Maximum Entropy Modeling

4274: Maschinelles Lernen in der Sprachverarbeitung

WS 2004/05

Karl-Michael Schneider

Universität Passsau

Modellierung von Wahrscheinlichkeiten

Klassifizierung = Abschätzen von Wahrscheinlichkeiten

- $p(\text{Klasse}, \text{Ereignis})$
- $p(\text{Klasse}|\text{Ereignis})$

Ereignis = Merkmalsvektor $\vec{x} = \langle x_1 \ldots x_n \rangle$

Naive Bayes Annahme: Merkmale x_i sind unabhängig

$$p(\vec{x}|\text{Klasse}) = \prod_i p(x_i|\text{Klasse})$$

In der Realität sind Merkmale nicht unabhängig

Parameterglättung, um Null-Wahrscheinlichkeiten zu vermeiden

Dies führt zu Modelleigenschaften, die nicht durch die beobachteten (Trainings-) Daten gerechtfertigt sind (**Bias**).

Prinzip der maximalen Entropie

Entropie einer Wahrscheinlichkeitsverteilung: Maß für die Unvorhersagbarkeit eines Ereignisses

- $p(x_0) = 1, p(x) = 0 \, \forall x \neq x_0$ hat Entropie 0.
- Die uniforme Verteilung $p(x) =$ *konstant* hat maximale Entropie.

Aufgabe: Abschätzen von $p(a, b)$

Trainingsdaten enthalten Evidenz für das gemeinsame Auftreten von a und b, aber nicht genug, um p vollständig zu spezifizieren.

Gesucht: Modell $\hat{p}$ für p

Prinzip der maximalen Entropie:
Wähle die Verteilung $\hat{p}$ mit der größten Entropie, die mit der beobachteten Evidenz (Trainingsdaten) vereinbar ist.

Verteilung mit maximaler Entropie

Beispiel

- $a \in \{x, y\}, b \in \{0, 1\}$
- Beobachtet: $\tilde{p}(x, 0) + \tilde{p}(y, 0) = \tilde{p}(b = 0) = 0.6$

Mögliche Verteilungen, die mit der Evidenz vereinbar sind:

$p_1(a,b)$	0	1	
x	0.5	0.1	
y	0.1	0.3	
$\sum$	0.6	0.4	1.0

$p_2(a,b)$	0	1	
x	0.3	0.2	
y	0.3	0.2	
$\sum$	0.6	0.4	1.0

$H(p_1) = 1.168$, $H(p_2) = 1.366$

p_2 = Maximum-Entropie-Verteilung

Repräsentation von Evidenz

Kodieren von Eigenschaften der Daten durch Merkmale

Merkmal = binäre Funktion $f_i : \mathcal{E} \to \{0, 1\}$

Erwartung von f_i bzgl. p: $\quad E_p f_i = \sum_{x \in \mathcal{E}} p(x) f_i(x)$

Beispiel: $a \in \{x, y\}$, $b \in \{0, 1\}$

$$f_1(a, b) = \begin{cases} 1 & \text{falls } b = 0 \\ 0 & \text{sonst} \end{cases}$$

Evidenz: $\quad \tilde{p}(x, 0) + \tilde{p}(y, 0) = 0.6 \quad$ i.e. $E_{\tilde{p}} f_1 = 0.6$

Repräsentation von Evidenz

Betrachte Evidenz als **Bedingungen** (Constraints) für Modelle

Modell p ist konsistent mit Evidenz $\tilde{p}$, falls für alle $i = 1, \ldots, n$ gilt:

$$E_p f_i = E_{\tilde{p}} f_i$$

Menge der konsistenten Modelle:

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i,\ i = 1, \ldots, n\}$$

Prinzip der maximalen Entropie: wähle das konsistente Model mit der größten Entropie:

$$p^* = \operatorname*{argmax}_{p \in P} H(p)$$

Modellklasse

Darroch & Ratcliff (1972): p^* ist eindeutig bestimmt

p^* hat die Form:

$$p^*(x) = \frac{1}{Z} \prod_{i=1}^{n} \alpha_i^{f_i(x)} \qquad \text{(conditional exponentional model)}$$

$$= \frac{1}{Z} \exp\Big(\sum_{i=1}^{n} f_i(x) \log \alpha_i\Big) \qquad \text{(log linear model)}$$

α_i = Gewichte, Z = Normalisierungskonstante

Für die Gewichte α_i existiert im allgemeinen keine geschlossene Form (bei überlappenden Merkmalen)

Bestimmen der Gewichte mit iterativem Algorithmus

Maximum Likelihood

Log-Likelihood der empirischen Verteilung $\tilde{p}$ bzgl. des Modells p:

$$L_{\tilde{p}}(p) = \sum_x \tilde{p}(x) \log p(x)$$

Maximum Entropy und Maximum Likelihood sind duale Eigenschaften:

$$p^* = \operatorname*{argmax}_{p \in P} L_{\tilde{p}}(p)$$

d.h. das Modell mit der maximalen Entropie ist gleichzeitig das Modell, unter dem die empirische Verteilung maximale Wahrscheinlichkeit besitzt.

Generalized Iterative Scaling

Generalized Iterative Scaling (Darroch & Ratcliff 1972): Algorithmus zum Bestimmen von p^* (unter Beachtung der Bedingungen $E_{p^*} f_i = E_{\tilde{p}} f_i$)

Zusatzbedingung: $\forall x \sum_{i=1}^{n} f_i(x) = C \quad (C = \text{Konstante})$

Verwende Korrekturmerkmal f_{n+1}:

- $C = \max_x \sum_{i=1}^{n} f_i(x)$
- $f_{n+1}(x) = C - \sum_{i=1}^{n} f_i(x)$

Improved Iterative Scaling (Berger et al. 1996): kommt ohne die Zusatzbedingung aus

Generalized Iterative Scaling

1. Initialisierung: $k = 1$, $\alpha_i^{(1)} = 1$ $(i = 1, \ldots, n+1)$

2. $p^{(k)}(x) = \frac{1}{Z^{(k)}} \prod_{i=1}^{n+1} (\alpha_i^{(k)})^{f_i(x)}, \quad Z^{(k)} = \sum_x \prod_{i=1}^{n+1} (\alpha_i^{(k)})^{f_i(x)}$

3. $E_{p^{(k)}} f_i = \sum_x p^{(k)}(x) f_i(x)$

4. $\alpha_i^{(k+1)} = \alpha_i^{(k)} \left(\frac{E_{\tilde{p}} f_i}{E_{p^{(k)}} f_i} \right)^{\frac{1}{C}}$

5. Stop falls $L_{\tilde{p}}(p^{(k)}) - L_{\tilde{p}}(p^{(k-1)}) < \varepsilon$ oder $k \geq$ *MaxIterations*

6. $k := k + 1$, weiter bei 2.

Darroch & Ratcliff (1972): $p^{(k)}$ konvergiert zu p^*

Generalized Iterative Scaling

Bei Klassifikationsproblemen ist p von der Form $p(y|x)$

Trainingsbeispiele: $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

Log-Likelihood:

$$L_{\tilde{p}}(p) = \sum_{x,y} \tilde{p}(x, y) \log p(y|x) = \frac{1}{N} \sum_{j=1}^{N} \log p(y_j|x_j)$$

Empirische Erwartung:

$$E_{\tilde{p}} f_i = \sum_{j=1}^{N} \tilde{p}(x_j, y_j) f_i(x_j, y_j) = \frac{1}{N} \sum_{j=1}^{N} f_i(x_j, y_j)$$

Generalized Iterative Scaling

$E_{p(k)} f_i$ kann im allgemeinen nicht effizient berechnet werden: erfordert Summierung über alle (x, y)

Für n (überlappende) Merkmale f_i können bis zu 2^n unterscheidbare Ereignisse (x, y) existieren

Lösung: Verwende Approximation:

$$\begin{aligned} E_{p^{(k)}} f_i &\approx \sum_{x,y} \tilde{p}(x) p^{(k)}(y|x) f_i(x, y) \\ &= \frac{1}{N} \sum_{j=1}^{N} \sum_{y} p^{(k)}(y|x_j) f_i(x_j, y) \end{aligned}$$

$$p^{(k)}(y|x) = \frac{1}{Z^{(k)}(x)} \prod_{i=1}^{n+1} (\alpha_i^{(k)})^{f_i(x,y)} \qquad Z^{(k)}(x) = \sum_{y} \prod_{i=1}^{n+1} (\alpha_i^{(k)})^{f_i(x,y)}$$