EM-Algorithmus

Eric Kunze

5. Februar 2021

Mit dieser Lösung ist keine Garantie auf Vollständigkeit und/oder Korrektheit verbunden!

Aufgabe 1

Wir betrachten den Wurf zweier Münzen, wobei die erste Münze auch auf dem Rand landen kann. Dementsprechend gibt es für die erste Münze die Möglichkeiten $\{K, Z, R\}$, für die zweite “normale” Münze nur die Möglichkeiten $\{K, Z\}$. Insgesamt gibt es also die Möglichkeiten

$$X := \{K, Z, R\} \times \{K, Z\} = \{(K, K), (K, Z), (Z, K), (Z, Z), (R, K), (R, Z)\} \quad .$$

Wir gewinnen ein Spiel, wenn beide Münzen gleich landen.

(a) Gesucht ist nun der Analysator für dieses Spiel. In der Vorlesung wurde eine Beobachtungsfunktion

$$\text{yield}\colon X \to \{\text{Gewinn}, \text{keinGewinn}\}$$

eingeführt, die jedem Ergebnis der Ergebnismenge eine Beobachtung zuordnet, d.h. ob man bei einem Ergebnis gewinnt oder nicht. Beispielsweise ist $\text{yield}(K, K) = \text{Gewinn}$ aber $\text{yield}(R, Z) = \text{keinGewinn}$. Da wird nun aber nicht den konkreten Ausgang des Spiels erfahren, sondern nur die Beobachtung, ob gewonnen wurde oder nicht, benötigen wir alle Ergebnisse, die zu dieser Beobachtung geführt haben könnten. Das wird mathematisch gesehen das Urbild einer Beobachtung unter der yield-Abbildung. Dieses definiert uns eine neue Abbildung

$$A\colon \{\text{Gewinn}, \text{keinGewinn}\} \to \mathcal{P}(X) \quad ,$$

die wir **Analysator** nennen. Der Analysator liefert also für eine gegebene Beobachtung

die Menge der zugehörigen Ergebnisse. Somit ist also

$$A(\text{Gewinn}) = \{x \in X : \text{yield}(x) = \text{Gewinn}\} = \{(K,K),(Z,Z)\} \ ,$$
$$A(\text{keinGewinn}) = \{x \in X : \text{yield}(x) = \text{keinGewinn}\} = \{(K,Z),(Z,K),(R,K),(R,Z)\} \ .$$

(b) Nun können wir nicht mehr den Korpus über X betrachten, weil wir kennen nicht mehr die exakten Ergebnisse, sondern wir müssen auf den Korpus über $Y := \{\text{Gewinn}, \text{keinGewinn}\}$ ausweichen. Das nennt man dann Korpus mit unvollständigen Daten (weil wir eben nicht mehr alles wissen).

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.

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

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

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

Diese Verteilung können wir nun noch ergänzen, da sich die Wahrscheinlichkeiten zu 1 aufaddieren müssen. Somit ist also

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

Nun können wir das unabhängige Produkt nutzen. Dieses ist ungefähr das, was man aus der Schule als Pfadregel kennt, wo man entlang eines Pfades multipliziert. Damit erhalten wir dann

$$q_0(K,K) = q_0^1(K) \cdot q_0^2(K) = \frac{2}{5} \cdot \frac{1}{3} = \frac{2}{15} \qquad q_0(K,Z) = q_0^1(K) \cdot q_0^2(Z) = \frac{2}{5} \cdot \frac{2}{3} = \frac{4}{15}$$
$$q_0(Z,K) = q_0^1(Z) \cdot q_0^2(K) = \frac{2}{5} \cdot \frac{1}{3} = \frac{2}{15} \qquad q_0(Z,Z) = q_0^1(Z) \cdot q_0^2(Z) = \frac{2}{5} \cdot \frac{2}{3} = \frac{4}{15}$$
$$q_0(R,K) = q_0^1(R) \cdot q_0^2(K) = \frac{1}{5} \cdot \frac{1}{3} = \frac{1}{15} \qquad q_0(R,Z) = q_0^1(R) \cdot q_0^2(Z) = \frac{1}{5} \cdot \frac{2}{3} = \frac{2}{15}$$

Im E-Schritt des EM-Algorithmus muss nun der Korpus h zu einem Korpus h_1 erweitert

werden. Dies geschieht mit folgender Formel:

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

Das sieht jetzt zwar kompliziert aus, aber lässt sich relativ einfach vereinfachen bzw. einfacher lesen. Der Summationsbereich $A(\text{yield}(x))$ besteht genau aus den Elementen, die die gleiche Beobachtung haben wie x, also zum Beispiel

$$A(\text{yield}(K,K)) = A(\text{Gewinn}) = \{(K,K),(Z,Z)\} \ .$$

Der Vorfaktor $h(\text{yield}(x))$ lässt sich auch relativ leicht berechnen, wir nehmen einfach die Beobachtung von x und deren Korpus, d.h. zum Beispiel

$$h(\text{yield}(K,K)) = h(\text{Gewinn}) = 6 \ .$$

Damit ergibt sich dann

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

Mit gleicher Rechnung erhält man für die restlichen Ereignisse

$$\begin{aligned}
h_1(K,Z) = 8 \ , \qquad & h_1(Z,K) = 4 \ , \qquad & h_1(R,K) = 2 \ , \\
& h_1(Z,Z) = 4 \ , \qquad & h_1(R,Z) = 4 \ .
\end{aligned}$$

(d) Nun führen wir noch den M-Schritt aus und bestimmen die Teilkorpora h_1^1 bzw. h_1^2 durch **Marginalisierung**:

$$h_1^1(K) = \sum_{x \in \{K,Z\}} h_1(K,x) = h_1(K,K) + h_1(K,Z) = 2 + 8 = 10 \ .$$

Effektiv addiert man also die Korpora, wo K in der *ersten* Komponenten vorkommt.

Schließlich erhält man h_1^1 vollständig mit

$$h_1^1(K) = 10 \ , \qquad h_1^1(Z) = 8 \ , \qquad h_1^1(R) = 6 \ .$$

Für h_1^2 erhält man

$$\begin{aligned} h_1^2(K) &= \sum_{x \in \{K,Z,R\}} h_1(x,K) = h_1(K,K) + h_1(Z,K) + h_1(R,K) = 2+4+2 = 8 \ , \\ h_1^2(Z) &= \sum_{x \in \{K,Z,R\}} h_1(x,Z) = h_1(K,Z) + h_1(Z,Z) + h_1(R,Z) = 8+4+4 = 16 \ , \end{aligned}$$

wobei man wiederum die Korpora addiert, wo K in der *zweiten* Komponenten vorkommt. Damit sind h_1^1 und h_1^2 vollständig bestimmt. Man kann die Marginalisierung auch in Matrixschreibweise nachvollziehen, was auch ein bisschen übersichtlicher ist:

$X_1 \backslash X_2$	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

(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

$$\begin{aligned} q_1^1(K) = \mathrm{rfe}(h_1^1)(K) &= \frac{h_1^1(K)}{h_1^1(K) + h_1^1(Z) + h_1^1(R)} = \frac{10}{24} = \frac{5}{12} \ , \\ q_1^1(Z) = \mathrm{rfe}(h_1^1)(Z) &= \frac{h_1^1(Z)}{h_1(K) + h_1^1(Z) + h_1^1(R)} = \frac{8}{24} = \frac{1}{3} \ , \\ q_1^1(R) = \mathrm{rfe}(h_1^1)(R) &= \frac{h_1^1(R)}{h_1^1(K) + h_1^1(Z) + h_1^1(R)} = \frac{6}{24} = \frac{1}{4} \end{aligned}$$

und

$$\begin{aligned} q_1^2(K) = \mathrm{rfe}(h_1^2)(K) &= \frac{h_1^2(K)}{h_1^2(K) + h_1^2(Z)} = \frac{8}{24} = \frac{1}{3} \ , \\ q_1^2(Z) = \mathrm{rfe}(h_1^2)(Z) &= \frac{h_1^2(Z)}{h_1^2(K) + h_1^2(Z)} = \frac{16}{24} = \frac{2}{3} \ . \end{aligned}$$

Zusatzaufgabe 2

Die Personen A und B spielen ein Spiel mit einer Münze mit den Beschriftungen 1 und 2, sowie mit einem dreiseitigen "Würfel" mit den Beschriftungen 1, 2, und 3. In jeder Runde werden die Münze und der Würfel geworfen. Der Spieler A gewinnt die Runde, falls die gefallene Zahl des Würfels kleiner gleich der auf der Münze ist. Ansonsten gewinnt B die Runde. Die Menge der möglichen Ergebnisse ist somit $X = \{1, 2\} \times \{1, 2, 4\}$.

(a) Gesucht ist der Analysator A für dieses Szenario.

$$A(\text{"A gewinnt"}) = \{(1,1), (2,1), (2,2)\}$$
$$A(\text{"B gewinnt"}) = \{(1,2), (1,3), (2,3)\}$$

(b) Der Korpus mit unvollständigen Daten ist gegeben durch

$$h(\text{"A gewinnt"}) = 21 \ ,$$
$$h(\text{"B gewinnt"}) = 10 \ .$$

(c) Gegeben ist eine initiale Wahrscheinlichkeitsverteilung $q_0 = q_0^M \times q_0^W$. Wir vervollständigen zuerst die Angaben von q_0^M und q_0^W:

$$q_0^M(1) = \frac{1}{3} \ , \qquad q_0^W(1) = \frac{1}{4} \ ,$$
$$q_0^M(2) = \frac{2}{3} \ , \qquad q_0^W(2) = \frac{1}{2} \ ,$$
$$q_0^W(3) = \frac{1}{4} \ .$$

Nun können wir q_0 als unabhängiges Produkt berechnen:

$$q_0(1,1) = q_0^M(1) \cdot q_0^W(1) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12} \qquad q_0(2,1) = q_0^M(2) \cdot q_0^W(1) = \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6}$$
$$q_0(1,2) = q_0^M(1) \cdot q_0^W(2) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} \qquad q_0(2,2) = q_0^M(2) \cdot q_0^W(2) = \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$$
$$q_0(1,3) = q_0^M(1) \cdot q_0^W(3) = \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12} \qquad q_0(2,3) = q_0^M(2) \cdot q_0^W(3) = \frac{2}{3} \cdot \frac{1}{4} = \frac{1}{6}$$

(d) Wir führen den E-Schritt aus und Vervollständigen den Korpus h zu einem Korpus h_1:

$$h_1(1,1) = 3 \ , \qquad h_1(2,1) = 6 \ ,$$
$$h_1(1,2) = 4 \ , \qquad h_1(2,2) = 12 \ ,$$
$$h_1(1,3) = 2 \ , \qquad h_1(2,3) = 4 \ .$$

(e) M-Schritt — Wir bestimmen zuerst die Teilkorpora h_1^M und h_1^W durch Marginalisierung:

$M\backslash W$	1	2	3	
1	$h_1(1,1)$	$h_1(1,2)$	$h_1(1,3)$	$h_1^M(1)$
2	$h_1(2,1)$	$h_1(2,2)$	$h_1(2,3)$	$h_1^M(2)$
	$h_1^W(1)$	$h_1^W(2)$	$h_1^W(3)$	

$\rightsquigarrow$

$M\backslash W$	1	2	3	
1	3	4	2	9
2	6	12	4	22
	9	16	6	*31*

(f) Wir schätzen nun die Wahrscheinlichkeitsverteilung q_1^M und q_1^W als relative Häufigkeit der Teilkorpora.

$$q_1^W(1) = \mathrm{rfe}(h_1^W)(1) = \frac{h_1^W(1)}{h_1^W(1) + h_1^W(2) + h_1^W(3)} = \frac{9}{31} \ ,$$
$$q_1^W(2) = \mathrm{rfe}(h_1^W)(2) = \frac{h_1^W(1)}{h_1^W(1) + h_1^W(2) + h_1^W(3)} = \frac{16}{31} \ ,$$
$$q_1^W(3) = \mathrm{rfe}(h_1^W)(3) = \frac{h_1^W(1)}{h_1^W(1) + h_1^W(2) + h_1^W(3)} = \frac{6}{31}$$

und

$$q_1^M(1) = \mathrm{rfe}(h_1^M)(1) = \frac{h_1^M(1)}{h_1^M(1) + h_1^M(2)} = \frac{9}{31} \ ,$$
$$q_1^M(2) = \mathrm{rfe}(h_1^M)(2) = \frac{h_1^M(2)}{h_1^M(1) + h_1^M(2)} = \frac{22}{31} \ .$$