Naivni Bayesov klasifikator in nomogrami

Naivni bayesov klasifikator je verjetnosti klasifikator. Njegova naivnost je v predpostavki, da so atributi med seboj pogojno neodvisni. Kljub navidez nerealistični predpostavki, se algoritem v praksi pogosto izkaže kot dober, priljubljen pa je predvsem zaradi svoje preprostosti in hitrosti.

Temelji na Bayesovem izreku:

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)},$$

kjer je:

- $P(y|x)$ posteriorna verjetnost razreda y pri danih vrednostih atributa x,
- $P(y)$ apriorna verjetnost razreda y,
- $P(x|y)$ verjetje (*angl. likelihood*) oz. pogojna verjetnost x pri danem y in
- $P(x)$ apriorna verjetnost vrednosti atributa.

V primeru več atributov:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}.$$

Imenovalec v zgornji formuli ni odvisen od razreda in ga za potrebe klasifikatorja zanemarimo, števec pa razpišemo z verižno uporabo pravila:

$$P(A \wedge B|C) = P(A|C)\ P(B|C \wedge A).$$

Pogojno verjetnost v števcu razpišemo takole:

$$\begin{aligned} P(x_1, \ldots, x_n|y) &= P(x_1|y)\ P(x_2, \ldots, x_n|y \wedge x_1) \\ &= P(x_1|y)\ P(x_2|y \wedge x_1)\ P(x_3, \ldots, x_n|y \wedge x_1 \wedge x_2) \\ &= P(x_1|y)\ P(x_2|y \wedge x_1)\ P(x_3|y \wedge x_1 \wedge x_2)\ P(x_4, \ldots, x_n|y \wedge x_1 \wedge x_2) \\ &= \vdots \\ &= P(x_1|y)\ P(x_2|y \wedge x_1) \ldots P(x_n|y \wedge x_1 \wedge \ldots \wedge x_{n-1}) \end{aligned}$$

Privzemimo, da je vsak atribut x_i pogojno neodvisen od x_j, za $i \neq j$, torej $P(x_i|y \wedge x_j) = P(x_i|y)$. Zdaj lahko poenostavimo zadnjo vrstico zgornje izpeljave:

$$P(x_1, \ldots, x_n|y) = P(x_1|y)\ P(x_2|y) \ldots P(x_n|y)$$

Verjetnost razreda pri danih vrednostih atributov je ob predpostavki pogojne neodvisnosti

$$P(y|x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\ldots,x_n)}.$$

Algoritem nov primer klasificira v razred z največjo verjetnostjo:

$$\text{argmax}_{y\in Y} \;\; P(y)\prod_{i=1}^{n} P(x_i|y),$$

kjer je Y množica vrednosti razreda y.

1. naloga

Dani so naslednji učni podatki:

X	Y	Z	R
0	1	a	0
0	1	a	0
1	0	b	0
1	0	b	0
2	1	b	0
1	1	b	1
1	1	b	1
1	0	b	1
2	0	a	1
2	1	a	1

Klasificirajte naslednje primere z naivnim Bayesom. Verjetnosti računajte z relativno frekvenco.

- $X = 1, Y = 1, Z = a$
- $X = 0, Y = 0, Z =?$
- $X = 2, Y = 0, Z = b$

Rešitev

Z naivnim Bayesom klasificiramo tako, da za vsak razred c izračunamo produkt pogojnih verjetnosti danih vrednosti atributov in vse skupaj pomnožimo z verjetnostjo razreda. Na koncu klasificiramo v tisti razred, kjer ima zgornji produkt največjo vrednost:

$$R = argmax_{c\in R}\, P(c)\prod_{i=1}^{n} P(A_i|c)$$

- $X = 1, Y = 1, Z = a$

	$c = 0$	$c = 1$
$P(c)$	1/2	1/2
$P(X = 1\|c)$	2/5	3/5
$P(Y = 1\|c)$	3/5	3/5
$P(Z = a\|c)$	2/5	2/5
$\prod$	6/125	9/125

  Primer $(X = 1, Y = 1, Z = a)$ klasificiramo v razred $c = 1$. Opozorimo, da izračunana vrednost R ni verjetnost $P(c = 1|X = 1, Y = 1, Z = a)$ - to verjetnost lahko izračunamo takole:

$$P(c = 1|X = 1, Y = 1, Z = a) = \frac{9/125}{6/125 + 9/125} = 0.6$$

- $X = 0, Y = 0, Z =?$

	$c = 0$	$c = 1$
$P(c)$	1/2	1/2
$P(X = 0\|c)$	2/5	0
$P(Y = 0\|c)$	2/5	2/5
$P(Z =?\|c)$	/	/
$\prod$	2/25	0

  Primer $(X = 0, Y = 0, Z =?)$ klasificiramo v razred $c = 0$. Ker vrednost atributa Z ni podana, tega atributa pri računanju ne upoštevamo.

- $X = 2, Y = 0, Z = b$

	$c = 0$	$c = 1$
$P(c)$	1/2	1/2
$P(X = 2\|c)$	1/5	2/5
$P(Y = 0\|c)$	2/5	2/5
$P(Z = b\|c)$	3/5	3/5
$\prod$	3/125	6/125

  Primer $(X = 2, Y = 0, Z = b)$ klasificiramo v razred $c = 1$.

V tej nalogi smo pogojne verjetnosti računali z relativno frekvenco, lahko pa bi uporabili Laplaceovo oceno verjetnosti ali m-oceno.

2. naloga

SKUPINA	SPOL		VIŠINA		TEŽA		$\sum$
	M	Ž	< 175 cm	≥ 175 cm	< 65 kg	≥ 65 kg	
A	102	98	96	104	5	195	200
B	53	47	96	4	49	51	100
	155	145	192	108	54	246	300

a) V katero skupino bi uvrstili 198 cm visokega moškega, ki tehta 80kg?

b) V katero skupino bi uvrstili 165 cm visoko žensko, ki tehta 60kg?

c) Kakšna je verjetnost, da ženska pripada skupini A?

Pogojne verjetnosti ocenjujte z m-oceno ($m = 2$), apriorne pa z relativno frekvenco.

Rešitev[1]

a) P(SKUPINA | (M, ≥ 175, ≥ 65))

	SKUPINA = A	SKUPINA = B
apriorna	$P(A) = 200/300 = 2/3 = 0.667$	$P(B) = 100/300 = 1/3 = 0.333$
SPOL	$P(M\|A) = \frac{102+m155/300}{200+m} = 0.51$	$P(M\|B) = \frac{53+m155/300}{100+m} = 0.529$
VIŠINA	$P(\geq 175\|A) = \frac{104+m108/300}{200+m} = 0.518$	$P(\geq 175\|B) = \frac{4+m108/300}{100+m} = 0.046$
TEŽA	$P(\geq 65\|A) = \frac{195+m246/300}{200+m} = 0.973$	$P(\geq 65\|B) = \frac{51+m246/300}{100+m} = 0.516$
$\prod$	$0.667 \cdot 0.51 \cdot 0.518 \cdot 0.973 = 0.171$	$0.333 \cdot 0.529 \cdot 0.046 \cdot 0.516 = 0.004$

Moškega, visokega 198 cm in težkega 80kg uvrstimo v skupino A.

b) P(SKUPINA | (Ž, < 175, < 65))

	SKUPINA = A	SKUPINA = B
apriorna	$P(A) = 200/300 = 2/3 = 0.667$	$P(B) = 100/300 = 1/3 = 0.333$
SPOL	$P(Z\|A) = \frac{98+m145/300}{200+m} = 0.49$	$P(Z\|B) = \frac{47+m145/300}{100+m} = 0.47$
VIŠINA	$P(< 175\|A) = \frac{96+m192/300}{200+m} = 0.48$	$P(< 175\|B) = \frac{96+m192/300}{100+m} = 0.95$
TEŽA	$P(< 65\|A) = \frac{5+m54/300}{200+m} = 0.026$	$P(< 65\|B) = \frac{49+m54/300}{100+m} = 0.48$
$\prod$	$0.667 \cdot 0.49 \cdot 0.48 \cdot 0.026 = 0.004$	$0.333 \cdot 0.47 \cdot 0.95 \cdot 0.48 = 0.07$

Žensko, visoko 165 cm in težko 60kg uvrstimo v skupino B.

c) P(A | Ž) = ?
$P(A) = 2/3$
$P(Z|A) * P(A) = \frac{98}{200}\frac{2}{3} = 0.326$
$P(Z|B) * P(B) = \frac{47}{100}\frac{1}{3} = 0.156$
P(A | Ž) = $\frac{0.326}{0.326+0.156} = 0.676$

[1]Zaradi utesnjenosti v tabelah krajšamo imena in vrednosti atributov.