TECHNISCHE UNIVERSITÄT DRESDEN

ALGORITHMEN UND DATENSTRUKTUREN

ÜBUNG 14: EM-ALGORITHMUS

Eric Kunze

eric.kunze@mailbox.tu-dresden.de

TU Dresden, 24.01.2022

letzte Änderung:
22.01.2022, 19:03

Aufgabe 1

WAHRSCHEINLICHKEITSTHEORIE

Wir betrachten ein Zufallsexperiment (X, p) mit

- Ergebnismenge X und
- einer Funktion $p\colon X \to [0,1]$ mit $\sum_{x \in X} p(x) = 1$ (**Wahrscheinlichkeitsverteilung von** X)

WAHRSCHEINLICHKEITSTHEORIE

Wir betrachten ein Zufallsexperiment (X, p) mit

- Ergebnismenge X und
- einer Funktion $p\colon X \to [0, 1]$ mit $\sum_{x \in X} p(x) = 1$ (**Wahrscheinlichkeitsverteilung von X**)

Die Menge aller Wahrscheinlichkeitsverteilungen über X sei $\mathcal{M}(X)$. Jede Teilmenge $\mathcal{M} \subseteq \mathcal{M}(X)$ heißt **Wahrscheinlichkeitsmodell**.

WAHRSCHEINLICHKEITSTHEORIE

Wir betrachten ein Zufallsexperiment (X, p) mit

- Ergebnismenge X und
- einer Funktion $p\colon X \to [0,1]$ mit $\sum_{x \in X} p(x) = 1$ (**Wahrscheinlichkeitsverteilung von** X)

Die Menge aller Wahrscheinlichkeitsverteilungen über X sei $\mathcal{M}(X)$. Jede Teilmenge $\mathcal{M} \subseteq \mathcal{M}(X)$ heißt **Wahrscheinlichkeitsmodell**.

Ein Wahrscheinlichkeitsmodell $\mathcal{M}$ heißt **beschränkt**, falls $\mathcal{M} \neq \mathcal{M}(X)$; andernfalls unbeschränkt.

WAHRSCHEINLICHKEITSTHEORIE

Wir betrachten ein Zufallsexperiment (X, p) mit

- Ergebnismenge X und
- einer Funktion $p: X \to [0,1]$ mit $\sum_{x \in X} p(x) = 1$ (**Wahrscheinlichkeitsverteilung von X**)

Die Menge aller Wahrscheinlichkeitsverteilungen über X sei $\mathcal{M}(X)$. Jede Teilmenge $\mathcal{M} \subseteq \mathcal{M}(X)$ heißt **Wahrscheinlichkeitsmodell**.

Ein Wahrscheinlichkeitsmodell $\mathcal{M}$ heißt **beschränkt**, falls $\mathcal{M} \neq \mathcal{M}(X)$; andernfalls unbeschränkt.

Führen wir nun zwei Zufallsexperimente nacheinander aus und nehmen dabei an, dass die beiden Experimente *unabhängig* voneinander sind. Folge das erste Experiment einer Verteilung $p_1 \in \mathcal{M}(X_1)$ und das zweite Experiment einer Verteilung $p_2 \in \mathcal{M}(X_2)$, dann ist $p_1 \times p_2 \in \mathcal{M}(X_1 \times X_2)$ eine Verteilung auf der Ergebnismenge $X_1 \times X_2$ unseres zweistufigen Experiments:

$$(p_1 \times p_2)(a, b) = p_1(a) * p_2(b).$$

„Einzelwahrscheinlichkeiten multiplizieren / erste Pfadregel“

KORPORA UND KORPUSWAHRSCHEINLICHKEITEN

Oftmals wissen wir aber die zugrundeliegende Verteilung nicht, sondern können lediglich die Ergebnisse des Experiments wahrnehmen. Zählen wir diese Beobachtungen, dann nennen wir das einen X**-Korpus** modelliert durch eine Funktion $h\colon X \to \mathbb{R}^{\geq 0}$. Man definiert die **Korpuswahrscheinlichkeit** / **Likelihood** von h unter einer Verteilung p als

$$L(h, p) = \prod_{x \in X} p(x)^{h(x)}.$$

Nun kennen wir aber die Verteilung p nicht und müssen sie daher aus den beobachteten Daten schätzen. Dies macht der **Maximum-Likelihood-Schätzer** (MLE)

$$\mathrm{mle}(h, \mathcal{M}) = \operatorname*{arg\,max}_{p \in \mathcal{M}} L(h, p).$$

Solange das Modell unbeschränkt gewählt wird, d.h. es werden alle Verteilungen über X zugelassen, dann wird der MLE zur relativen Häufigkeit von h.

UNVOLLSTÄNDIGE DATEN

Bisher sind wir davon ausgegangen, dass die Daten stets vollständig waren, d.h. wir konnten jedes Ergebnis beobachten. In der Realität können aber oftmals nur Gruppen von Ergebnissen beobachtet werden; z.B. gewinne oder verliere ich bei einem Spiel. Wir wissen aber nicht, welches Ergebnis genau erzielt wurde.

Sei Y die Menge der Beobachtungen. Die Beobachtungsfunktion $\mathrm{yield}\colon X \to Y$ ordnet jedem Ergebnis seine Beobachtung zu. Die Umkehrabbildung ordnet dann jeder Beobachtung eine Menge von möglichen Ergebnissen zu, die zu dieser Beobachtung führen, d.h.

$$A\colon Y \to \mathcal{P}(X) \quad \text{mit} \quad A(y) = \{x \in X : \mathrm{yield}(x) = y\}\,.$$

Diese Funktion heißt **Analysator**.

Sei h ein Y-Korpus, d.h. h zählt Beobachtungen (nicht Ergebnisse). Die **Korpuswahrscheinlichkeit** / **Likelihood** von h unter einer Verteilung p ist

$$L(h, p) = \prod_{y \in Y} \left(\sum_{x \in A(y)} p(x) \right)^{h(y)}\,.$$

Der MLE bleibt wie er war: $\mathrm{mle}(h, \mathcal{M}) = \arg\max_{p \in \mathcal{M}} L(h, p)$.

AUFGABE 1

Bestimmen Sie für die folgenden Szenarien die Menge X der Ergebnisse und die Menge Y der Beobachtungen. Bestimmen Sie außerdem den Analysator.

(a) Werfen zweier unabhängiger Münzen. Sie können nur beobachten, ob beide Münzen dieselbe oder verschiedene Seiten zeigen.

(b) Werfen zweier Würfel, wobei Sie nur die Summe der Augenzahlen beobachten.

(c) Zwei Spieler spielen Schere-Stein-Papier. Sie beobachten lediglich, welcher Spieler gewonnen hat bzw. ob das Spiel unentschieden ausging.

AUFGABE 1

(a) **zweimaliger Münzwurf – Beobachtung der Gleichheit**

$$X = \{K, Z\} \times \{K, Z\} \qquad \text{und} \qquad Y = \{\text{gleich}, \text{ungleich}\}$$

Der Analysator ordnet jeder Beobachtung $y \in Y$ die Menge der Ergebnisse aus X zu, die zur Beobachtung y führen, also

$$A(\text{gleich}) = \{(K, K), (Z, Z)\}$$
$$A(\text{ungleich}) = \{(K, Z), (Z, K)\}\,.$$

(b) **zweimaliger Würfelwurf – Beobachtung der Augensumme**

$$X = \{1, \ldots, 6\} \times \{1, \ldots, 6\} \qquad \text{und} \qquad Y = \{2, \ldots, 12\}$$

Analysator: $A(x) = \{(i, j) \in X : i + j = x\}$, d.h. konkret

$$A(2) = \{(1, 1)\}$$
$$A(3) = \{(1, 2), (2, 1)\}$$
$$A(4) = \{(1, 3), (3, 1), (2, 2)\}$$
$$\vdots$$
$$A(12) = \{(6, 6)\}$$

AUFGABE 1

(c) **Schere, Stein, Papier – Beobachtung des Gewinners**

$$X = \{\text{Schere}, \text{Stein}, \text{Papier}\}^2$$
$$Y = \{\text{Spieler1}, \text{Spieler2}, \text{Unentschieden}\}$$

Analysator:

$$A(\text{Spieler1}) = \{(\text{Schere}, \text{Papier}), (\text{Stein}, \text{Schere}), (\text{Papier}, \text{Stein})\}$$
$$A(\text{Spieler2}) = \{(\text{Papier}, \text{Schere}), (\text{Schere}, \text{Stein}), (\text{Stein}, \text{Papier})\}$$
$$A(\text{Unentschieden}) = \{(\text{Papier}, \text{Papier}), (\text{Stein}, \text{Stein}), (\text{Schere}, \text{Schere})\}$$

Aufgabe 2

PROBLEME MIT UNVOLLSTÄNDIGEN DATEN

Problem 1: Wir wollen die Verteilung der *Ergebnisse* schätzen. Um den MLE nutzen zu können, brauchen wir aber einen X-Korpus. Jedoch ist uns nur ein Y-Korpus gegeben, da wir nur *Beobachtungen* wahrnehmen können.

Ausweg: Erweiterung des Y-Korpus h zu einem X-Korpus h_1 auf vollständigen Daten

$$h_i(x) = h(\mathrm{yield}(x)) \cdot \frac{p_{i-1}(x)}{\sum_{x' \in A(\mathrm{yield}(x))} p_{i-1}(x)} \qquad \text{für alle } x \in X$$

Dazu benötigen wir eine gewisse Vorkenntnis mit der Verteilung p_{i-1}, die wir aus dem vorherigen Iterationsschritt bzw. einer initialen Vermutung bekommen.

Problem 2: Oftmals haben wir zwei unabhängige Teilversuche mit Ergebnismengen X_1 und X_2 und das zugehörige Modell besteht ebenso aus dem unabhängigen Produkt der Modelle auf den Teilversuchen. Wie wir sehen werden, ist dann ein X-Korpus h nicht unbedingt hilfreich – wir wollen vielmehr die Teilkorpora h^1 auf X_1 und h^2 auf X_2.

Ausweg: Marginalisierung

Wir betrachten die zwei Ergebnismengen X_1 und X_2. Das Modell sei gegeben durch das unabhängige Produkt der Modelle auf X_1 und der Modelle auf X_2, d.h. $\mathcal{M} = \left\{p^1 \times p^2 : p^1 \in \mathcal{M}(X_1), p^2 \in \mathcal{M}(X_2)\right\}$. Weiter sei h ein $X_1 \times X_2$-Korpus. Die Teilkorpora h^1 auf X_1 und h^2 auf X_2 erhalten wir durch **Marginalisierung**

$$h^1(x_1) = \sum_{x_2 \in X_2} h(x_1, x_2) \quad \text{für alle } x_1 \in X_1$$

$$h^2(x_2) = \sum_{x_1 \in X_1} h(x_1, x_2) \quad \text{für alle } x_2 \in X_2$$

Die Summen entsprechen dabei gerade Zeilen- bzw. Spaltensummen, wenn man h in einer Tabelle notiert.

$X_1 \backslash X_2$	α	$\ldots$	ω	
a	$h(a,\alpha)$	$\ldots$	$h(a,\omega)$	$h^1(a)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
z	$h(z,\alpha)$	$\ldots$	$h(z,\omega)$	$h^1(z)$
	$h^2(\alpha)$	$\ldots$	$h^2(\omega)$	

Der MLE auf $X_1 \times X_2$ im unbeschränkten Modell ist gegeben durch die relativen Häufigkeiten auf den Teilkorpora, d.h. $\mathrm{mle}(h, \mathcal{M}) = \mathrm{rfe}(h^1) \times \mathrm{rfe}(h^2)$.

EM-ALGORITHMUS

Input: initiale Wahrscheinlichkeitsverteilung p_0

Eine Iteration des Algorithmus besteht aus den folgenden beiden Schritten:

E-Schritt *Expectation*
Bestimmte die versteckten Eigenschaften mithilfe der Parameter aus der vorherigen Iteration.

$$h_i(x) = h(\mathrm{yield}(x)) \cdot \frac{p_{i-1}(x)}{\sum_{x' \in A(\mathrm{yield}(x))} p_{i-1}(x)}$$

M-Schritt *Maximization*
Bestimmte die neuen Parameter mithilfe des vollständigen Eigenschaften aus dem E-Schritt.

$$p_i = \arg\max_{p \in \mathcal{M}} L(h_i, p)$$

AUFGABE 2 — TEIL (A)

Das Spiel wird gewonnen, wenn beide Münzen auf der gleichen Seite landen.

AUFGABE 2 — TEIL (A)

Das Spiel wird gewonnen, wenn beide Münzen auf der gleichen Seite landen.

Damit ist der Analysator $A\colon \{\mathsf{Gewinn}, \mathsf{keinGewinn}\} \to \mathcal{P}(X)$ gegeben durch

$$\begin{aligned} A(\mathsf{Gewinn}) &= \{x \in X : \mathsf{yield}(x) = \mathsf{Gewinn}\} \\ &= \{(K, K), (Z, Z)\} \\ A(\mathsf{keinGewinn}) &= \{x \in X : \mathsf{yield}(x) = \mathsf{keinGewinn}\} \\ &= \{(K, Z), (Z, K), (R, K), (R, Z)\} \end{aligned}$$

AUFGABE 2 — TEIL (B)

Wir können nur die Beobachtungen Gewinn und $\mathsf{keinGewinn}$ feststellen.

Wir spielen das Spiel 24 Mal und gewinnen 6 Mal. Gesucht ist nun der Y-Korpus h, d.h. wie oft beobachten wir die Ereignisse Gewinn und $\mathsf{keinGewinn}$.

AUFGABE 2 — TEIL (B)

Wir können nur die Beobachtungen Gewinn und keinGewinn feststellen.

Wir spielen das Spiel 24 Mal und gewinnen 6 Mal. Gesucht ist nun der Y-Korpus h, d.h. wie oft beobachten wir die Ereignisse Gewinn und keinGewinn.

$$h(\text{Gewinn}) = 6 \qquad\qquad h(\text{keinGewinn}) = 18$$

AUFGABE 2 — TEIL (C)

Gegeben ist nun eine initiale Wahrscheinlichkeitsverteilung $q_0 = q_0^1 \times q_0^2$ über den vollständigen Daten mit

$$q_0^1(K) = \frac{2}{5} \qquad q_0^2(K) = \frac{1}{3}$$

$$q_0^1(R) = \frac{1}{5}$$

$$\Rightarrow q_0^1(Z) = 1 - q_0^1(K) - q_0^1(R) = \frac{2}{5} \qquad q_0^2(Z) = 1 - q_0^1(K) = \frac{2}{3}$$

AUFGABE 2 — TEIL (C)

Gegeben ist nun eine initiale Wahrscheinlichkeitsverteilung $q_0 = q_0^1 \times q_0^2$ über den vollständigen Daten mit

$$q_0^1(K) = \frac{2}{5} \qquad q_0^2(K) = \frac{1}{3}$$

$$q_0^1(R) = \frac{1}{5}$$

$$\Rightarrow q_0^1(Z) = 1 - q_0^1(K) - q_0^1(R) = \frac{2}{5} \qquad q_0^2(Z) = 1 - q_0^1(K) = \frac{2}{3}$$

Mit dem unabhängigen Produkt erhalten wir

$$q_0(K,K) = q_0^1(K) \cdot q_0^2(K) = \frac{2}{15} \qquad q_0(K,Z) = q_0^1(K) \cdot q_0^2(Z) = \frac{4}{15}$$

$$q_0(Z,K) = q_0^1(Z) \cdot q_0^2(K) = \frac{2}{15} \qquad q_0(Z,Z) = q_0^1(Z) \cdot q_0^2(Z) = \frac{4}{15}$$

$$q_0(R,K) = q_0^1(R) \cdot q_0^2(K) = \frac{1}{15} \qquad q_0(R,Z) = q_0^1(R) \cdot q_0^2(Z) = \frac{2}{15}$$

AUFGABE 2 — TEIL (C)

E-Schritt: Erweiterung von h auf h_1 mit folgender Formel:

$$h_1(x) = h(\text{yield}(x)) \cdot \frac{q_0(x)}{\sum\limits_{x' \in A(\text{yield}(x))} q_0(x')}$$

AUFGABE 2 — TEIL (C)

E-Schritt: Erweiterung von h auf h_1 mit folgender Formel:

$$h_1(x) = h(\text{yield}(x)) \cdot \frac{q_0(x)}{\sum\limits_{x' \in A(\text{yield}(x))} q_0(x')}$$

Damit ergibt sich dann zum Beispiel für das Ergebnis (K, K)

$$\begin{aligned} h_1(K,K) &= h(\text{Gewinn}) \cdot \frac{q_0(K,K)}{\sum\limits_{x' \in \{(K,K),(Z,Z)\}} q_0(x')} \\ &= h(\text{Gewinn}) \cdot \frac{q_0(K,K)}{q_0(K,K) + q_0(Z,Z)} \\ &= 6 \cdot \frac{\frac{2}{15}}{\frac{2}{15} + \frac{4}{15}} \\ &= 2 \end{aligned}$$

AUFGABE 2 — TEIL (C)

E-Schritt: Erweiterung von h auf h_1 mit folgender Formel:

$$h_1(x) = h(\text{yield}(x)) \cdot \frac{q_0(x)}{\sum_{x' \in A(\text{yield}(x))} q_0(x')}$$

Damit ergibt sich dann zum Beispiel für das Ergebnis (K, K)

$$\begin{aligned} h_1(K, K) &= h(\text{Gewinn}) \cdot \frac{q_0(K, K)}{\sum_{x' \in \{(K,K),(Z,Z)\}} q_0(x')} \\ &= h(\text{Gewinn}) \cdot \frac{q_0(K, K)}{q_0(K, K) + q_0(Z, Z)} \\ &= 6 \cdot \frac{\frac{2}{15}}{\frac{2}{15} + \frac{4}{15}} \\ &= 2 \end{aligned}$$

Mit gleicher Rechnung erhalten wir für die restlichen Ereignisse

$$\begin{array}{ccc} & h_1(Z, K) = 4 & h_1(R, K) = 2 \\ h_1(K, Z) = 8 & h_1(Z, Z) = 4 & h_1(R, Z) = 4 \end{array}$$

AUFGABE 2 — TEIL (D)

M-Schritt: Bestimmung der Teilkorpora h_1^1 bzw. h_1^2 durch *Marginalisierung*:

AUFGABE 2 — TEIL (D)

M-Schritt: Bestimmung der Teilkorpora h_1^1 bzw. h_1^2 durch *Marginalisierung*:

$X_1 \backslash X_2$	K	Z	
K	$h_1(K,K)$	$h_1(K,Z)$	$h_1^1(K)$
Z	$h_1(Z,K)$	$h_1(Z,Z)$	$h_1^1(Z)$
R	$h_1(R,K)$	$h_1(R,Z)$	$h_1^1(R)$
	$h_1^2(K)$	$h_1^2(Z)$	

$\rightsquigarrow$

$X_1 \backslash X_2$	K	Z	
K	2	8	10
Z	4	4	8
R	2	4	6
	8	16	24

AUFGABE 2 — TEIL (E)

Nun bestimmen wir noch die relativen Häufigkeiten mit der Formel

$$\mathrm{rfe}(h)(x) := \frac{h(x)}{|h|} \quad \text{mit} \quad |h| := \sum_{x \in X} h(x)$$

AUFGABE 2 — TEIL (E)

Nun bestimmen wir noch die relativen Häufigkeiten mit der Formel

$$\mathrm{rfe}(h)(x) := \frac{h(x)}{|h|} \quad \text{mit} \quad |h| := \sum_{x \in X} h(x)$$

Wenden wir dies nun auf h_1 und h_2 an, so erhalten wir

$$q_1^1(K) = \mathrm{rfe}(h_1^1)(K) = \frac{h_1^1(K)}{h_1^1(K) + h_1^1(Z) + h_1^1(R)} = \frac{10}{24} = \frac{5}{12}$$

$$q_1^1(Z) = \mathrm{rfe}(h_1^1)(Z) = \frac{h_1^1(Z)}{h_1(K) + h_1^1(Z) + h_1^1(R)} = \frac{8}{24} = \frac{1}{3}$$

$$q_1^1(R) = \mathrm{rfe}(h_1^1)(R) = \frac{h_1^1(R)}{h_1^1(K) + h_1^1(Z) + h_1^1(R)} = \frac{6}{24} = \frac{1}{4}$$

und

$$q_1^2(K) = \mathrm{rfe}(h_1^2)(K) = \frac{h_1^2(K)}{h_1^2(K) + h_1^2(Z)} = \frac{8}{24} = \frac{1}{3}$$

$$q_1^2(Z) = \mathrm{rfe}(h_1^2)(Z) = \frac{h_1^2(Z)}{h_1^2(K) + h_1^2(Z)} = \frac{16}{24} = \frac{2}{3}$$