

Alexander Hofmann

Machine Learning – Naïve Bayes

Naïve Bayes

- Einer der wichtigsten Klassifizierungsalgorithmen
- Über die Frequenztafel wird eine Wahrscheinlichkeit Tabelle aufgebaut
- Ein neuer Datensatz wird über den Satz von Bayes klassifiziert
 - Es werden die Wahrscheinlichkeiten für alle Klassen berechnet, auf Grund der vorliegenden Daten des neuen Datensatzes
 - Die Klasse mit der größten Wahrscheinlichkeit wird genommen
- Funktioniert nur für echt unabhängige Attribute! („naïvely assumes independence“)
- Einfache Verarbeitung von sehr großen Datenmengen
- Unterschieden wird der Umgang mit Namens- und Zahlenwerten

Satz von Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

c ... Classifier value

x ... Attribute value

Klassifikation

- Neuer Datensatz kommt rein:
 - Outlook = sunny (der Rest fehlt) – wie klassifizieren?

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(x|c) = P(\text{sunny}|\text{yes}) = \frac{2}{9} = 0.22$$

$$P(c) = P(\text{yes}) = \frac{9}{14} = 0.64$$

$$P(x) = P(\text{sunny}) = \frac{5}{14} = 0.36$$

$$P(c|x) = P(\text{yes}|\text{sunny}) = \frac{0.22 * 0.64}{0.36} = 0.39$$

$$P(c|x) = P(\text{no}|\text{sunny}) = \frac{0.6 * 0.36}{0.36} = 0.61$$

Likelihood Table		Play		P(x)
		yes	no	
Outlook	sunny	2/9	3/5	5/14
	overcast	4/9	0/5	4/14
	rainy	3/9	2/5	5/14
P(c)		9/14	5/14	

 no

Remember again?

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Frequency & Likelihood Table

Outlook

Frequency Table		Play		
		yes	no	
Outlook	sunny	2	3	5
	overcast	4	0	4
	rainy	3	2	5
		9	5	
Likelihood Table		Play		
		yes	no	
Outlook	sunny	2/9	3/5	5/14
	overcast	4/9	0/5	4/14
	rainy	3/9	2/5	5/14
		9/14	5/14	
Likelihood Table		Play		
		yes	no	
Outlook	sunny	0,22	0,6	0,36
	overcast	0,44	0	0,29
	rainy	0,33	0,4	0,36
		0,64	0,36	

Frequency & Likelihood Table

Temp

Frequency Table		Play		
		yes	no	
Temp	cool	3	1	4
	mild	4	2	6
	hot	2	2	4
		9	5	
Likelihood Table		Play		
		yes	no	
Temp	cool	3/9	1/5	4/14
	mild	4/9	2/5	6/14
	hot	2/9	2/5	4/14
		9/14	5/14	
Likelihood Table		Play		
		yes	no	
Temp	cool	0,33	0,20	0,29
	mild	0,44	0,40	0,43
	hot	0,22	0,40	0,29
		0,64	0,36	

Frequency & Likelihood Table

Humidity

Frequency Table		Play		
		yes	no	
Humidity	normal	6	1	7
	high	3	4	7
		9	5	
Likelihood Table		Play		
		yes	no	
Humidity	normal	6/9	1/5	7/14
	high	3/9	4/9	7/14
		9/14	5/14	
Likelihood Table		Play		
		yes	no	
Humidity	normal	0,67	0,20	0,50
	high	0,33	0,44	0,50
		0,64	0,36	

Frequency & Likelihood Table

Windy

Frequency Table		Play		
		yes	no	
Windy	true	3	3	6
	false	6	2	8
		9	5	
Likelihood Table		Play		
		yes	no	
Windy	true	3/9	3/5	6/14
	false	6/9	2/5	8/14
		9/14	5/14	
Likelihood Table		Play		
		yes	no	
Windy	true	0,33	0,60	0,43
	false	0,67	0,40	0,57
		0,64	0,36	

Klassifikation

- Neuer Datensatz kommt rein:

$x_1 \dots Outlook = rainy, x_2 \dots Temp = mild, x_3 \dots Humidity = normal, x_4 \dots Windy = true$

- Klassifikation?

Naïve Bayes

*Likelihood: $P(c|X) = P(x_1|c) * P(x_2|c) * \dots * P(x_n|c) * P(c)$*

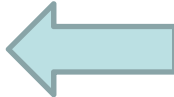
Normalisiert ergibt sich die Wahrscheinlichkeit

$$P(c|X) = \frac{P(c) * \prod_{i=0}^n P(x_i|c)}{P(X)}$$

Naïve Bayes

$x_1 \dots Outlook = rainy, x_2 \dots Temp = mild, x_3 \dots Humidity = normal, x_4 \dots Windy = true$

$$P(yes|X) = \frac{P(x_1|yes) * P(x_2|yes) * P(x_3|yes) * P(x_4|yes) * P(yes)}{P(X)} = \frac{\frac{3}{9} * \frac{4}{9} * \frac{6}{9} * \frac{2}{9} * \frac{9}{14}}{\left(\frac{3}{9} * \frac{4}{9} * \frac{6}{9} * \frac{3}{9} * \frac{9}{14}\right) + \left(\frac{2}{5} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}\right)}$$

= 0.67  **yes**

$$P(no|X) = \frac{P(x_1|no) * P(x_2|no) * P(x_3|no) * P(x_4|no) * P(no)}{P(X)} = \frac{\frac{2}{5} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}}{\left(\frac{3}{9} * \frac{4}{9} * \frac{6}{9} * \frac{3}{9} * \frac{9}{14}\right) + \left(\frac{2}{5} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}\right)} = 0.33$$

Numerical Values

■ Diskretisieren

- 1R (Attribut nach values sortieren, bei jedem Klassenwechsel neuer Range)

64	65	68	69	70	71	72	72	75	75	80	81	83	85
yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

66.5 < x < 70.5

- Optional auch mit Mindestgröße (z.B. 2)
- Equal-width binning, equal-frequency binning, proportional k-interval discretization (unsupervised)
- Entropy-based discretization (+with MDL stopping) (supervised)
- Error-based discretization

Numeric Values in Naïve Bayes

Temp.		
	yes	no
	83	85
	70	80
	68	65
	64	72
	69	71
	75	
	75	
	72	
	81	
μ	73,0	74,6
σ	6,2	7,9

- Wir gehen von einer Normalverteilung oder Gauß'schen Verteilung aus
- Probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

x ... zu testender Wert

σ ... Standardabweichung

μ ... Mittelwert

- Beispiel:

$$f(temp = 66|yes) = \frac{1}{6.2 * \sqrt{2\pi}} * e^{-\frac{(66-73)^2}{2*6.2^2}} = 0,034$$

$$f(temp = 66|no) = \frac{1}{7.9 * \sqrt{2\pi}} * e^{-\frac{(66-74.6)^2}{2*7.9^2}} = 0,0279$$

$$P(temp = 66|yes) = \frac{0,034}{0,034 + 0,0279} = 0,549$$

$$P(temp = 66|no) = \frac{0,0279}{0,034 + 0,0279} = 0,451$$

 yes

Numeric Values in Naïve Bayes

- Die probability density function kann im Naïve Bayes Classifier verwendet werden

$x_1 \dots Outlook = rainy, x_2 \dots Temp = 66, x_3 \dots Humidity = normal, x_4 \dots Windy = true$

$$P(yes|X) = \frac{\frac{2}{9} * 0.034 * \frac{6}{9} * \frac{3}{9} * \frac{9}{14}}{\left(\frac{2}{9} * 0.034 * \frac{6}{9} * \frac{3}{9} * \frac{9}{14}\right) + \left(\frac{3}{5} * 0.0279 * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}\right)} = 0,6$$

$$P(no|X) = \frac{\frac{3}{5} * 0.0279 * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}}{\left(\frac{2}{9} * 0.034 * \frac{6}{9} * \frac{3}{9} * \frac{9}{14}\right) + \left(\frac{3}{5} * 0.0279 * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}\right)} = 0,4$$

Multinomial Naïve Bayes

Document Classification,

- spam or ham data collection
 - spam XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> <http://wap.xxxmobilemovieclub.com?n=QJKGIGHJJGCBL>
 - ham I HAVE A DATE ON SUNDAY WITH WILL!!
- Wörter, die darin vorkommen, klassifizieren die Dokumente (Binär: kommt vor, kommt nicht vor)
- Mehrfachvorkommen sind auch ein interessantes Attribut
- Multinomial Naïve Bayes

Multinomial Naïve Bayes

Document Classification

n ... Anzahl der Vorkommen eines Wortes einer Instanz

$$N = n_1 + n_2 + \dots + n_k$$

P_1, P_2, \dots, P_k Wahrscheinlichkeiten, dass ein Wort i in einer Instanz der Klasse c vorkommt

$$P(x|c) = N! * \prod_{i=1}^k \frac{P_i^{n_i}}{n_i!}$$

Multinomial Naïve Bayes

Document Classification

- Beispiel:

2 Wörter (*credit*, *date*),

$P_1(\text{credit}, \text{spam}) = 75\%$, $P_2(\text{date}, \text{spam})$

$= 25\%$, 1 neu zu klassifizierende Instanz {*date*, *credit*, *date*},

Länge $N = 3$

$$P(\{\text{credit}, \text{credit}, \text{credit}\}|\text{spam}) = 3! * \frac{0,75^3}{3!} * \frac{0,25^0}{0!} = \frac{27}{64} = 0,42$$

$$P(\{\text{credit}, \text{credit}, \text{date}\}|\text{spam}) = 3! * \frac{0,75^2}{2!} * \frac{0,25^1}{1!} = \frac{27}{64} = 0,42$$

$$P(\{\text{credit}, \text{date}, \text{date}\}|\text{spam}) = 3! * \frac{0,75^1}{1!} * \frac{0,25^2}{2!} = \frac{9}{64} = 0,14$$

$$P(\{\text{date}, \text{date}, \text{date}\}|\text{spam}) = 3! * \frac{0,75^0}{0!} * \frac{0,25^3}{0!} = \frac{1}{64} = 0,016$$

