Naive Bayes

Ein einfacher Klassifikator

Wolfgang Konen
Fachhochschule Köln
November 2007

Inhalt

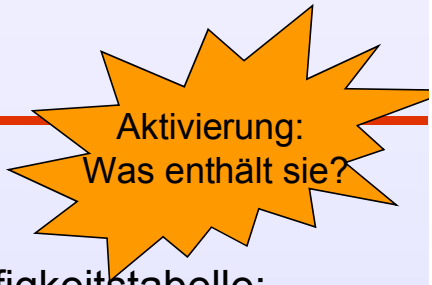
Naive Bayes

- Der Ansatz
- Beispiel Wetterdaten
- Bayes‘sche Regel
- Das Problem der Häufigkeit 0
- Fehlende Werte
- Numerische Werte
- Diskussion

Der Ansatz für Naive Bayes

- Naive Bayes benutzt alle Attribute
- Zwei Annahmen über Attribute:
  - Alle gleich wichtig
  - Statistisch unabhängig (in Bezug auf Klassenwert)
    - D. h. Kenntnis des Wertes eines Attributs sagt nichts über den Wert eines anderen Attributs (wenn die Klasse bekannt ist)
- Unabhängigkeitsannahme stimmt nie!
- Aber … dieses Verfahren funktioniert gut in der Praxis
- Kann auch gut mit fehlenden Werten umgehen

Beispiel Wetterdaten



Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- Häufigkeitstabelle:

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Beispiel Wetterdaten

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Ein neuer Tag:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

„Likelihood“ für beide Klassen:

Für „yes“: 2/9 x 3/9 x 3/9 x 6/9 x 9/14 = 0.0053

Für „no“: 3/5 x 1/5 x 4/5 x 3/5 x 5/14 = 0.0206

Umwandlung in Wahrscheinlichkeiten durch Normalisierung:

Pr(„yes“) = 0.0053 / (0.0053 + 0.0206) = 0.205 = 20.5%

Pr(„no“) =0.0206 / (0.0053 + 0.0206) = 0.795 = 79.5%

Bayes‘sche Regel

- Wahrscheinlichkeit des Ereignisses *H, wenn die Evidenz E gegeben ist* :

$$\Pr(H|E) = \frac{\Pr(E|H)\Pr(H)}{\Pr(E)}$$

Aktivierung: Was besagt sie?

- *Apriori*-Wahrscheinlichkeit von *H* : $\Pr(H)$
  - Wahrscheinlichkeit eines Ereignisses, *bevor* die Evidenz bekannt ist
- *Aposteriori*-Wahrscheinlichkeit *von H* : $\Pr(H|E)$
  - Wahrscheinlichkeit eines Ereignisses, *nachdem* die Evidenz bekannt ist

Thomas Bayes, 1702-1761, England

Naive Bayes für die Klassifikation

- Lernen von Klassifikationen:
  Was ist die Wahrscheinlichkeit einer Klasse, wenn die Instanz gegeben ist?
  - Evidenz *E* = Instanz = Input-Variablen
  - Ereignis *H* = Klassenwert der Instanz = Soll-Output
- Naive Annahme: Evidenz kann aufgeteilt werden auf die Attribute, die *unabhängig* voneinander sind

$$\Pr(H \mid E) = \frac{\Pr(E_1 \mid H) \cdot \Pr(E_2 \mid H) \cdot \ldots \cdot \Pr(E_n \mid H) \cdot \Pr(H)}{\Pr(E)}$$

Wetterdaten-Beispiel

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidenz E*

$$\Pr[yes \mid E] = \Pr[Outlook=Sunny \mid yes] \times \Pr[Temperature=Cool \mid yes] \times \Pr[Humidity=High \mid yes] \times \Pr[Windy=True \mid yes] \times \frac{\Pr[yes]}{\Pr[E]}$$

Wahrscheinlichkeit der Klasse "yes"

$$= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}$$

Was tun mit Pr(E)?

- Die Wahrscheinlichkeit Pr(E) ist allen Faktoren gemeinsam, sie ist aber i.d.R. unbekannt.
- Wir brauchen sie auch überhaupt nicht, wenn wir erst die Zähler ausrechnen, dann die Summe aller Resultatwahrscheinlichkeiten (Zähler) auf 1 normieren:

$$\Pr(\text{"yes"} \mid E) + \Pr(\text{"no"} \mid E) = 1$$

Das Problem der Häufigkeit 0

- Was tun, wenn ein Attributwert nicht mit jedem Klassenwert vorkommt?
  (z. B. “Outlook=overcast” für die Klasse “no”)
  - Wahrscheinlichkeit ergibt sich zu 0!
  - *Aposteriori*–Wahrscheinlichkeit ergibt sich ebenfalls zu 0!
    (Egal, wie groß die anderen Wahrscheinlichkeiten sind!)

$$\Pr(\text{Overcast} \mid \text{"no"}) = 0 \quad \overset{\text{Bayes}}{\Rightarrow} \quad \Pr(\text{"no"} \mid \text{Overcast}) = 0$$

Das Problem der Häufigkeit 0

- Abhilfe: Addiere (kleinen) Wert λ zu den Häufigkeiten in Zähler und Nenner von Pr(E|H):
  (z. B. “Outlook=overcast” für die Klasse “no”)

$$\frac{3+\lambda/3}{5+\lambda}\ \text{"Sunny"} \qquad \frac{0+\lambda/3}{5+\lambda}\ \text{"Overcast"} \qquad \frac{2+\lambda/3}{5+\lambda}\ \text{"Rainy"}$$

- sog. *Laplace-Schätzer*
- Ergebnis: Wahrscheinlichkeiten können nie 0 werden!
  (außerdem: die Wahrscheinlichkeitsschätzungen werden stabilisiert)

Umgang mit fehlenden Werte

- Training:
  - Instanz wird bei der Häufigkeitszählung für die Attributwert-Klassenkombination nicht berücksichtigt
- Klassifikation:
  - Attribut wird bei der Berechnung ausgelassen
- Beispiel:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of “yes” = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238

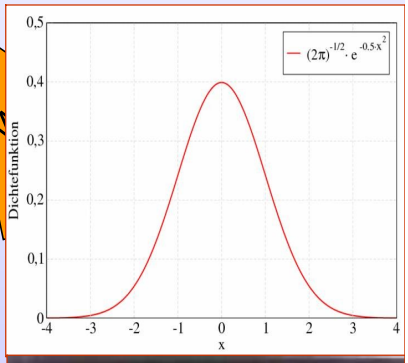
Likelihood of “no” = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343

P(“yes”) = 0.0238 / (0.0238 + 0.0343) = 41%

P(“no”) = 0.0343 / (0.0238 + 0.0343) = 59%

Numerische Werte bei Naive Bayes

- Übliche Annahme: Attributwerte sind normalverteilt (innerhalb jeder Klasse)
- Die *Dichtefunktion* für die Normalverteilung enthält zwei Parameter:
  - *Mittelwert* $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - *Standardabweichung* $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$
  - Dichtefunktion $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Anwendung auf die Wetterdaten

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, …	85, …	80, …	95, …					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

- Beispiel für Dichtewert:

$$f(\text{temperature} = 66 \mid \text{yes}) = \frac{1}{\sqrt{2\pi}\cdot 6.2} e^{-\frac{(66-73)^2}{2\cdot 6.2^2}} = 0.0340$$

Klassifikation bei numerischem Input

- Ein neuer Tag:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of "yes" = $2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$

Likelihood of "no" = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$

P("yes") = 0.000036 / (0.000036 + 0. 000136) = 20.9%

P("no") = 0.000136 / (0.000036 + 0. 000136) = 79.1%

Naive Bayes: Diskussion

- Naiver Bayes funktioniert überraschend gut (selbst wenn die Unabhängigkeitsannahme klar verletzt ist)
- Warum? Weil die Klassifikation keine exakten Wahrscheinlichkeitsschätzungen benötigt, *solange die maximale Wahrscheinlichkeit der korrekten Klasse zugewiesen wird*
- Allerdings: Hinzufügen zu vieler redundanter Attribute führt zu Problemen (z. B. identische Attribute)
- Außerdem: viele numerische Attribute sind nicht normalverteilt (→ *kernel density-Schätzer*)
- Fazit: Naive Bayes eignet sich als „Plain-Vanilla"-Standardverfahren der Klassifikation
  („Wieviel besser als andere Standard-Klassifikationsverfahren ist mein Verfahren?")

Aufgabe: Naive Bayes in R

- Schreiben Sie in R zwei Funktionen
  - `naiveBayes <- function(respvar, d, laplace=1) …`
  - `predict.naiveBayes <- function(res.nb,newdata=d) …`

  die das Modell „Naive-Bayes“ implementieren.
- 1. Version: numerische Input-Variablen, keine MV‘s
- 2. Version: numerische oder nominale Inputs
- 3. Version: auch MV‘s (missing values) erlaubt

- Hinweis: (2 x nlev)-Matrix mit Zeilen- und Spaltennamen:
  - ```
    pmat <- matrix(0,nrow=2,ncol=nlev, dimnames=
            list(c("mean","sig"),levels(d[,respvar])))
    ```
    ```