Stochastische Lernalgorithmen

Gerhard Jäger

14. Mai 2003

Das Maximum-Entropy-Prinzip

Der Entropiebegriff

- *Entropie:* „Chaos“, Unordung, Nicht-Vorhersagbarkeit, ...
- Begriff kommt ursprünglich aus der Physik:
  - Entropie wird nur geringer wenn Energie zugeführt wird
- Maß der *Ungewissheit*:
  - geringe Entropie bedeutet geringe Unsicherheit und umgekehrt
  - hohe Entropie bedeutet hohe „Überraschung“
  - je überraschender das Ergebnis eines Experimentes ist, desto überraschter sind wir

Die Formel

- Sei $p(X)$ die Verteilung der Zufallsvariablen X
- Entropie von X („$H(X)$“): Erwartungswert des negativen Logarithmus der Wahrscheinlichkeit

$$H(X) \quad = \quad -\sum_x p(x) \log_2 p(x)$$

- Einheit: *Bit*
- Konventionen:
  - Wenn nicht explizit anders angegeben, benutzen wir den natürlichen Logarithmus
  - $0 \log 0 = 0$
- Notationsvarianten: $H(X) = H(p) = H_X(p) = H(p_X)$

Beispiele

- Bernoulli-Verteilung, basiert auf fairer Münze
  - $p(0) = p(1) = 0.5$
  - $\mathbf{H(p)} = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = -2 \cdot (0.5 \cdot -1) = \mathbf{1}$
- Augenzahl beim fairen Würfel
  - $\mathbf{H(p)} = -(6 \cdot (\frac{1}{6} \log_2 \frac{1}{6})) = \log_2 6 \approx 2.584962$
- Unfaire Münze:
  - $p(1) = 0.2 \ldots H(p) = 0.722; p(1) = 0.01 \ldots h(p) = 0.081$

Extreme

- Wenn für ein $x : p(x) = 1$, dann $H(p) = 0$
- Keine Überraschung $\Rightarrow$ keine Information
- keine oberste Schranke, aber $H(X) \leq \log_2 |\Omega|$
  - Wenn alle Elementarereignisse gleich wahrscheinlich sind, ist die Vorhersagbarkeit am geringsten

Kodierungs-Interpretation der Entropie

- minimale durchschnittliche Zahl von Bits, die nötig sind, um eine Botschaft (Zeichenkette, Sequenz, Serie, ...) zu kodieren (wobei jedes Element Resultat eines Zufallsprozesses mit Verteilung p ist) $:= H(p)$
- **Beispiel:**
  - 4 Symbole

    $$A\ C\ G\ T$$

  - Wahrscheinlichkeiten

    $$P(A) = 0.5,\ P(C) = 0.25,\ P(G) = 0.125,\ P(T) = 0.125$$

  - Negative binäre Logarithmen:

    A: 1 Bit, C: 2 Bit, G: 3 Bit, T: 3 Bit

- Entropie

$$\begin{aligned} H &= \sum p(x) - \log_2 p(x) \\ &= 0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.125 \cdot 3 \\ &= 0.5 + 0.5 + 0.375 + 0.375 \\ &= 1.75 \end{aligned}$$

- Mögliche Kodierung in binärem Alphabet:

A: 00; C: 01; G: 10; T: 11

- gewichtete durchschnittliche Kode-Länge: 2 ($> H$)
- Besserer Kode:

A: 1; C: 01; G: 000; T: 001

- gewichtete durchschnittliche Kode-Länge:

$$0.5 \cdot 1 + 0.25 \cdot 2 + 0.125 \cdot 3 + 0.25 \cdot 3 = 1.75$$

- Optimaler Kode: Kodelänge für Symbol $S = -log_2(P(S))$

Perplexität: Motivation

- Gleichverteilung:
  - bei 2 gleichwahrscheinlichen Ergebnissen: $H(p) = 1$
  - bei 32 gleichwahrscheinlichen Ergebnissen: $H(p) = 5$
  - bei 4.3 Milliarden gleichwahrscheinlichen Ergebnissen: $H(p) \approx 32$
- Wenn Ergebnissen nicht gleichwahrscheinlich sind:
  - 32 Elementarereignisse, 2 davon gleichwahrscheinlich, Rest unmöglich:
    - $H(p) = 1$
  - Gibt es eine Möglichkeit, den Grad der Unsicherheit/Informationsgehalt für Zufallsvariable zu vergleichen, wenn die zugrundeliegenden Ereignisräume unterschiedlich groß sind?

Perplexität

- Perplexität
  - $G(p) = 2^{H(p)}$

- D.h. wir sind wieder bei 32 (für 32 gleichwahrscheinliche Ergebnisse), 2 für Münzwurf usw.
- intuitiv leichter vorstellbar:
  - $G(p) = n$ bedeutet, dass p so gut vorhersagbar ist wie ein Experiment mit n gleichwahrscheinlichen Ausgängen
- Je mehr eine Verteilung von der Gleichverteilung abweicht, desto geringer sind Entropie und Perplexität

Eigenschaften der Entropie

- Entropie ist nicht-negative
  - $H(x) \geq 0$
  - Beweis:
    - $\log_2 x \leq 0$ für $0 < x \leq 1$
    - $p(x)$ ist nicht-negativ, also ist $p(x) \log_2 p(x)$ nicht-positiv
    - Summe nicht-positiver Werte ist nicht-positiv
    - Negation eines nicht-positiven Wertes ist nicht-negativ

Relative Entropie (Kullback-Leibler-Abstand)

- Effiziente Kodierung eines stocahstischen Prozesses setzt Kenntnis seiner Wahrscheinlichkeitsverteilung voraus
- Annahme falscher Verteilung führt zu Verschwendung von Bits
- *Relative Entropie* zweier Verteilungen p und q misst die minimale durchschnittliche Anzahl von verschwendeten Bits, wenn man einen Prozess mit Verteilung p auf der Basis von q kodiert

  Definition 1 (Relative Entropie) *Die* relative Entropie *(oder der* Kullback-Leibler-Abstand*) zwischen zwei Wahrscheinlichkeitsfunktionen p und q ist definiert als*

$$\begin{aligned} D(p\|q) &= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \\ &= E_p \log_2 \frac{p(x)}{q(x)} \end{aligned}$$

- Relative Entropie ist immer nicht-negativ
- $D(p\|q) = 0$ gdw. $p = q$

- Kann als „Abstand“ zwischen zwei Verteilungen aufgefasst werden
- Aber Vorsicht: Die relative Entropie (im Unterschied zum Abstand im geometrischen Sinn)
  - ist nicht immer symmetrisch
  - erfüllt die Dreiecksungleichung nicht

Maximum Entropy

- Gegeben sei eine Menge von Trainings-Ereignissen Y, die (via relative Häufigkeiten) eine Referenz-Verteilung q definieren
- Weiterhin gegeben: Menge von **Merkmalen** (*features*) f_i – Funktionen von Elementarereignissen in die reelen Zahlen.[1]
- Gesucht ist eine Wahrscheinlichkeitsverteilung p, die folgende Bedingungen (*constraints*) erfüllt

$$\sum_x p(x) f_i(x) = \sum_x q(x) f_i(x) \quad (1)$$

$$\sum_x p(x) = 1 \quad (2)$$

- Darüber hinaus sollen keine speziellen Annahmen über p gemacht werden.
- Im Allgemeinen gibt es unendlich viele Lösungen.
- Optimal ist die Lösung, deren Entropie maximal ist.
- Unter sehr allgemeinen Bedingungen existiert eine eindeutige Lösung.

Loglineare Modelle

- **Hauptproblem:** Sei $\mathcal{C}$ die Menge der Wahrscheinlichkeitsverteilungen, die die o.g. Constraints erfüllen. Gesucht is

$$p_* = \arg\max_{p \in \mathcal{C}} H(p)$$

- **Hilfsfunktion** (“Lagrangian”): Für jedes Merkmal f_i wird ein **Gewicht** λ_i eingeführt (sog. Lagrange-Multiplikator). Über die Gewichte ist zunächst nichts bekannt, außer dass es reelle Zahlen sind.

$$\Lambda(p, \lambda) = H(p) + \sum_i \lambda_i (\sum_x p(x) f_i(x) - \sum_x q(x) f_i(x))$$

[1] Nicht zu verwechseln mit linguistischen Merkmalen wie Genus, Kasus, Kongruenz usw. Letztere heißen der Eindeutigkeit halber in Zukunft “Attribute”.

- Suche Maximum von Λ bezüglich p (wobei $\mathcal{P}$ die unbeschränkte Menge aller Wahrscheinlichkeitsverteilungen ist):

$$p_\lambda = \arg\max_{p \in \mathcal{P}} \Lambda(p, \lambda)$$

- **Dualfunktion:**

$$\Psi(\lambda) = \Lambda(p_\lambda, \lambda)$$

- p_λ und $\Psi(\lambda)$ können mit Hilfe der Differentialrechnung bestimmt werden:

$$p_\lambda(x) = \frac{1}{Z_\lambda} \exp(\sum_i \lambda_i f_i(x)) \quad (3)$$

$$Z_\lambda = \sum_x \exp(\sum_i \lambda_i f_i(x)) \quad (4)$$

$$\Psi(\lambda) = -\log Z_\lambda + \sum_x q(x) \sum_i \lambda_i f_i(x) \quad (5)$$

- **Dualproblem**: Finde die Gewichte, die die Dualfunktion maximieren:

$$\lambda^* = \arg\max_\lambda \Psi(\lambda) \quad (6)$$

- Lösung des Dualproblems liefert Lösung des Hauptproblems:

$$\boxed{\boxed{p_* = \Psi(\lambda^*)}}$$

- Dualproblem nicht einfach analytisch oder numerisch lösbar
- diverse iterative Optimierungalgorithmen (siehe nächste Woche, Sitzung am 11. Juni)
- Warum "loglinear":
  - Wahrscheinlichkeit nach p^* ist exponentiell in den Merkmalswerten
  - Anders gesagt:

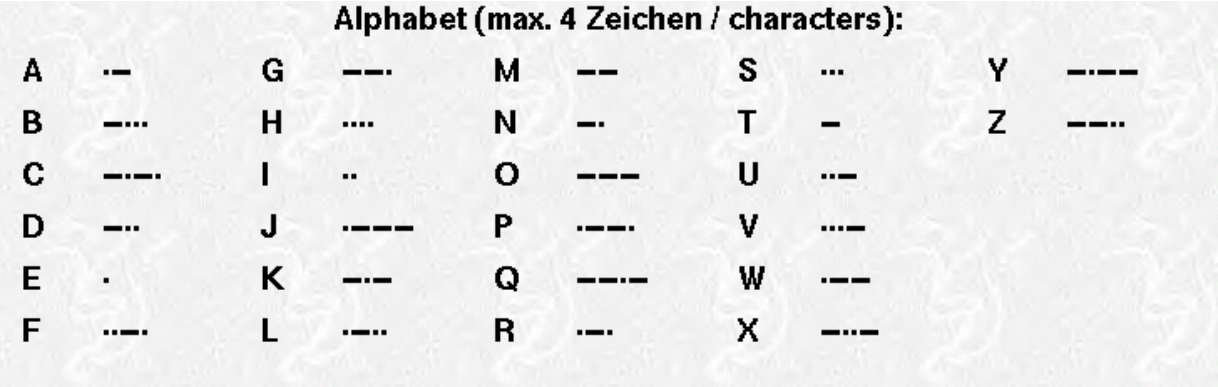
$$p(x) = \frac{1}{Z} \exp(\sum_i \lambda_i f_i(x)) \quad (7)$$

$$\log p(x) = -\log Z + \sum_i \lambda_i f_i(x) \quad (8)$$

  - **Log**arithmus der Wahrscheinlichkeit ist **lineare** Funktion der Merkmalswerte (mit Gewichten als Koeffizienten)

Hausaufgaben

Das Morsealphabet kodiert Buchstaben des lateinischen Alphabets im Binärkode (der Einfachheit halber ignorieren wir das Leerzeichen zwischen den kodierten Buchstaben).



In deutschen Texten sind die einzelnen Buchstaben folgendermaßen verteilt:

1.	E	17.40%	14.	M	2.53%
2.	N	9.78%	15.	O	2.51%
3.	I	7.55%	16.	B	1.89%
4.	S	7.27%	17.	W	1.89%
5.	R	7.00%	18.	F	1.66%
6.	A	6.51%	19.	K	1.21%
7.	T	6.15%	20.	Z	1.13%
8.	D	5.08%	21.	P	0.79%
9.	H	4.76%	22.	V	0.67%
10.	U	4.35%	23.	J	0.27%
11.	L	3.44%	24.	Y	0.04%
12.	C	3.06%	25.	X	0.03%
13.	G	3.01%	26.	Q	0.02%

1. Was ist die Entropie eines Buchstabens in einem deutschen Text, wenn man Groß-/Kleinschreibung, Interpunktion und Leerzeichen ignoriert und die Wahrscheinlichkeit eines Buchstaben mit seiner relativen Häufigkeit gleichsetzt?

2. Was ist die erwartete Länge des Kodes eines deutschen Buchstaben im Morsealphabet?

3. Was ist die erwartete Länge des Kodes eines deutschen Buchstaben im ASCII-Code?

4. (freiwillig) Benutzen Sie ein Komprimierungsprogramm wie „gzip“ und einen selbstgewählten deutschen Text[2], um eine obere Schranke für die tatsächliche Entropie eines Buchstaben im Deutschen zu ermitteln.

[2] Zum Beispiel das Negra-Corpus; zugänglich von /home/jaeger/NEGRA/negra-corpus.sent auf hyperion