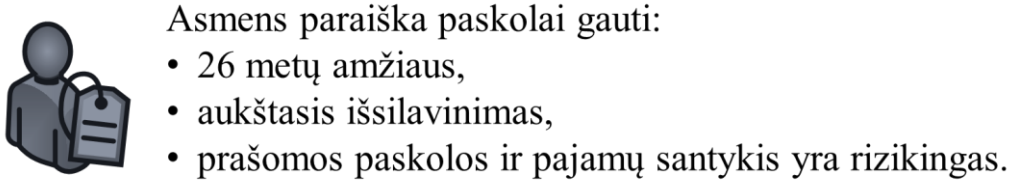
Bayeso klasifikatorius – tikimybėmis grįstas mokymasis

II dalis. Naivusis Bayeso klasifikatorius

Naivusis Bayeso klasifikatorius pagrįstas prielaida, jog duomenų imtyje esantys požymiai vienas nuo kito nepriklauso. Šioje dalyje pavyzdžiu pademonstruosime, kaip kuriamas toks modelis.

Tarkime, kad asmuo nori gauti paskolą. Apie asmenį yra žinoma tokia informacija: jis yra 26 metų amžiaus, turi aukštąjį išsilavinimą, prašomos paskolos dydžio ir asmens pajamų santykis yra rizikingas, t. y. apskaičiuota santykio reikšmė patenka į tokį intervalą, kuris rodo, jog šiuo metu asmens pajamos nėra pakankamos (ribinės) paskolai gauti (1 pav.).



1 pav. *Demonstracinio pavyzdžio atvejis*

Kreditus teikianti įmonė svarsto klausimą, ar išduoti paskolą naujam klientui?

Sprendimui priimti įmonė tikisi gauti rekomendaciją iš modelio. Šiuo atveju – tai naivusis Bayeso klasifikatorius.

Sukaupti istoriniai duomenys atvaizduoti duomenų matricoje – lentelėje (2 pav.). Tai buvusių įmonės klientų sąrašas: ID, amžius (X_1), išsilavinimas (X_2), paskolos dydžio ir pajamų santykis (X_3). Gali būti žinoma ir daugiau informacijos, bet šį kartą Bayeso klasifikatorius bus mokomas pagal šiuos požymius, išskyrus ID. Taip pat jau yra žinoma, ar konkretus klientas grąžino paskolą laiku. Ši informacija naudojama kaip tikslinis kintamasis (Y) – tai yra klasė, įgyjant dvi reikšmės TAIP ir NE.

ID	Amžius	Išsilavinimas	Paskolos dydžio ir pajamų santykis	Paskola grąžinta laiku?
1	24	Aukštasis	Normalus	Taip
2	36	Profesinis	Normalus	Ne
3	49	Profesinis	Normalus	Taip
4	37	Aukštasis	Normalus	Ne
5	28	Aukštasis	Normalus	Taip
6	40	Profesinis	Rizikingas	Ne
7	41	Profesinis	Rizikingas	Taip
8	33	Aukštasis	Rizikingas	Taip
9	51	Aukštasis	Rizikingas	Ne
10	29	Profesinis	Rizikingas	Ne
11	37	Vidurinysis	Normalus	Taip
12	27	Profesinis	Rizikingas	Taip
13	50	Profesinis	Normalus	Taip
14	30	Aukštasis	Rizikingas	Taip

Požymiai $X = (X_1, X_2, X_3)$ Klasės kintamasis Y

2 pav. *Demonstracinio pavyzdžio duomenų imtis*

Prisiminkime iš pirmos temos dalies, jog Bayeso klasifikatoriaus pagrindas yra Bayeso teorema $\mathrm{P}(Y|X) = \frac{\mathrm{P}(X|Y)\cdot\mathrm{P}(Y)}{\mathrm{P}(X)}$, kuri yra išplečiama trims imtyje esantiems požymiams X_1, X_2, X_3 , t. y. modelio prognozės tikimybinis įvertis apskaičiuojamas taip:

$$\mathrm{P}(Y|(X_1, X_2, X_3)) = \frac{\mathrm{P}((X_1, X_2, X_3)|Y) \cdot \mathrm{P}(Y)}{\mathrm{P}(X)}.$$

Aposteriorinei tikimybei $\mathrm{P}(Y|(X_1, X_2, X_3))$ apskaičiuoti reikia įvertinti tris sąlygines tikimybes iš duomenų, t. y. tikėtinumo $\mathrm{P}((X_1, X_2, X_3)|Y)$ ir dvi apriorines tikimybes $\mathrm{P}(X)$ ir $\mathrm{P}(Y)$.

Tarp jų – viena svarbiausių tikimybių yra tikėtinumas $\mathrm{P}((X_1, X_2, X_3)|Y)$. Kadangi požymiai tarpusavyje nepriklausomi (tai naivojo Bayeso klasifikatoriaus prielaida!), tai tikėtinumo formulė gerokai supaprastėja, nes nebetaikomas grandinėlės principas. Todėl tikėtinumas apskaičiuojamas kiekvienam požymiui keliant vienintelę sąlygą, t. y. kokia yra Y klasė:

$$\mathrm{P}((X_1, X_2, X_3)|Y) = \mathrm{P}(X_1|Y) \cdot \mathrm{P}(X_2|Y) \cdot \mathrm{P}(X_3|Y).$$

Kaip žinome, apriorinė požymių X tikimybė $\mathrm{P}(X)$=$\mathrm{P}(X_1, X_2, X_3)$ turi normuojančio daugiklio vaidmenį, kuris neturi įtakos klasifikavimui, ir jį galime ignoruoti siekdami sparčiau atlikti skaičiavimus. Įvertinę naiviojo Bayeso klasifikatoriaus prielaidą ir atsisakę normuojančio daugiklio, gauname gerokai paprastesnę formulę aposteriorinei tikimybei, t. y. modelio prognozei gauti:

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y) \cdot \mathrm{P}(X_2|Y) \cdot \mathrm{P}(X_3|Y) \cdot \mathrm{P}(Y).$$

Taigi, ši formulė tampa pagrindine ir galiausiai apskaičiuosime šias keturias tikimybes naudodami duomenų imtį apie kredito įmonės buvusius klientus panašiu principu, kaip buvo demonstruota pirmoje temos dalyje.

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y) \cdot \mathrm{P}(X_2|Y) \cdot \mathrm{P}(X_3|Y) \cdot \mathrm{P}(Y)$$

Pradėkime nuo tikimybės $\mathrm{P}(Y)$ skaičiavimo. Y – tai tikslinis kintamasis – klasė, rodanti, ar paskola grąžinta laiku. Jai apskaičiuoti duomenų matricoje suskaičiuojame (3 pav.), kiek yra eilučių su reikšme TAIP ir kiek yra eilučių su reikšme NE. Gauname 9 ir 5 eilutes atitinkamai, todėl tikimybės yra $\mathrm{P}(Y = \text{Taip}) = \frac{9}{14}$ ir $\mathrm{P}(Y = \text{Ne}) = \frac{5}{14}$. Gautas reikšmes užfiksuokime lentelėje (4 pav.).

ID	Amžius	Išsilavinimas	Paskolos dydžio ir pajamų santykis	Paskola grąžinta laiku?
1	24	Aukštasis	Normalus	Taip
2	36	Profesinis	Normalus	Ne
3	49	Profesinis	Normalus	Taip
4	37	Aukštasis	Normalus	Ne
5	28	Aukštasis	Normalus	Taip
6	40	Profesinis	Rizikingas	Ne
7	41	Profesinis	Rizikingas	Taip
8	33	Aukštasis	Rizikingas	Taip
9	51	Aukštasis	Rizikingas	Ne
10	29	Profesinis	Rizikingas	Ne
11	37	Vidurinysis	Normalus	Taip
12	27	Profesinis	Rizikingas	Taip
13	50	Profesinis	Normalus	Taip
14	30	Aukštasis	Rizikingas	Taip

Požymiai $X = (X_1, X_2, X_3)$ Klasės kintamasis Y

3 pav. *Demonstracinio pavyzdžio duomenų imtis*

Y	
Taip	Ne
9/14	5/14

4 pav. *Apriorinės tikimybės $P(Y)$ įverčiai*

Skaičiuokime sąlyginę tikimybę $\mathrm{P}(X_1|Y)$, kai X_1 žymi amžių. Galime pastebėti tai, kad šis požymis yra tolydusis, t. y. įgyja beveik nesikartojančias reikšmes. Todėl jei skaičiuotume tikimybes kiekvienai amžiaus reikšmei, tai gautume jas labai mažas arba nulines. Todėl reikia įsivesti tam tikrus intervalus, kitaip tariant, tolydųjį kintamąjį diskretizuoti. Kadangi duomenų turime nedaug, tai sudarykime du intervalus: $[18; 33]$ ir $(33; 99]$. Kadangi klasės kintamasis yra dvireikšmis, tai gauname keturis atvejus ir kiekvienam iš jų apskaičiuojame tikimybes:

$$\mathrm{P}(X_1 \le 33 \mid Y = \text{Taip}) = \frac{5/14}{9/14} = \frac{5}{9}; \quad \mathrm{P}(X_1 > 33 \mid Y = \text{Taip}) = \frac{4/14}{9/14} = \frac{4}{9};$$

$$\mathrm{P}(X_1 \le 33 \mid Y = \text{Ne}) = \frac{1/14}{5/14} = \frac{1}{5}; \quad \mathrm{P}(X_1 > 33 \mid Y = \text{Ne}) = \frac{4/14}{5/14} = \frac{4}{5}.$$

Jas užfiksuojame lentelėje (5 pav.). Galime pasitikrinti, jog gautų tikimybių suma turi būti lygi vienam.

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_1	≤ 33	5/9	1/5
	> 33	4/9	4/5
Suma		1	1

5 pav. *Sąlyginės tikimybės $P(X_1|Y)$ įverčiai*

Skaičiuokime sąlyginę tikimybę $\mathrm{P}(X_2|Y)$, čia X_2 žymi išsilavinimą. Turime tris kategorijas: aukštasis, profesinis, vidurinysis. Kiekvienai šių kategorijų skaičiuojame tikimybes ir dar papildomai nustatome pagal klasę TAIP ir NE. Todėl iš viso turime šešis variantus. Kiekvienam iš jų naudodami duomenų matricą apskaičiuojame tikimybės įvertį:

$$\mathrm{P}(X_2 = \text{Aukštasis} \mid Y = \text{Taip}) = \frac{5/14}{9/14} = \frac{4}{9}; \ \mathrm{P}(X_2 = \text{Aukštasis} \mid Y = \text{Ne}) = \frac{2/14}{5/14} = \frac{2}{5};$$

$$\mathrm{P}(X_2 = \text{Profesinis} \mid Y = \text{Taip}) = \frac{4/14}{9/14} = \frac{4}{9}; \ \mathrm{P}(X_2 = \text{Profesinis} \mid Y = \text{Ne}) = \frac{3/14}{5/14} = \frac{3}{5};$$

$$\mathrm{P}(X_2 = \text{Vidurinysis} \mid Y = \text{Taip}) = \frac{1/14}{9/14} = \frac{1}{9}; \ \mathrm{P}(X_2 = \text{Vidurinysis} \mid Y = \text{Ne}) = \frac{0/14}{5/14} = 0.$$

Taip pat lentelėje reziumuojame apskaičiuotas tikimybes (6 pav.).

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_2	Aukštasis	4/9	2/5
	Profesinis	4/9	3/5
	Vidurinysis	1/9	0/5
Suma		1	1

6 pav. *Sąlyginės tikimybės $P(X_2|Y)$ įverčiai*

Analogiškai apskaičiuojamos sąlyginės tikimybės požymiui X_3 – paskolos dydžio ir pajamų santykis, t. y. $\mathrm{P}(X_3|Y)$. Šis požymis įgyja tik dvi reikšmes, todėl iš viso galimi keturi variantai:

$$\mathrm{P}(X_3 = \text{Normalus} \mid Y = \text{Taip}) = \frac{5/14}{9/14} = \frac{5}{9};\quad \mathrm{P}(X_3 = \text{Rizikingas} \mid Y = \text{Taip}) = \frac{4/14}{9/14} = \frac{4}{9};$$

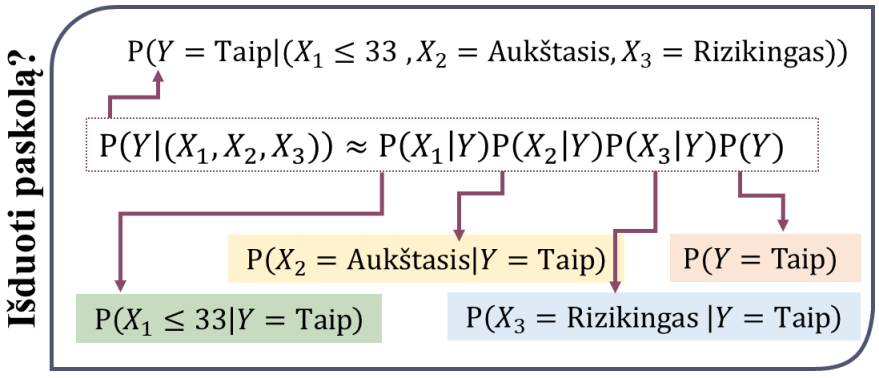
$$\mathrm{P}(X_3 = \text{Normalus} \mid Y = \text{Ne}) = \frac{2/14}{5/14} = \frac{2}{5};\quad \mathrm{P}(X_3 = \text{Rizikingas} \mid Y = \text{Ne}) = \frac{3/14}{5/14} = \frac{3}{5}.$$

Gauti rezultatai apibendrinti lentelėje (7 pav.).

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_3	Normalus	5/9	2/5
	Rizikingas	4/9	3/5
	Suma	1	1

7 pav. *Sąlyginės tikimybės $P(X_3|Y)$ įverčiai*

Visi tarpiniai veiksmai atlikti. Dabar tereikia atsirinkti apskaičiuotas tikimybes ir įrašyti jas į aposteriorinės tikimybės formulę. Prisiminkime, jog asmuo yra 26 metų amžiaus, turi aukštąjį išsilavinimą, o prašomos paskolos ir pajamų santykis yra rizikingas (1 pav.). Kitaip tariant, pirmasis požymis X_1 yra mažiau nei 33, antrasis požymis X_2 yra aukštasis, o trečiasis požymis X_3 yra rizikingas. Panagrinėkime atvejį $Y = \text{Taip}$, t. y. išduoti paskolą. Atitinkamai pagal naujo kliento požymius yra užpildoma supaprastinta naiviojo Bayeso klasifikatoriaus formulė: kairėje formulės pusėje yra aposteriorinė tikimybė – tai modelio prognozė, o dešinėje pusėje – sąlyginės tikimybės kiekvienam kliento požymiui su sąlyga Y =TAIP ir apriorinė tikimybė Y =TAIP (8 pav.).



8 pav. *Aposteriorinės tikimybės skaičiavimas, kai $Y = Taip$*

Paveiksle (8 pav.) spalvotai pažymėtų tikimybių reikšmes tereikia susirasti lentelėse, kurias jau esame gavę atlikdami tarpinius veiksmus. Šias reikšmes atitinkamai sužymėjome spalvomis (9 pav.).

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_1	≤ 33	5/9	1/5
	> 33	4/9	4/5
Suma		1	1

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_3	Normalus	5/9	2/5
	Rizikingas	4/9	3/5
Suma		1	1

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_2	Aukštasis	4/9	2/5
	Profesinis	4/9	3/5
	Vidurinysis	1/9	0/5
Suma		1	1

Y	
Taip	Ne
9/14	5/14

9 pav. *Tikimybiniai įverčiai pagal kliento požymius ir klasę* $Y = Taip$

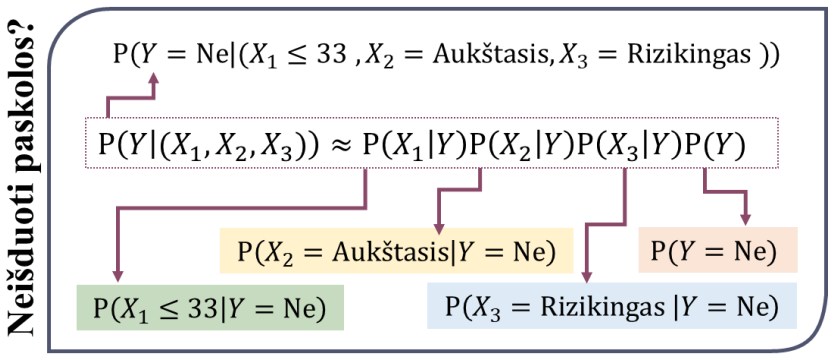
Kaip matome, iš kiekvienos lentelės pasirenkame vieną tikimybę atitinkamai pagal požymio reikšmę ir keliamą sąlygą $Y =$ Taip. Turint tokias paruoštas lenteles, keičiant kliento požymius ar keliamas sąlygas, galima greitai tikimybiškai įvertinti bet kurį variantą,.

Sustatome šias reikšmes į naiviojo Bayeso klasifikatoriaus formulę ir apskaičiuojame tikimybės įvertį, kai $Y =$ Taip:

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y)\mathrm{P}(X_2|Y)\mathrm{P}(X_3|Y)\mathrm{P}(Y) =$$

$$= \frac{5}{9} \cdot \frac{4}{9} \cdot \frac{4}{9} \cdot \frac{9}{14} \approx 0{,}0705.$$

Dabar gaukime tikimybės įvertį, jei norėtume įvertinti neišduotos paskolos atvejį. Kadangi kliento požymių nekeičiame, tai visos reikšmės išlieka tos pačios, tačiau Y klasės reikšmė vietoj TAIP keičiama į NE (10 pav.).



10 pav. *Aposteriorinės tikimybės skaičiavimas, kai* $Y = Ne$

Vėlgi iš kiekvienos lentelės atsirankame po vieną reikšmę ir įstatome į naiviojo Bayeso klasifikatoriaus formulę. Atsirinkę atitinkamai pagal spalvas (11 pav.), gauname, jog tikimybės įvertis šiek tiek mažesnis nei dvi šimtosios:

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y)\mathrm{P}(X_2|Y)\mathrm{P}(X_3|Y)\mathrm{P}(Y) = \frac{1}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{5}{14} \approx 0{,}0171.$$

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_1	≤ 33	5/9	1/5
	> 33	4/9	4/5
Suma		1	1

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_3	Normalus	5/9	2/5
	Rizikingas	4/9	3/5
Suma		1	1

Sąlyginių tikimybių lentelė		Y	
		Taip	Ne
X_2	Aukštasis	4/9	2/5
	Profesinis	4/9	3/5
	Vidurinysis	1/9	0/5
Suma		1	1

Y	
Taip	Ne
9/14	5/14

11 pav. *Tikimybiniai įverčiai pagal kliento požymius ir klasę* $Y = Ne$

Palyginame abu atvejus (12 pav.).

Išduoti paskolą?

$\mathrm{P}(Y = \text{Taip}|(X_1 \leq 33\ , X_2 = \text{Aukštasis}, X_3 = \text{Rizikingas}))$

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y)\mathrm{P}(X_2|Y)\mathrm{P}(X_3|Y)\mathrm{P}(Y) = $$

$$= \frac{5}{9}\cdot\frac{4}{9}\cdot\frac{4}{9}\cdot\frac{9}{14} \approx 0{,}0705.$$

Neišduoti paskolos?

$\mathrm{P}(Y = \text{Ne}|(X_1 \leq 33\ , X_2 = \text{Aukštasis}, X_3 = \text{Rizikingas}))$

$$\mathrm{P}(Y|(X_1, X_2, X_3)) \approx \mathrm{P}(X_1|Y)\mathrm{P}(X_2|Y)\mathrm{P}(X_3|Y)\mathrm{P}(Y) = $$

$$= \frac{1}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{5}{14} \approx 0{,}0171.$$

12 pav. Apskaičiuotos prognozių tikimybės

Atvejis su didžiausia tikimybe – naiviojo Bayeso klasifikatoriaus siūlomas sprendimas, t. y. paskolą išduoti (12 pav.).

Programinėse priemonėse šios tikimybės gan dažnai standartizuojamos, t. y. kiekviena tikimybė atskirai padalijama iš visų apskaičiuotų tikimybių sumos (13 pav.). Taip gaunamos prognozės , kurių tikimybių suma lygi vienam. Tačiau toks standartizavimas patiės sprendimo su maksimalia reikšme nepakeis.

Išduoti paskolą?

$$\frac{0{,}0705}{0{,}0705+0{,}0171} \approx 0{,}8048.$$

Neišduoti paskolos?

$$\frac{0{,}0171}{0{,}0705+0{,}0171} \approx 0{,}1952.$$

13 pav. *Standartizuotos apskaičiuotų prognozių tikimybės*

Tad šiuo atveju modelio rekomendacija yra išduoti paskolą nepaisant to, jog prašomos paskolos ir pajamų santykis yra rizikingas. Norime atkreipti dėmesį į tai, jog modelis pasiūlė sprendimą pagal tai, kokia duomenų imtis buvo panaudojama jam mokyti. Turbūt turimoje kredito įmonės patirtyje rizikingas paskolos ir pajamų santykis neturėjo didelės įtakos paskolos grąžinimui laiku. Tačiau galutinis sprendimas yra priimamas įvertinant daugelį kitų faktorių.