Maschinelles Lernen II

4. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Christoph Sawade

Sommer 2011

Ausgabe am: 19.05.11
Besprechung am: 26.05.11

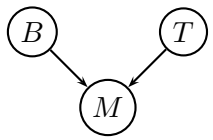
Aufgabe 1 *Unabhängigkeit, bedingte Unabhängigkeit*

Wir betrachten die folgende Domäne: eine faire Münze wird zwei Mal geworfen; die Ergebnisse der Würfe sind repräsentiert durch die Zufallsvariablen $X, Y \in \{0, 1\}$. Wir definieren eine dritte Zufallsvariable Z durch $Z = xor(X, Y)$, dh. Z hat den Wert Eins falls genau eine der beiden Variablen X, Y den Wert Eins hat.

1. Geben Sie die gemeinsame Verteilung $p(X, Y, Z)$ über die Variablen X, Y, Z an.
2. (a) Sind X, Y, Z paarweise unabhängig, dh. gilt $p(X, Y) = p(X)p(Y)$, $p(X, Z) = p(X)p(Z)$, und $p(Y, Z) = p(Y)p(Z)$?
   (b) Sind X, Y, Z unabhängig, dh. gilt $p(X, Y, Z) = p(X)p(Y)p(Z)$?
   (c) Geben Sie alle Unabhängigkeiten an, die in der Verteilung $p(X, Y, Z)$ gelten, als Menge
   $$I(p) := \{(A \perp B|C) \mid p(A|B, C) = p(A|C)\}$$
   wobei A, B, C beliebige Teilmengen von $\{X, Y, Z\}$ sind.
3. Geben Sie ein Bayessches Netz für die Verteilung $p(X, Y, Z)$ an. Gilt $I(G) = I(p)$?

Aufgabe 2 *Parameterlernen*

Wir betrachten das folgende einfache graphisches Modell über den drei binären Variablen T (Tank gefüllt), B (Batteriespannung ok) und M (Motor startet):



$p(B = 1)$
θ_B

$p(T = 1)$
θ_T

$p(M = 1\|B, T)$	B	T
$\theta_{M\|00}$	0	0
$\theta_{M\|01}$	0	1
$\theta_{M\|10}$	1	0
$\theta_{M\|11}$	1	1

Die gemeinsame Verteilung $p(B, T, M) = p(B)p(T)p(M|B, T)$ ist entsprechend den gegebenen Tabellen mit 6 Parametern parametrisiert.

Leider kennen wir die echten Parameterwerte nicht. Wir haben aber die folgenden 10 Beobachtungen des Systems gemacht:

B	T	M
1	1	1
1	1	1
1	0	0
1	0	0
1	0	0
1	0	1
0	1	0
0	1	0
0	1	1
0	0	0

1. Wir möchten die echten Parameterwerte aus den Beobachtungen schätzen. Leiten Sie dazu die Likelihood der Beobachtungen als Funktion der 6 Parameter her. Wir nehmen wie üblich an, dass einzelne Beobachtungen unabhängig sind gegeben das Modell. Berechnen Sie die Parameterwerte $\hat{\theta}_B$, $\hat{\theta}_T$,$\hat{\theta}_{M|00}, ..., \hat{\theta}_{M|11}$ die die Likelihood maximieren.

2. Nehmen Sie an, dass einige Aufzeichnungen über Beobachtungen unleserlich sind, und wir deshalb nur die Tabelle

B	T	M
1	1	1
?	1	1
1	?	0
1	?	0
1	0	?
?	0	1
?	1	0
0	1	0
0	1	1
0	0	0

   zur Verfügung haben, wobei ein "?" bedeutet, dass der entsprechende Wert nicht bekannt ist. Berechnen Sie Schätzungen für die 6 unbekannten Parameter mit den folgenden Verfahren:

   (a) Wir löschen alle Records, in denen ein Wert unbeobachtet ist; anschliessend bestimmen wir die Maximum-Likelihood Parameter.

   (b) Wir schätzen Maximum-Likelihood Parameter. Zum Schätzen eines Parameters verwenden wir einen Record immer dann, wenn alle für die Schätzung relevanten Variablen beobachtet sind. Beispiel: für die Schätzung von $p(B)$ verwenden wir alle Records, für die B beobachtet ist.

(c) Zunächst verfahren wir wie Methode (b). Danach verwenden wir das (vorläufig) gelernte Modell, um mittels Inferenz für alle Records mit unbeobachteten Variablen die jeweils wahrscheinlichsten Werte für diese Variablen zu schätzen, und entsprechend in die Tabelle einzutragen. Wenn Inferenz eine Gleichverteilung ergibt, wird der Wert der Variable auf "?" belassen. Anschliessend schätzen wir mit der "komplettierten" Tabelle das Modell neu.

3. Maximum-Likelihood Parameterschätzungen führen oft zu (unrealistischen) Schätzungen von Null oder Eins für Wahrscheinlichkeiten. Argumentieren Sie, welche Klasse von Prior-Verteilungen geeignet wären, um diese Fälle zu verhindern. Wie sähe die entsprechende a-posteriori Verteilung über Parameter aus, und wie könnte man die entsprechenden maximum-a-posteriori Parameter berechnen?.

Aufgabe 3 *Repräsentation von bedingten Verteilungen*

In der 1. Übung wurde ein modifiziertes Naïve Bayes Modell diskutiert, dass die Verteilung über Merkmale und die Klasse durch

$$p(\mathbf{x}, y) = \prod_{i=1}^{n} p(x_i) p(y|x_1, ..., x_n)$$

modelliert, wobei

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}$$

der Merkmalsvektor ist.

1. Nehmen Sie $n = 3$ an, und die bedingte Verteilung

$p(y = 1\|x_1, x_2, x_3)$	x_1	x_2	x_3
1/2	0	0	0
2/3	1	0	0
3/4	0	1	0
6/7	1	1	0
4/5	0	0	1
8/9	1	0	1
12/13	0	1	1
24/25	1	1	1

Bekanntlich hat dieses Modell das Problem, dass die Verteilung $p(y \mid x_1, ..., x_n)$ exponentiell viele Parameter enthält. Es ist allerdings auch nicht immer nötig, eine solche bedingte Verteilung explizit als Tabelle zu repräsentieren. Die oben dargestellte Verteilung zeigt gewisse Regelmässigkeiten, und kann viel einfacher mit nur drei Parametern charakterisiert werden.

Geben Sie eine Repräsentation der Verteilung an, die mit drei Parametern auskommt. Hinweis: die Sigmoid-Funktion, gegeben durch

$$\sigma(z) = \frac{1}{1 + exp(-z)}$$

könnte hier hilfreich sein, ebenso wie lineare Modelle.

2. Nehmen wir an, wir möchten eine Verteilung $p(y \mid x_1, ..., x_n)$ repräsentieren, die nicht die Regelmässigkeiten der oben gegeben Verteilung zeigt. Wir wollen trotzdem nur wenige Parameter im Modell haben, und sind dafür bereit, die Verteilung nur approximativ zu repräsentieren. Wie könnte man in diesem Fall vorgehen? Hinweis: maschinelles Lernen.