Mustererkennung: Bayessche Entscheidungen

D. Schlesinger – TUD/INF/KI/IS

Erkennung

Gegeben sei zwei statistische Größen.
Typischerweise ist eine davon diskret (d.h. $k \in K$) und heißt „Klasse".
Die andere ist allgemein (sehr oft kontinuierlich, d.h. $x \in \mathbb{R}^n$) und heißt „Beobachtung".

„Gegeben" sei die Verbundwahrscheinlichkeitsverteilung $p(x, k)$.
Da die k diskret ist, wird oft $p(x, k)$ durch $p(x, k) = p(k) \cdot p(x|k)$ spezifiziert.

Die Erkennungsaufgabe:
man beobachtet x, man sage etwas über k
– „welche Klasse hat die Beobachtung x verursacht".

Menge der Entscheidungen D, Entscheidungsstrategie $e : X \to D$
Kostenfunktion $C : D \times K \to \mathbb{R}$

Das Bayessche Risiko:

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

Spezialfall (fast immer) – die Menge der Entscheidungen ist nicht eingeschränkt:

$$R\Big(e(x)\Big) = \sum_k p(x,k) \cdot C\Big(e(x), k\Big) \to \min_{e(x)}$$

Noch spezieller (sehr oft) – $D = K$
die Menge der Entscheidungen ist die Menge der Klassen:

$$k^* = \arg\min_k \sum_{k'} p(x,k') \cdot C(k,k')$$

Maximum Aposteriori Entscheidung (MAP)

Die Kostenfunktion ist (die einfachste)

$$C(k, k') = \mathbb{1}(k \neq k')$$

Daraus folgt die Maximum Aposteriori Entscheidung

$$\begin{aligned} R(k) &= \sum_{k'} p(k'|x) \cdot \mathbb{1}(k \neq k') = \\ &= \sum_{k'} p(k'|x) - p(k|x) = 1 - p(k|x) \to \min_k \end{aligned}$$

$$p(k|x) \to \max_k$$

MAP, Beispiel

$K = \{1, 2\}$, $x \in \mathbb{R}^2$, $p(k)$ sei gegeben, bedingte Wahrscheinlichkeit ist Gaussch:

$$p(x|k) = \frac{1}{2\pi\sigma_k^2} \exp\left[-\frac{\|x-\mu_k\|^2}{2\sigma_k^2}\right]$$

Kostenfunktion ist $\mathbb{1}(k \neq k')$, d.h. MAP.
Wie sieht die Entscheidungsstrategie (Abbildung) $e : X \to K$ aus?

$$p(1) \cdot \frac{1}{2\pi\sigma_1^2} \exp\left[-\frac{\|x-\mu_1\|^2}{2\sigma_1^2}\right] > p(2) \cdot \frac{1}{2\pi\sigma_2^2} \exp\left[-\frac{\|x-\mu_2\|^2}{2\sigma_2^2}\right]$$

Spezialfall: $\sigma_1 = \sigma_2$ (Ableitung an der Tafel)

$$\langle x, \mu_2 - \mu_1 \rangle > const$$

Linearer Klassifikator – die zum Vektor $\mu_2 - \mu_1$ senkrechte Trennebene.

Mehrere Klassen, gleiche σ $\to$ Voronoi Diagramm
Zwei Klassen, unterschiedliche σ $\to$ eine quadratische Kurve
usw.

Entscheidungen mit Rückweisung

Die Menge der Entscheidungen $D = K \cup \{rw\}$, die Kostenfunktion

$$C(d,k) = \begin{cases} \mathbb{1}(d \neq k) & \text{wenn} \quad d \in K \\ \varepsilon & \text{wenn} \quad d = rw \end{cases}$$

Fallunterscheidung:

1) $d \in K$, Entscheidung: $d = k^*$ (MAP, wie vorher), Kosten dafür: $1 - p(k^*|x)$
2) $d = rw$, Kosten dafür: ε

Vergleiche $p(k^*|x)$ mit $1 - \varepsilon$ und entscheide für die Variante mit größerem Wert.

Andere einfache Kostenfunktionen

Die Menge K sei strukturiert.

Beispiel:
Wahrscheinlichkeitsmodell: $p(x,y) = p(y) \cdot p(y|x)$, $x, y \in \mathbb{R}$, Sei $D = Y = \mathbb{R}$

Das Bayessche Risiko:

$$R\Big(d(x)\Big) = \int_{-\infty}^{\infty} p(y|x) \cdot C\Big(d(x), y\Big)\, dy$$

($p(y|x)$ sei beherrschbar).

Kostenfunktion berücksichtigt die Abweichung der Entscheidung vom „wahren“ Wert
Zum Beispiel $C(d,y) = (d-y)^2$, dann (Ableitung an der Tafel)

$$\begin{aligned} e(x) &= \arg\min_d \int_{-\infty}^{\infty} p(y|x) \cdot (d-y)^2\, dy \\ &= \int_{-\infty}^{\infty} y \cdot p(y|x)\, dy \quad = \quad \mathbb{E}_{p(y|x)}\, y \end{aligned}$$

Andere Möglichkeiten:
$C(d,y) = |d-y|$, $C(d,y) = \mathbb{1}(|d-y| > \delta)$, Kombination mit Rückweisung usw.

Die Klasse ist ein Vektor $\bar{k} = (k_1, k_2, \ldots, k_n) \in K^n$, die Entscheidungsmenge sei $D = K$
Die a-posteriori Wahrscheinlichkeit $p(k_1, k_2, \ldots, k_n|x)$ sei bekannt.

Variante 1: MAP, d.h. $C(k, k') = 1\!\!1(k \neq k')$

$$(k_1, k_2, \ldots, k_n)^* = \arg\max_k p(k_1, k_2, \ldots, k_n|x)$$

Variante 2: Kostenfunktionen gibt es für jeden Element k_i:

$$C(k, k') = \sum_i c_i(k_i, k'_i)$$

$$\begin{aligned}
R(\bar{k}) &= \sum_{\bar{k}'} p(\bar{k}') \cdot \sum_i c_i(k_i, k'_i) = \sum_i \sum_{\bar{k}'} c_i(k_i, k'_i) \cdot p(\bar{k}') = \\
&= \sum_i \sum_k \sum_{\bar{k}' : k'_i = k} c_i(k_i, k) \cdot p(\bar{k}') = \\
&= \sum_i \sum_k c_i(k_i, k) \sum_{\bar{k}' : k'_i = k} p(\bar{k}') = \sum_i \sum_k c_i(k_i, k) p(k'_i = k)
\end{aligned}$$

Additive Kostenfunktionen

1) Man berechne

$$p(k_i = k) = \sum_{\bar{k}: k_i = k} p(\bar{k}) \quad \forall i, k$$

2) Man treffe die Entscheidung für alle Elemente „unabhängig"

$$k_i^* = \arg\min_k \sum_{k'} p(k'_i = k) \cdot c_i(k, k')$$

$c(k, k') = \mathbb{1}(k \neq k')$ – Max-Marginal Entscheidung.

Andere Möglichkeiten: Kombination mit Rückweisung, „metrische" additive Kostenfunktionen usw.