom{n}{r} p^r (1-p)^{n-r}$$

wobei $\binom{n}{r} = \frac{n!}{(n-r)!r!} \quad 0 \leq r \leq n$

$\binom{n}{r}$ (Binomialkoeffizient, gesprochen "n über r") beschreibt die Anzahl der Möglichkeiten wie man r Objekte aus n auswählen kann, ohne die Reihenfolge zu beachten.

$E(X) = np$ und $V(X) = np(1-p)$

Binomialverteilung (2)

Für NLP ist die Binomialverteilung immer eine Approximation (keine echte Unabhängigkeit), wird aber trotzdem oft verwendet.

Beispiele

- mehrere Münzwürfe (Anzahl von 'Zahl')
- Anzahl der Passivkonstruktionen in einem Korpus

Multinomialverteilung

Die Multinomialverteilung ist eine Generalisierung der Binomialverteilung, in der k Ergebnisse möglich sind (mit $k > 2$).

$E(X_i) = np_i$ und $V(X_i) = np_i(1 - p_i)$

wobei n die Anzahl der Experimente ist

Beispiele

Multinomialverteilung

Die Multinomialverteilung ist eine Generalisierung der Binomialverteilung, in der k Ergebnisse möglich sind (mit $k > 2$).

$E(X_i) = np_i$ und $V(X_i) = np_i(1 - p_i)$

wobei n die Anzahl der Experimente ist

Beispiele

- n-gram-Modell

Zipf-Verteilung (1)

Das Zipfsche Gesetz (engl. Zipf's law) beschreibt den Zusammenhang zwischen der Häufigkeit f eines Elements, e und seines Rangs r in einer nach Häufigkeit geordneten Liste:

$f \propto \frac{1}{r}$ (oder auch $p(e) \propto \frac{1}{r}$)

Z.B. sollte das 50-häufigste Wort dreimal so häufig sein wie das 150-häufigste Wort.

Dies ist eine Ausprägung von Zipfs Prinzip des geringsten Aufwands.

Zipf-Verteilung (2)

Weitere Beispiele des Prinzip des geringsten Aufwands

- je häufiger ein Wort desto mehr Bedeutungen hat es
- je häufiger ein Wort desto kürzer ist es
- Inhaltswörter gruppieren sich zusammen

$\Rightarrow$ Zipf-Verteilungen spielen eine große Rolle für linguistischen Phänomene.

$\Rightarrow$ Dies hat Auswirkungen auf die statistische Modellierung (viele seltene Ereignisse)

Normalverteilung (1)

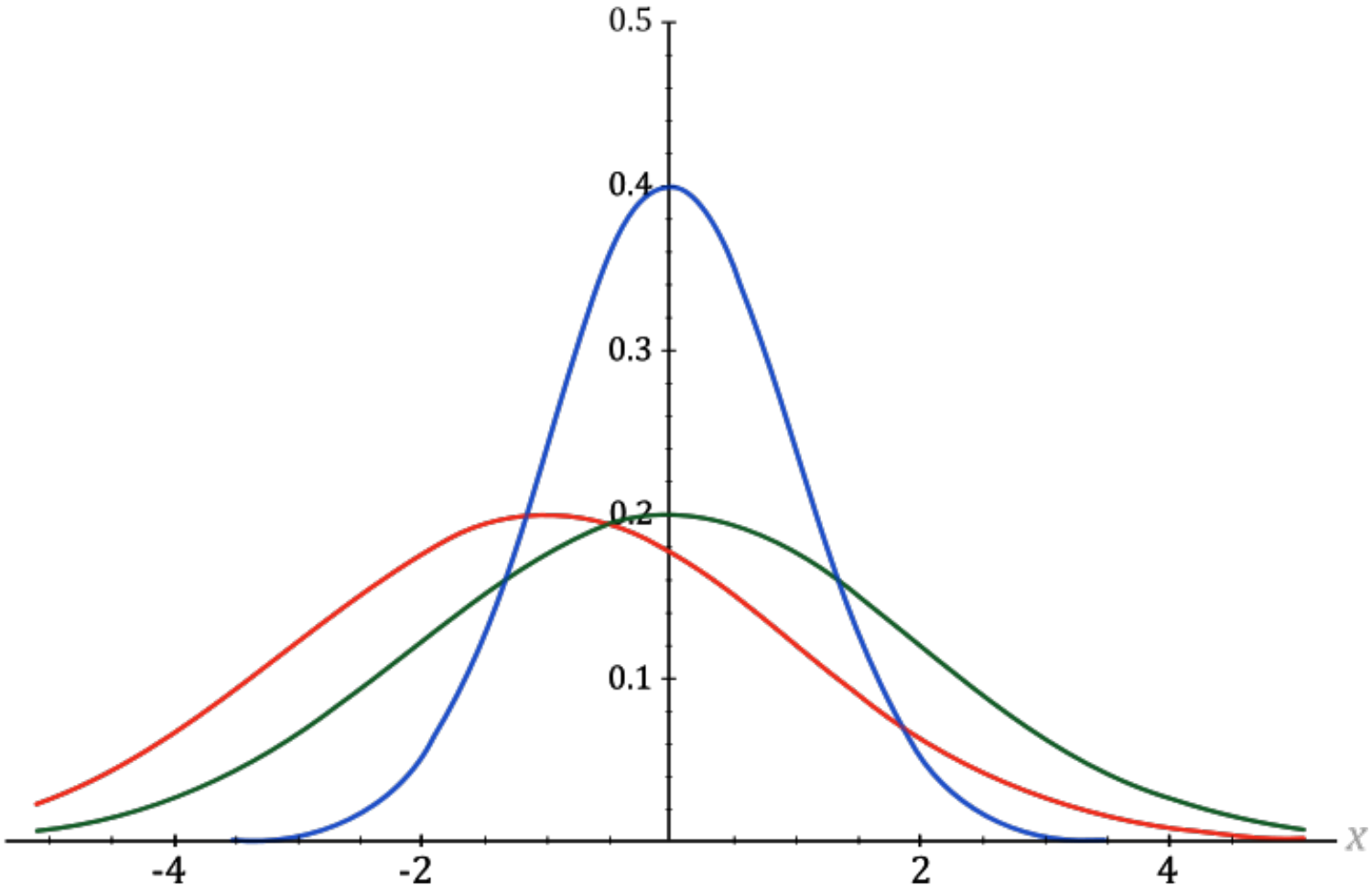
Die Normalverteilung (engl. normal distribution), auch Gauß-Verteilung (engl. Gaussian distribution) genannt, ist die bekannteste kontinuierliche Wahrscheinlichkeitsverteilung.

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

Wenn $\mu = 0$ und $\sigma^2 = 1$ nennt man die Verteilung Standardnormalverteilung (engl. standard normal distribution).

Normalverteilung (2)

Beispiele für Normalverteilungen*



$N(0,1)$ (blau); $N(0,2)$ (grün); $N(-1,2)$ (rot)

*(Quelle: http://de.wikipedia.org/wiki/Normalverteilung)

Normalverteilung (3)

Beispiele

Normalverteilung (3)

Beispiele

viele Phänomene sind (annähernd) normal-verteilt, z.B.:

- Körpergröße
- Schulnoten
- Anzahl der Eier, die Hühner legen
- durchschnittliche Lebensdauer von Glühbirnen

. . . linguistische Phänomene aber typischerweise nicht!

Zusammenfassung

Was ihr gelernt haben solltet:

- Zufallsvariable
- parametrische vs. nicht-parametrische Methoden
- verschiedene Wahrscheinlichkeitsfunktionen
- das Zipfsche-Gesetz