Maschinelles Lernen: Symbolische Ansätze

Übungsblatt für den 19.12.2006

Aufgabe 1

Ein Patient weiß folgendes über einen bestimmten Krebstest: Falls jemand Krebs hat, ist der Test in 98% der Fälle korrekt. Falls jemand keinen Krebs hat, ist der Test in 97% der Fälle korrekt. Insgesamt haben $0,8\%$ der gesamten Bevölkerung Krebs.

Der Patient erhält nun die Nachricht, daß sein Test positiv ist. Was sagt ihm das?

Lösung: Der Patient möchte natürlich wissen, wie hoch die Wahrscheinlichkeit ist, dass er mit der positiven Testvorhersage auch wirklich Krebs hat. Wir kodieren nun folgendes:

- $+ \equiv$ Test ist positiv (sagt Krebs vorher)
- $- \equiv$ Test ist negativ (sagt keinen Krebs vorher)
- $K \equiv$ Krebs vorhanden
- $\neg K \equiv$ Krebs nicht vorhanden

Bekannt sind folgende Wahrscheinlichkeiten:

- $\Pr(K) = 0,008 \Rightarrow \Pr(\neg K) = 0,992$
  (Wahrscheinlichkeit für Krebs und keinen Krebs in der gesamten Bevölkerung)
- $\Pr(+|K) = 0,98 \Rightarrow \Pr(-|K) = 0,02$
  (Wahrscheinlichkeit, dass der Test positiv ist, wenn jemand Krebs hat & Wahrscheinlichkeit, dass der Test negativ ist, wenn jemand Krebs hat)
- $\Pr(-|\neg K) = 0,97 \Rightarrow \Pr(+|\neg K) = 0,03$
  (Wahrscheinlichkeit, dass der Test negativ ist, wenn jemand keinen Krebs hat & Wahrscheinlichkeit, dass der Test positiv ist, wenn jemand keinen Krebs hat)

Der Patient ist an $\Pr(K|+)$ interessiert. Wir wenden das Bayes'sche Theorem an und erhalten:

$$\Pr(K|+) = {}^{\Pr(+|K)\cdot\Pr(K)}/_{\Pr(+)},$$

wobei alle Werte bis auf $\Pr(+)$ bekannt sind. Diese Wahrscheinlichkeit können wir mit dem Theorem der totalen Wahrscheinlichkeiten (Folie 5) berechnen, da sich die Ereignisse K (Krebs vorhanden) und $\neg K$ (Krebs nicht vorhanden) wechselseitig ausschließen. Das Theorem läßt sich in unserem Fall wie folgt anwenden:

$$\begin{aligned}\Pr(+) &= \Pr(+|K) \cdot \Pr(K) + \Pr(+|\neg K) \cdot \Pr(\neg K) \\ &= 0,98 \cdot 0,008 + 0,03 \cdot 0,992 = 0,0376\end{aligned}$$

Nun können wir weiterrechnen:

$$\begin{aligned}\Pr(K|+) &= {}^{0,98 \cdot 0,008}/_{0,0376} = 0,2085 \\ \Rightarrow \Pr(\neg K|+) &= 0,7915\end{aligned}$$

Aufgabe 2

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No
D15	Sunny	Mild	Normal	Weak	No

a) Berechnen Sie die Tabelle der bedingten Wahrscheinlichkeiten, wie sie Naïve Bayes erzeugt.

Lösung: Die Tabelle der bedingten Wahrscheinlichkeiten enthält jede mögliche Klasse, also in unserem Fall die bedingten Wahrscheinlichkeiten für jeden Attributwert und beide Klassen (“yes” und “no”). Möchte man zum Beispiel den Eintrag für $outlook = sunny$ für “yes” berechnen, so muss man $\Pr(sunny|yes)$ ausrechnen.

	outlook			temperature			humidity		wind	
	sunny	overcast	rain	hot	mild	cool	high	normal	weak	strong
yes	2/9	4/9	3/9	2/9	4/9	3/9	3/9	6/9	6/9	3/9
no	4/6	0	2/6	2/6	3/6	1/6	4/6	2/6	3/6	3/6

b) Welchen Klassifikationswert gibt Naïve Bayes für die folgenden Instanzen aus?
Lösung: Um ein Beispiel zu klassifizieren, wird die Klasse ausgegeben, die bei den Attributwerten des Beispiels den höchsten Wert im Produkt mit allen Attributwerten und der a priori-Klassenverteilung erreicht. Wir müssen also einmal für die Klasse "yes" die bedingten Wahrscheinlichkeiten der Attributwerte des Beispiels multiplizieren und einmal für die Klasse "no". Zu beachten ist, dass wir nicht die Wahrscheinlichkeit $\Pr(yes|Bsp.i)$ bzw. $\Pr(no|Bsp.i)$ berechnen, da diese für die Klassifikation nicht notwendig ist. Wir benötigen hierfür nur das Produkt der Wahrscheinlichkeiten der Attributwerte unter Beobachtung der jeweiligen Klasse und der Wahrscheinlichkeit der Klasse. Jedes Beispiel, das wir nun klassifizieren wollen, wird durch seine Attributwerte beschrieben: $(a_1, ..., a_n)$.

1. Outlook=Sunny, Temperature=Mild, Humidity=High, Wind=Strong
   Betrachten wir nun das erste Beispiel:

$$\prod_i \Pr(a_i|yes)\Pr(yes) = 2/9 \cdot 4/9 \cdot 1/3 \cdot 1/3 \cdot 9/15 \approx 0,006584$$
$$\prod_i \Pr(a_i|no)\Pr(no) = 2/3 \cdot 1/2 \cdot 2/3 \cdot 1/2 \cdot 6/15 \approx 0,04444$$

   In diesem Fall hat die Klasse "no" den höheren Wert, weshalb das Beispiel als "no" klassifiziert werden würde. Ein Vorteil vom Naïve Bayes Klassifizierer ist, dass er Wahrscheinlichkeiten für die Klassen ausrechnen kann. Möchte man lediglich klassifizieren, so ist eine Normierung nicht nötig. Möchte man hingegen eine Wahrscheinlichkeit für die Klassen, normiert man mit der Summe der Werte:

$$\Pr(yes|Bsp.1) = {}^{0,006584}/_{0,0510284} \approx 0,129$$
$$\Pr(no|Bsp.1) = {}^{0,04444}/_{0,0510284} \approx 0,871$$

   Bei Multiklassenproblemen ist daher eine Rangfolge möglich. Wenn beispielsweise bei einem Problem mit 20 Klassen eine Klasse eine Wahrscheinlichkeit von $0,9$, kann man sich sehr sicher sein, dass dies die richtige Klasse ist. Bei einem Regellerner würde man beispielsweise einfach diese Klasse vorhergesagt bekommen, ohne zu wissen, wie sicher die Vorhersage ist.

2. Outlook=Rain, Humidity=Normal
   Wir gehen wie beim ersten Beispiel vor:

$$\prod_i \Pr(a_i|yes)\Pr(yes) = 1/3 \cdot 2/3 \cdot 3/5 \approx 0,13333$$

$$\prod_i \Pr(a_i|no)\Pr(no) = 1/3 \cdot 1/3 \cdot 2/5 \approx 0,04444$$

Das zweite Beispiel würde demnach als “yes” klassifiziert werden.

3. Temperature=High

$$\prod_i \Pr(a_i|yes)\Pr(yes) = 2/9 \cdot 3/5 \approx 0,13333$$
$$\prod_i \Pr(a_i|no)\Pr(no) = 1/3 \cdot 2/5 \approx 0,13333$$

In diesem Fall ist der Wert bei beiden Klassen gleich. Hier würde man wie bei anderen Klassifizierern entweder die Majority-Klasse (“yes”) oder eine zufällige Klasse vorhersagen.

c) Wie würden Sie mit fehlenden Attributwerten umgehen?

Fehlende Attributwerte werden bei der Klassifikation einfach übersprungen. Beim Training werden sie nicht mitgezählt.

Aufgabe 3

Gegeben sei folgende Beispielmenge:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	26	High		No
D2	Sunny	28	High	Strong	No
D3	Overcast	29	High	Weak	Yes
D4	Rain	23	High	Weak	Yes
D5	Rain		Normal	Weak	Yes
D6	Rain	12	Normal	Strong	No
D7	Overcast	8		Strong	Yes
D8	Sunny	25	High	Weak	No
D9	Sunny	18	Normal	Weak	Yes
D10	Rain	20	Normal	Weak	Yes
D11	Sunny	20	Normal	Strong	
D12	Overcast	21	High	Strong	Yes
D13		26	Normal	Weak	Yes
D14	Rain	24	High	Strong	No
D15	Sunny	23	Normal	Weak	No
D16	Sunny	21	Normal	Weak	Yes

Überlegen Sie sich, wie Sie diesen Datensatz, der numerische Werte enthält, mit Naïve Bayes behandeln würde.

Lösung: Es existieren u.a. drei Hauptmethoden zur Behandlung von numerischen Attributen bei der Anwendung von Naïve Bayes (siehe auch http://www.cs.waikato.ac.nz/%7Eremco/disc.ps). Die ersten beiden Methoden, *Normal-* und *Kernelmethode*, verwenden zur Approximation der gesuchten Wahrscheinlichkeiten $\Pr(a|c_i)$ (für den Attributwert a des numerischen Attributes A und alle Klassen c_i; in unserem Beispiel: $a = 22$, $A = Temperature$ und c_i ist *yes* oder *no*) eine (nicht-)parametrisierte Verteilung. Wir verwenden die Normalverteilung, die wie folgt berechnet wird:

$$N(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x-\mu^2}{2\sigma^2}}$$

mit dem Mittelwert μ und Varianz σ.

Bei der Normalmethode berechnen wir für jede Klasse c_i den Mittelwert μ_{c_i} und die Varianz σ_{c_i} (deren Berechnung können Sie der Stochastistik-Literatur entnehmen). Mit diesen können wir die gesuchten Wahrscheinlichkeiten wie folgt berechnen:

$$\Pr(a|c_i) = N(a|\mu_{c_i}, \sigma_{c_i})$$

Bei der Kernelmethode approximiert man $\Pr(a|c_i)$ durch eine Anzahl von Normalverteilungen (eine Normalverteilung für jedes Beispiel der Klasse c_i). Der Mittelwert der einzelnen Normalverteilungen wird auf den Attributwert des entsprechenden Beispiels

gesetzt. Die Varianz ist für alle Normalverteilungen gleich und wird auf $\frac{1}{\sqrt{\#c_i}}$ gesetzt, wobei $\#c_i$ der Anzahl der Beispiele der Klasse c_i entspricht. Zur Bestimmung werden dann die Werte der einzelnen Normalverteilungen addiert:

$$\Pr(a|c_i) = \sum_{D \in c_i} N(a|a_D, \frac{1}{\sqrt{\#c_i}}),$$

wobei D ein Beispiel der Klasse c_i ist, a_D dessen Wert für das Attribut A ist.

Bei der dritten Methode, Diskretisierung, handelt es sich eigentlich um eine Gruppe von Methoden. Alle haben gemeinsam, daß sie numerische Attribute in nominale Attribute konvertieren. Die Werte eines numerischen Attributes werden in mehrere, disjunkte Intervalle unterteilt. Die genaue Vorgehensweise wird in den kommenden Vorlesungen erklärt und kann auch dem oben genannten Paper entnommen werden. Die Diskretisierung erfolgt, bevor Naïve Bayes die Tabelle der bedingten Wahrscheinlichkeiten bestimmt. Mit den nun diskreten Daten kann die Tabelle wie gehabt aufgestellt werden. Jedes Intervall entspricht dann jeweils einem nominalen Wert. Bei der Klassifikation von Beispielen müssen diese auch diskretisiert werden. Danach kann die Wahrscheinlichkeit des entsprechenden Intervalles wie gewohnt der Tabelle entnommen werden.

Aufgabe 4

Betrachten sie folgende Regeln:

1. Outlook = Sunny $\rightarrow$ Yes else No
2. Wind = Weak $\rightarrow$ No else Yes
3. Humidity = Normal and $16 <$ Temperature $< 25 \rightarrow$ Yes else No
4. Temperature $> 28 \rightarrow$ Yes else No

Ohne auf die Daten zu schauen, schätzen Sie bitte die Plausibilität jeder einzelnen Regel ein. Weisen Sie jeder Regel diesen Wert als Wahrscheinlichkeit zu.

Betrachten Sie nun den Datensatz:

a) Welche der Regeln ist h_{MAP}, welche h_{ML}?

Lösung: Wir legen wahllos die Wahrscheinlichkeiten der Hypothesen ($\Pr(h)$) wie folgt fest:

$$\Pr(h_1) = 0,5$$
$$\Pr(h_2) = 0,2$$
$$\Pr(h_3) = 0,2$$
$$\Pr(h_4) = 0,1$$

Die Wahrscheinlichkeiten der Daten unter Beobachtung der Hypothesen ($\Pr(D|h)$) schätzen wir ab, indem wir sie durch die relative Häufigkeit der durch die Hypothese korrekt klassifizierten Beispiele abschätzen. Beispiele, die wir mit einer Hypothese nicht klassifizieren können oder deren Klassewert unbekannt ist, ignorierien wir bei der Bestimmung dieser Häufigkeit. Die Vorhersagen der einzelnen Hypothesen können der folgenden Tabelle entnommen werden:

Day	O	T	H	W	Play	h_1	h_2	h_3	h_4
D1	S	26	H		No	Yes	—	**No**	**No**
D2	S	28	H	S	No	Yes	Yes	**No**	**No**
D3	O	29	H	W	Yes	No	No	No	**Yes**
D4	R	23	H	W	Yes	No	No	No	No
D5	R		N	W	Yes	No	No	—	—
D6	R	12	N	S	No	**No**	Yes	**No**	**No**
D7	O	8		S	Yes	No	**Yes**	—	No
D8	S	25	H	W	No	Yes	**No**	**No**	**No**
D9	S	18	N	W	Yes	**Yes**	No	**Yes**	No
D10	R	20	N	W	Yes	No	No	**Yes**	No
D11	S	20	N	S		—	—	—	—
D12	O	21	H	S	Yes	No	**Yes**	No	No
D13		26	N	W	Yes	—	No	No	No
D14	R	24	H	S	No	**No**	Yes	**No**	**No**
D15	S	23	N	W	No	Yes	**No**	Yes	**No**
D16	S	21	N	W	Yes	**Yes**	No	**Yes**	No

Die korrekt klassifizierten Beispiel sind wir die folgenden Wahrscheinlichkeiten: fett markiert. h_1 sagt 4 Beispiele von 14 korrekt voraus, h_2 4 von 14, h_3 8 von 13 und h_4 7 von 14. Damit erhalten

$$\begin{aligned}\Pr(D|h_1) &= \tfrac{4}{14} = \tfrac{2}{7}\\ \Pr(D|h_2) &= \tfrac{4}{14} = \tfrac{2}{7}\\ \Pr(D|h_3) &= \tfrac{8}{13}\\ \Pr(D|h_4) &= \tfrac{7}{14} = \tfrac{1}{2}\end{aligned}$$

Mit diesen Wahrscheinlichkeiten können wir bereits h_{ML} bestimmen:
$h_{ML} = h_3$

Für h_{MAP} müssen wir jedoch noch einige Berechnungen durchführen. Zuerst berechnen wir folgendes:

$$\begin{aligned}\Pr(D|h_1)\Pr(h_1) &= \tfrac{2}{7}\cdot\tfrac{1}{2} = \tfrac{1}{7} \approx 0,143\\ \Pr(D|h_2)\Pr(h_2) &= \tfrac{2}{7}\cdot\tfrac{1}{5} = \tfrac{2}{35} \approx 0,057\\ \Pr(D|h_3)\Pr(h_3) &= \tfrac{8}{13}\cdot\tfrac{1}{5} = \tfrac{8}{65} \approx 0,123\\ \Pr(D|h_4)\Pr(h_4) &= \tfrac{1}{2}\cdot\tfrac{1}{10} = \tfrac{1}{20} = 0,05\end{aligned}$$

Mit diesen Produkten können wir $\Pr(D)$ berechnen (siehe Satz der totalen Wahrscheinlichkeit):

$$\Pr(D) = \sum_{i=1}^{4} \Pr(D|h_i)\Pr(h_i) \approx 0,373$$

Damit können wir nun $\Pr(h|D)$ für alle Hypothesen berechnen:

$$\Pr(h_1|D) \approx 0,384$$
$$\Pr(h_2|D) \approx 0,154$$
$$\Pr(h_3|D) \approx 0,33$$
$$\Pr(h_4|D) \approx 0,134$$

Demnach ist $h_{MAP} = h_1$.

b) Wie lautet die Bayes'sche optimale Klassifikation für die Instanz
Outlook = Sunny, Temperature=22, Humidity=High, Wind=Normal?

Lösung: In a) haben wir die Wahrscheinlichkeiten $\Pr(h|D)$ bereits berechnet. Also fehlen uns für die Berechnung der Bayes'schen optimalen Klassifikation nur noch die Wahrscheinlichkeiten $\Pr(yes|h)$ und $\Pr(no|h)$. Da eine Regelmenge kein Maß für ihr Vertrauen in ihre Vorhersage zur Verfügung stellt. Können wir nur der vorhergesagten Klassen die Wahrscheinlichkeit 1 zuordnen. Gezwungenermaßen ist die Wahrscheinlichkeit der andere Klasse 0. Somit erhalten wir folgende Wahrscheinlichkeiten:

$$\Pr(yes|h_1) = 1 \quad \Pr(no|h_1) = 0$$
$$\Pr(yes|h_2) = 1 \quad \Pr(no|h_2) = 0$$
$$\Pr(yes|h_3) = 0 \quad \Pr(no|h_3) = 1$$
$$\Pr(yes|h_4) = 0 \quad \Pr(no|h_4) = 1$$

Nun können wir die Bayes'sche optimale Klassifikation des Beispiels berechnen:

$$yes : \sum_{i=1}^{4} \Pr(yes|h_i) \cdot \Pr(h_i|D) \approx 0,556$$
$$no : \sum_{i=1}^{4} \Pr(no|h_i) \cdot \Pr(h_i|D) \approx 0,444$$

Demnach ist die Bayes'sche optimale Klassifikation yes.