Mathematische Grundlagen III

Naiv-Bayes-Klassifikation

Garance PARIS

5. Juli 2011

Die Intuition

Die Wahrscheinlichkeit einer Zielkategorie für eine Instanz hängt von zwei Faktoren ab:

- die Gesamtwahrscheinlichkeit der Kategorie überhaupt, Engl. *prior probability (a-priori-Wahrscheinlichkeit)*

  Bsp.: Wie oft kommen *play=yes* bzw. *play=no* insgesamt im Datensatz vor?

- die Wahrscheinlichkeit der Kategorie bei der von der Instanz beigetragenen Evidenz oder Information, Engl. *posterior probability*

  Hier: Wahrscheinlichkeit der einzelnen Attributwerte der Instanz

Das Maximum-a-posteriori

Wir wollen für neue Instanzen die beste Kategorie gegeben der Evidenz finden, auf Englisch das *"maximum a posteriori"*:

$$c_{map} = \underset{c}{\text{argmax}}\, P(c|e) \qquad \text{Bayes'scher Satz}$$

$$= \underset{c}{\text{argmax}}\, \frac{P(c)P(e|c)}{P(e)} \qquad P(e) \text{ fällt weg, weil es konstant und unabhängig von der Hypothese ist}$$

$$= \underset{c}{\text{argmax}}\, P(c)P(e|c)$$

Parameter-Schätzung

- $P(c)$ schätzen ist einfach: Man berechnet die relative Häufigkeit jeder Hypothese in den Trainingsdaten
- $P(e|c)$ schätzen ist wegen Sparse-Data schwieriger, da man nie genug Daten hat, damit alle mögliche Attribut-Kombinationen vorkommen
- Es wird daher angenommen, dass die Attribute unabhängig voneinander sind

$$P(e|c) = P(a_1, a_2, \ldots, a_n|c) = \prod_{i=1}^{n} P(a_i|c)$$

- $P(a|c)$: Relative Häufigkeit des Attributs a unter den Instanzen, die Kategorie c angehören

Beispiel

- Klassifizierung von Instanz *outlook* $=$ *sunny*, *temperature* $=$ *cool*, *humidity* $=$ *high*, *windy* $=$ *true*:

$$\begin{aligned} c_{map} &= \underset{c \in yes,no}{\operatorname{argmax}} P(c) \cdot \prod_i P(a_i|c) \\ &= \underset{c \in yes,no}{\operatorname{argmax}} P(c) \cdot P(outlook = sunny|c) \\ &\quad \cdot P(temperature = cool|c) \\ &\quad \cdot P(humidity = high|c) \\ &\quad \cdot P(windy = true|c) \end{aligned}$$

- Berechnung der relativen Häufigkeit von $P(sunny|yes)$:

  Insgesamt gibt es 9 Instanzen mit *play* $=$ *yes*,
  davon 2 bei denen *outlook* $=$ *sunny*,
  also ist $P(sunny|yes) = \frac{2}{9}$

Vorgehen

- Zuerst die Wahrscheinlichkeit jeder Kategorie in den Trainingsdaten berechnen

  Gesamthäufigkeiten für "play": $yes = 9$, $no = 5$
  Wahrscheinlichkeiten: $yes = \frac{9}{14}$, $no = \frac{5}{14}$

- Für jede Kategorie die Wahrscheinlichkeit der einzelnen Attribute berechnen

		freq.		P	
		yes	*no*	*yes*	*no*
Outlook	sunny	2	3	$\frac{2}{9}$	$\frac{3}{5}$
	overcast	4	0	$\frac{4}{9}$	0
	rainy	3	2	$\frac{3}{9}$	$\frac{2}{5}$
	$\sum$	9	5	1	1
Temperature	...				
Humidity	...				
Windy	...				

Vorgehen, Teil 2

- Wahrscheinlichkeit jeder Kategorie für die vorliegende Instanz berechnen

  $P(yes) \cdot P(sunny|yes) \cdot P(cool|yes) \cdot P(high|yes) \cdot P(true|yes) = \frac{9}{14} * \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = 0.0053$

  $P(no) \cdot P(sunny|no) \cdot P(cool|no) \cdot P(high|no) \cdot P(true|no) = \frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = 0.0206$

- Kategorie zuweisen

$$\begin{aligned} c_{map} &= \underset{c \in yes,no}{\operatorname{argmax}} P(c) \cdot \prod_i P(a_i|c) \\ &= no \end{aligned}$$

Textklassifikation

Bei der Klassifikation von Dokumenten werden die Wörter des Textes als Attribute eingesetzt:

$$\begin{aligned} c_{map} &= \underset{c}{\operatorname{argmax}}\, P(c)P(d|c) \\ &= \underset{c}{\operatorname{argmax}}\, P(c)\prod_i P(w_i|c) \end{aligned}$$

$P(w|c)$: Wahrscheinlichkeit, dass Wort w in einem Dokument mit Klasse c vorkommt

Klassifikation nach Bayes

Anwendung: Spam herausfiltern

- Datensatz: Ein Korpus von E-Mail-Nachrichten, annotiert als “Spam” oder “Ham”
- Aufgabe: Eine neue Nachricht als “Spam” oder “Ham” klassifizieren
- Attribute: Vokabular der E-Mail-Nachrichten im Trainingskorpus

Bsp.: Eine E-Mail mit dem Inhalt “Get rich fast!!!”

$$
\begin{aligned}
c_{map} &= \underset{c \in spam, ham}{\operatorname{argmax}} \; P(c) \cdot \prod_i P(a_i|c) \\
&= \underset{c \in spam, ham}{\operatorname{argmax}} \; P(c) \cdot \\
&\quad P(get|c) \cdot P(rich|c) \cdot P(fast|c) \cdot P(!!!|c)
\end{aligned}
$$

Anwendung: Spam Herausfiltern

- Problem: “rich” und “!!!” sind in Ham-E-Mails selten
- Null-Werte führen dazu, dass Kategorien ununterscheidbar werden
  $P(ham) \cdot P(get|ham) \cdot P(rich|ham) \cdot P(fast|ham) \cdot P(!!!|ham)$
  $= P(ham) \cdot P(get|ham) \cdot 0 \cdot P(fast|ham) \cdot 0 = \boxed{0}$!
- Mögliche Lösung: Kleine Werte zu Zähler und Nenner hinzufügen, damit beide ungleich null werden

$$P(a_x|c) = \frac{n_x + \frac{1}{k}}{n + 1}$$

n : Anzahl Instanzen in Kategorie c
n_x : Anzahl Instanzen in Kategorie c, bei denen $a = x$
k : Anzahl Werte, die Attribut a annehmen kann