Mustererkennung: Diskriminatives Lernen

D. Schlesinger – TUD/INF/KI/IS

Diskriminatives Lernen – Idee

Gegeben sie eine (parametrisierte) Klasse der Wahrscheinlichkeitsverteilungen.

Jeder konkreten Wahrscheinlichkeitsverteilung entspricht ein Klassifikator.

Das eingentliche Ziel ist die Erkennung – die Anwendung des Klassifikators.

Generatives Lernen – der Ausgangspunkt ist die Klasse der WV:

1. Die Parameter der WV werden gesucht (z.B. nach dem ML).
2. Der Klassifikator wird daraus abgeleitet (z.B. nach dem Bayesschen Prinzip).
3. Der Klassifikator wird angewendet – Erkennung.

Diskriminatives Lernen – der Ausgangspunkt ist die Klasse der Klassifikatoren:

1. Die Parameter des Klassifikators werden „direkt“ gesucht.
2. Der Klassifikator wird angewendet – Erkennung.

Ist die (parametrisierte) Klasse der Klassifikatoren explizit angegeben,
so wird die Klasse der WV nicht mehr benötigt.

Diskriminatives Lernen – Vorgehensweise

„Die Wahrheit (Parameter des darunterliegenden Modells) ist egal. Wichtig ist nur, dass alles am Ende gut funktioniert (die vorgegebene Lernstichprobe wird richtig klassifiziert)“.

Beispiel:
zwei Klassen, Gausssiane gleicher Varianz als bedingte Wahrscheinlichkeiten für $x \in \mathbb{R}^n$:

$$p(x,k) = p(k) \cdot \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{\|x-\mu^k\|^2}{2\sigma^2}\right]$$

Klassifikator ist eine Hyperebene
$\rightarrow$ suche gleich nach einer „guten“ Hyperebene $\langle w, x\rangle < b$.

Vergleiche: $2 \cdot n + 2$ freie Parameter der WV gegen n freie Parameter des Klassifikators.

Konsequenz: einem Klassifikator entsprechen viele WV.

Beispiel: Die lage der Hyperebene im obigen Beispiel hängt von σ gar nicht ab.
Zentren μ^1 und μ^2 sind nicht relevant, sondern die Differenz $\mu^1 - \mu^2$.

Diskriminatives Lernen – Empirisches Risiko

Wie ist eine gute Hyperebene zu finden?

Das Bayessche Risiko:

$$R_b(e) = \sum_x \sum_k p(x,k)\, C\Big(e(x), k\Big) \to \min_e$$

Jetzt hat man aber keine $p(x,k)$. Man hat nur eine Lernstichprobe $L = \Big((x^l, k^l) \ldots\Big)$
Das Bayessche Risiko wird durch das **Empirische** Risiko ersetzt:

$$R_e(e) = \sum_l C\Big(e(x^l), k^l\Big) \to \min_{e \in \mathcal{E}}$$

wobei $\mathcal{E}$ die Klasse der Klassifikatoren ist.

Beispiel:

a) lineare Klassifikatoren, d.h. $\langle w, x\rangle < b$ mit unbekannten $w \in \mathbb{R}^n$ und $b \in \mathbb{R}$
b) $\mathbb{1}(k \neq k')$ als Kostenfunktion,
c) Annahme: es existiert ein Klassifikator, der die vorgegebene Klassifikation wiedergibt

$\to$ Perzeptron Algorithmus.

VC-Dimension

Die Frage: wie genau ist das Lernen?

Satz (Vapnik, Chervonenkis):
notwendige und hinreichende Bedingungen für Konsistenz des empirischen Risikos:

$$\lim_{|L|\to\infty} \{\sup_e |R_e - R_b| > \varepsilon\} = 0 \quad \forall \varepsilon > 0$$

Obere Schranke für die Differenz der Fehler:

$$P\left\{ |R_b - R_e| < \sqrt{\frac{h\Big(\log(2N/h) + 1\Big) - \log(\delta/4)}{N}} \right\} > 1 - \delta$$

mit VC-Dimension h.
(notw. und hinr. Bedingungen für Konsistenz des empirischen Risikos: $h < \infty$).

VC-Dimension:
Die kleinste Zahl n so, dass kein $(n+1)$-Tupel von Punkten beliebig Klassifizierbar ist.

Beispiel: lineare Klassifikatoren im $\mathbb{R}^n$, VC=$n+1$.

Diskriminatives Lernen – Bedingte Wahrscheinlichkeiten

Eine etwas andere Vorgehensweise:
$p(x)$ ist bei der Erkennung irrelevant, wichtig ist nur $p(k|x)$
$\rightarrow$ definiere die Klasse der **bedingten** WV explizit ($p(x)$ ist „beliebig“).

Maximum-Likelihood:

$$p(L;\Theta) = \prod_l \Big[p(x^l) \cdot p(k^l|x^l;\Theta) \Big]$$

$$\ln p(L;\Theta) = \sum_l \ln p(x^l) + \sum_l \ln p(k^l|x^l;\Theta)$$

Beispiel: wieder die Gaussiane gleicher Streuung:

$$p(k=1|x) = \frac{p(1)p(x|1)}{p(1)p(x|1) + p(2)p(x|2)} = \frac{1}{1 + \frac{p(2)p(x|2)}{p(1)p(x|1)}} =$$

$$= \frac{1}{1 + \exp\left[-\frac{\|x-\mu^2\|^2}{2\sigma^2} + \frac{\|x-\mu^1\|^2}{2\sigma^2} + \ln p(2) - \ln p(1) \right]} =$$

$$= \frac{1}{1 + \exp\Big(\langle w, x \rangle + b \Big)}$$

Fragen/Nachteile: kein unüberwachtes Lernen möglich, VC $= \infty$ (?).