

Mustererkennung: Maximum-Likelihood Prinzip

D. Schlesinger – TUD/INF/KI/IS

Das übergeordnete Thema – Lernen

Gegeben sei eine parametrisierte Klasse (Familie) der Wahrscheinlichkeitsverteilungen, d.h. $P(x; \Theta) \in \mathcal{P}$.

Beispiel – die Menge aller Gaussiane im $\mathbb{R}^n$

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

parametrisiert mit dem Mittelwert $\mu \in \mathbb{R}^n$ und $\sigma \in \mathbb{R}$, d.h. $\Theta = (\mu, \sigma)$

Eine Lernstichprobe steht zur Verfügung: z.B. $L = (x^1, x^2, \dots, x^{|L|})$ mit $x^l \in \mathbb{R}^n$.

Man entscheide sich für eine Wahrscheinlichkeitsverteilung aus der vorgegebenen Familie, d.h. für einen Parametersatz (z.B. $\Theta^* = (\mu^*, \sigma^*)$ für den Gaussian).

Die Lernstichprobe ist eine Realisierung der unbekannten Wahrscheinlichkeitsverteilung, sie ist entsprechend der Wahrscheinlichkeitsverteilung gewürfelt.



Das, was beobachtet wird, hat hohe Wahrscheinlichkeit



Maximiere die Wahrscheinlichkeit der Lernstichprobe bezüglich der Parameter:

$$p(L; \Theta) \rightarrow \max_{\Theta}$$

Allgemeine diskrete Wahrscheinlichkeitsverteilung für $k \in K$,
d.h. $\Theta = p(k) \in \mathbb{R}^{|K|}$, $p(k) \geq 0$, $\sum_k p(k) = 1$.

Lernstichprobe $L = (k^1, k^2, \dots, k^{|L|})$, $k^l \in K$.

Annahme (sehr oft):

Die Elemente der Lernstichprobe werden unabhängig von einander generiert.

ML:

$$p(L; \Theta) = \prod_l p(k^l) = \prod_k \prod_{l: k^l = k} p(k) = \prod_k p(k)^{n(k)}$$

mit den Häufigkeiten $n(k)$ der Werte k in der Lernstichprobe.

$$\ln p(L; \Theta) = \sum_k n(k) \ln p(k) \rightarrow \max_p$$

oder (bei einer unendlichen Lernstichprobe)

$$\ln p(L; \Theta) = \sum_k p^*(k) \ln p(k) \rightarrow \max_p$$

$$\sum_i a_i \ln x_i \rightarrow \max_x, \quad \text{s.t. } x_i \geq 0 \quad \forall i, \quad \sum_i x_i = 1 \quad \text{mit } a_i \geq 0$$

Methode der Lagrange Koeffizienten:

$$F = \sum_i a_i \ln x_i + \lambda \left(\sum_i x_i - 1 \right) \rightarrow \min_{\lambda} \max_x$$

$$\frac{dF}{dx_i} = \frac{a_i}{x_i} + \lambda = 0$$

$$\frac{dF}{d\lambda} = \sum_i x_i - 1 = 0$$

$$x_i = c \cdot a_i$$

$$\sum_i c \cdot a_i - 1 = 0$$

$$x_i = \frac{a_i}{\sum_{i'} a_{i'}}$$

Die optimale Wahrscheinlichkeitsverteilung für das Beispiel 1:

Zähle, wieviel mal jedes k in der Lernstichprobe vorhanden ist und normiere auf 1.

Zum Beispiel Gaussian, d.h. eine parametrisierte Wahrscheinlichkeitsdichte

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

d.h. $\Theta = (\mu, \sigma)$, mit $\mu \in \mathbb{R}^n$, $\sigma \in \mathbb{R}$.

Lernstichprobe $L = (x^1, x^2, \dots, x^{|L|})$ – jeder Wert von x ist nur ein Mal (?) vorhanden.

ML:

$$\begin{aligned} \ln p(L; \mu, \sigma) &= \sum_l \left[-n \ln \sigma - \frac{\|x^l - \mu\|^2}{2\sigma^2} \right] = \\ &= -|L| \cdot n \cdot \ln \sigma - \frac{1}{2\sigma^2} \sum_l \|x^l - \mu\|^2 \rightarrow \max_{\mu, \sigma} \end{aligned}$$

$$\frac{d \ln p(L; \mu, \sigma)}{d\mu} = 0 \quad \Rightarrow \quad \mu = \frac{1}{|L|} \sum_l x^l$$

$$\frac{d \ln p(L; \mu, \sigma)}{d\sigma} = 0 \quad \Rightarrow \quad \sigma = \frac{1}{n \cdot |L|} \sum_l \|x^l - \mu\|^2$$

Das für die Erkennung typische Modell: $p(x, k; \Theta) = p(k; \Theta_a) \cdot p(x|k; \Theta_k)$, mit $k \in K$ (Klassen, diskret) und $x \in X$ (Beobachtung, allgemein).

Die unbekannten Parameter sind $\Theta_a = p(k)$ und Klassenspezifische Θ_k

Die Lernstichprobe besteht aus Paaren: $L = ((x^1, k^1), (x^2, k^2), \dots, (x^{|L|}, k^{|L|}))$

ML:

$$\begin{aligned} \ln p(L; \Theta) &= \sum_l \left[\ln p(k^l) + \ln p(x^l | k^l; \Theta_{k^l}) \right] = \\ &= \sum_k n(k) \ln p(k) + \sum_k \sum_{l: k^l = k} \ln p(x^l | k; \Theta_k) \rightarrow \max_{p(k), \Theta_k} \end{aligned}$$

Kann bezüglich $\Theta_a, \Theta_1, \dots, \Theta_{|K|}$ getrennt optimiert werden.

Dies war ein überwachtes Lernen.

Unüberwachtes Lernen: das Modell ist dasselbe, d.h. $p(x, k; \Theta)$

In der Lernstichprobe ist die Information unvollständig, d.h. $L = (x^1, x^2, \dots, x^{|L|})$

ML:

$$\ln p(L; \Theta) = \sum_l \ln p(x^l) = \sum_l \ln \sum_k p(x, k; \Theta) \rightarrow \max_{\Theta}$$

Expectation-Maximization Algorithmus:

Man führt eine „Nährhafte Eins“ wie folgt ein:

$$\sum_l \left[\sum_k \alpha_l(k) \ln p(k, x; \Theta) - \sum_k \alpha_l(k) \ln \frac{p(k, x; \Theta)}{\sum_{k'} p(k', x; \Theta)} \right]$$

mit $\alpha_l(k) \geq 0$, $\sum_k \alpha_l(k) = 1$ für alle l .

Dann ist dieser Ausdruck dem oberen äquivalent (Beweis auf der Tafel).

$$\ln p(L; \Theta) = F(\Theta, \alpha) - G(\Theta, \alpha), \quad \text{mit}$$

$$F(\Theta, \alpha) = \sum_l \sum_k \alpha_l(k) \ln p(k, x; \Theta)$$

$$G(\Theta, \alpha) = \sum_l \sum_k \alpha_l(k) \ln \frac{p(k, x; \Theta)}{\sum_{k'} p(k', x; \Theta)} = \sum_l \sum_k \alpha_l(k) \ln p(k|x; \Theta)$$

Man starte mit einem beliebigen Parametersatz $\Theta^{(0)}$ und wiederhole:

Expectation Schritt – „die Fehlenden Daten vervollständigen“:

Man wähle $\alpha^{(t)}$ so, dass das Maximum von G bezüglich Θ genau an der Stelle $\Theta^{(t)}$ eintritt.
Laut Schannonsches Lemma:

$$\alpha_l^{(t)}(k) = p(k|x^l; \Theta^{(t)})$$

Achtung!!! Das ist keine Optimierung, das ist eine Abschätzung der oberen Schranke für G .

Maximization Schritt – „überwachtes Lernen“:

Man maximiere F bezüglich Θ :

$$\Theta^{(t+1)} = \arg \max_{\Theta} F(\Theta, \alpha^{(t)})$$

Der Maximum-Likelihood Schätzer ist **konsistent**,
d.h. er liefert die tatsächlichen Parameter bei unendlichen Lernstichproben.

Der Maximum-Likelihood Schätzer ist nicht immer **erwartungswerttreu**,
d.h. bei endlichen Lernstichproben stimmt der Mittelwert des geschätzten Parameters
nicht unbedingt mit dem tatsächlichen überein.

Beispiele: ML für μ ist erwartungswerttreu, ML für σ – nicht.

Expectation-Maximization Algorithmus konvergiert immer,
aber nur zum lokalen Optimum (nicht global).