Bayeso klasifikatorius – tikimybėmis grįstas mokymasis

I dalis. Bayeso klasifikatoriaus mokymas

Bayeso klasifikatorius yra mokomas pagal tikimybes, išskaičiuotas iš duomenų imties. Šis modelis yra išvestas iš ***Bayeso teoremos***, kuri neišvengiamai yra kiekvieno tikimybių teorijos kurso dalis.

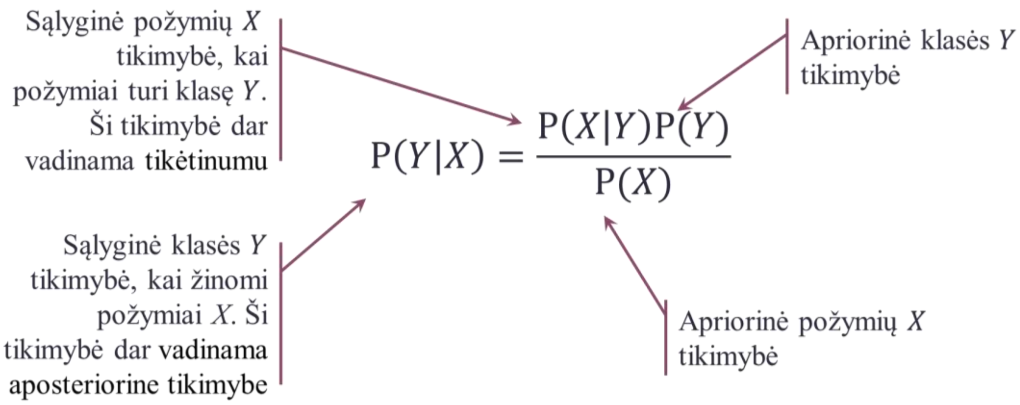
Sakykime, kad klasifikatoriaus mokymui yra paruošta duomenų imtis (1 pav.), kuri atvaizduota kaip matrica. Ją sudaro aibė X – tai kintamieji, nusakantys stebėjimų požymius. Pavyzdžiui, jei mūsų tyrimo objektas yra reklama socialiniame tinkle, tai požymiais galėtų būti reklamos rodymo trukmė, dažnis, tam tikros vaizdavimo charakteristikos ir pan. Bayeso klasifikatoriui išmokyti būtinai duomenų imtyje turi būti ir tikslinis kintamasis, kuris žymi stebėjimo klasę. Reklamos pavyzdžiui, tikslinis kintamasis galėtų būti sukuriamas iš istorinių duomenų, kiek dažnai reklama buvo peržiūrėta, t. y. populiari ar nepopuliari.

X_1	X_2	X_3	...	X_d	Y
x_1^1	x_1^2	x_1^3	...	x_1^d	y_1
x_2^1	x_2^2	x_2^3	...	x_2^d	y_2
...	...	...	...	...	...
x_n^1	x_n^2	x_n^3	...	x_n^d	y_n

X – kintamieji (požymiai) Y – tikslinis kintamasis - klasė

1 pav. *Duomenų matricos lentelė*

Iš suformuotos duomenų imties yra skaičiuojami tam tikrų tikimybinių įverčiai pagal Bayeso teoremą, kurios klasikinė formulė X ir Y įvykiams yra



čia $P(X)$ – apriorinė požymių X tikimybė, $P(Y)$ – apriorinė klasės Y tikimybė, $\mathrm{P}(X|Y)$ – sąlyginė požymių X tikimybė, kai požymiai turi klasę Y, $\mathrm{P}(Y|X)$ – sąlyginė klasės Y tikimybė, kai žinomi požymiai X. $\mathrm{P}(X|Y)$ tikimybė dar vadinama tikėtinumu, o $\mathrm{P}(Y|X)$ tikimybė dar žinoma kaip aposteriorinė tikimybė.

Kadangi modelis-klasifikatorius mokomas pagal daug požymių $X = (X_1, X_2, \ldots, X_d)$, tai Bayeso teoremos formulė yra modifikuojama įtraukiant visus šiuos požymius. Tokiu būdu gaunama ***apibendrintoji Bayeso teoremos formulė:***

$$\mathrm{P}(Y|(X_1, X_2, \ \dots, \ X_d)) = \frac{\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y\big) \cdot \mathrm{P}(Y)}{\mathrm{P}(X_1, X_2, \ \dots, \ X_d)}.$$

Klasifikatoriaus mokymo metu gauta apibendrintoji Bayeso teoremos formulė yra taikoma kiekvienai Y klasei $k \in \{k_1, \ k_2, \ \dots, k_K\}$ atskirai.

Gan dažnai klasifikavimo uždaviniuose klasės kintamasis yra dvireikšmis arba tiesiog sakoma, yra dvinaris ar binaris, t. y. įgyja dvi reikšmes. Pavyzdžiui, jei yra dvi klasės TAIP ir NE, tai taikant apibendrintąją Bayeso teoremos formulę turi būti apskaičiuojamos dvi aposteriorinės tikimybės atskirai klasei TAIP ir atskirai klasei NE:

$$\mathrm{P}(Y = \mathrm{TAIP}|(X_1, X_2, \ \dots, \ X_d)) = \frac{\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{TAIP}\big)\mathrm{P}(Y = \mathrm{TAIP})}{\mathrm{P}(X_1, X_2, \ \dots, \ X_d)},$$

$$\mathrm{P}(Y = \mathrm{NE}|(X_1, X_2, \ \dots, \ X_d)) = \frac{\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{NE}\big)\mathrm{P}(Y = \mathrm{NE})}{\mathrm{P}(X_1, X_2, \ \dots, \ X_d)}.$$

Panagrinėkime gautos formulės dešinėje lygybės pusėje esančias tikimybes. Pradėkime nuo ***tikėtinumo*** $\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{k}\big)$. <u>Laikoma, kad objektų požymiai yra tarpusavyje priklausomi.</u> Todėl tikėtinumas apskaičiuojamas ***grandinėlės principu***, t. y. tikėtinumas esant klasei k yra gaunamas sudauginant sąlygines tikimybes, apskaičiuotas kiekvienam iš požymių atskirai, kai keliama sąlyga vis griežtėja įtraukiant daugiau reikalavimų, t. y. pradedama tik nuo sąlygos $Y = \mathrm{k}$, o baigiama nariu su sąlyga $Y = \mathrm{k}$ ir įtraukiant visus požymius X, išskyrus vieną (šiuo atveju X_d):

$$\begin{aligned}\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{k}\big) &= \\ &= \mathrm{P}(X_1|Y = \mathrm{k}) \cdot \mathrm{P}(X_2|(X_1, Y = \mathrm{k})) \cdot \mathrm{P}(X_3|(X_2, \ X_1, Y = \mathrm{k})) \cdot \dots \\ &\cdot \mathrm{P}(X_d|(X_{d-1}, \ \dots, X_2, X_1, Y = \mathrm{k})).\end{aligned}$$

Dabar panagrinėkime ***apriorinę klasės Y tikimybę*** $\mathrm{P}(Y = \mathrm{k})$. Sprendimo metu šios tikimybės apskaičiavimas yra visiškai nesudėtingas ir bus pademonstruota pavyzdyje.

Apriorinė požymių X tikimybė $\mathrm{P}(X_1, X_2, \ \dots, \ X_d)$ turi normuojančio daliklio paskirtį ir klasifikavimo rezultatams jokios įtakos neturi. Todėl eksperimentuose gali būti taikoma ***supaprastinta apibendrintoji Bayeso teoremos formulė:***

$$\mathrm{P}(Y = \mathrm{k}|(X_1, X_2, \ \dots, \ X_d)) = \eta \cdot \mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{k}\big)\mathrm{P}(Y = \mathrm{k}),$$

čia η – normavimo daugiklis, kuris gali būti ignoruojamas.

Bayeso klasifikatoriaus siūlomas sprendimas randamas maksimizuojant posteriorinę tikimybę:

$$\underset{k \in \{k_1, \ k_2, \ \dots, k_K\}}{\mathrm{argmax}} \ \mathrm{P}(Y = \mathrm{k}|(X_1, X_2, \ \dots, \ X_d)) = \underset{k \in \{k_1, \ k_2, \ \dots, k_K\}}{\mathrm{argmax}} \ \frac{\mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{k}\big)\mathrm{P}(Y = \mathrm{k})}{\mathrm{P}(X_1, X_2, \ \dots, \ X_d)}.$$

Eliminavus normavimo tikimybes, formulė tampa dar paprastesnė:

$$\underset{k \in \{k_1, \ k_2, \ \dots, k_K\}}{\mathrm{argmax}} \ \mathrm{P}(Y = \mathrm{k}|(X_1, X_2, \ \dots, \ X_d)) = \underset{k \in \{k_1, \ k_2, \ \dots, k_K\}}{\mathrm{argmax}} \ \eta \cdot \mathrm{P}\big((X_1, X_2, \ \dots, \ X_d)|Y = \mathrm{k}\big)\mathrm{P}(Y = \mathrm{k}).$$

Pereikime prie pavyzdžio. Tarkime, kad knyga yra vertinama skaitytojo atsiliepimu ☺ (teigiamas) arba ☹ (neigiamas). Kiekviena knyga taip pat yra vertinama pagal tris kriterijus: (a) ar knygos turinys informatyvus? (b) ar knygos turinys aiškus? (c) ar knygos turinys originalus? Į šiuos tris klausimas atsakoma TAIP arba NE. Paprastumo dėlei, tarkime, kad yra 10 knygų su gautais skaitytojų atsiliepimais ir trijų kriterijų įvertinamais (2 pav.).

ID	Informatyvumas?	Aiškumas?	Originalumas?	Atsiliepimas
1	NE	TAIP	NE	☹
2	TAIP	NE	TAIP	☺
3	TAIP	NE	TAIP	☹
4	TAIP	NE	TAIP	☹
5	TAIP	NE	TAIP	☺
6	TAIP	TAIP	NE	☹
7	NE	TAIP	NE	☹
8	TAIP	NE	TAIP	☹
9	TAIP	NE	TAIP	☹
10	NE	TAIP	NE	☺
	TAIP	NE	TAIP	?

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

2 pav. *Pavyzdžio duomenų imtis*

Keliame klausimą, koks bus naujos knygos atsiliepimas iš skaitytojų, jei informatyvumas yra TAIP, aiškumas yra NE, originalumas yra TAIP ?

Pirmiausiai, formaliai užrašome aposteriorinę tikimybę:

$$\mathrm{P}(Y = ☺ \,|(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP}))$$

klausdami, kiek tikėtina, jog atsiliepimas bus teigiamas, jei informatyvumas (X_1) yra TAIP, aiškumas (X_2) yra NE, originalumas (X_3) yra TAIP.

Aposteriorinei tikimybei apskaičiuoti pritaikome apibendrintąją Bayeso teoremos formulę požymiams X priskirdami reikšmes pagal tai, kokiais kriterijais įvertina nauja knyga, o Y priskiriame teigiamą atsiliepimą pagal suformuluotą klausimą, jog atsiliepimas bus teigiamas:

$$P\big(Y = ☺ \,\big|(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})\big) =$$

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

Pradėkime nuo tikėtinumo $\mathrm{P}((X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})|Y = ☺)$ apskaičiavimo pagal grandinėlės principą:

$$\mathrm{P}((X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})|Y = ☺) =$$

$$\mathrm{P}(X_1 = \mathrm{TAIP}|Y = ☺) \cdot \mathrm{P}(X_2 = \mathrm{NE}|(X_1 = \mathrm{TAIP}, Y = ☺)) \cdot$$

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

Kiekviena iš formulėje esančių sąlyginių tikimybių yra apskaičiuojama pagal matricoje turimus istorinius duomenis. Taigi sąlyginė tikimybė $\mathrm{P}(X_1 = \mathrm{TAIP}|Y = ☺)$ yra gaunama suskaičiuojant, kiek yra eilučių, kai informatyvumas yra TAIP ir atsiliepimas yra teigiamas. Iš lentelės (2 pav.) matyti, jog tokių eilučių yra 2 iš 10. Taip pat suskaičiuojame, kiek iš viso yra eilučių su teigiamas atsiliepimais. Turime tik 3 iš 10. Surašome į formulę ir gauname:

$$\mathrm{P}(X_1 = \mathrm{TAIP} | Y = ☺) = \frac{\mathrm{P}(X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(Y = ☺)} = \frac{2}{3}.$$

Skaičiuokime kitą tikimybę, t. y. griežtinam sąlygą ir skaičiuojam sąlyginę tikimybę

$$\mathrm{P}\big(X_2 = \mathrm{NE} \big| (X_1 = \mathrm{TAIP}, Y = ☺)\big) = \frac{\mathrm{P}(X_2 = \mathrm{NE},\ X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(X_1 = \mathrm{TAIP}, Y = ☺)}.$$

Naudodamiesi šia formule, tuo pačiu principu suskaičiuojame duomenų matricoje eilutes, tenkinančias duotas sąlygas. Atvejui $(X_2 = \mathrm{NE},\ X_1 = \mathrm{TAIP}, Y = ☺)$ randame dvi eilutes, o atvejui $(X_1 = \mathrm{TAIP}, Y = ☺)$ taip pat randame dvi eilutes. Todėl apskaičiuota tikimybė yra

$$\mathrm{P}(X_2 = \mathrm{NE} | (X_1 = \mathrm{TAIP}, Y = ☺)) = \frac{\mathrm{P}(X_2 = \mathrm{NE},\ X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(X_1 = \mathrm{TAIP}, Y = ☺)} = \frac{2}{2} = 1.$$

Analogiškai apskaičiuojamas paskutinysis tikėtinumo formulėje esantis narys $\mathrm{P}(X_3 = \mathrm{TAIP} | (X_2 = \mathrm{NE}, X_1 = \mathrm{TAIP}, Y = ☺))$ ir gaunamas tikimybė lygi vienam, t. y.

$$\mathrm{P}(X_3 = \mathrm{TAIP} | (X_2 = \mathrm{NE}, X_1 = \mathrm{TAIP}, Y = ☺)) = \frac{\mathrm{P}(X_3 = \mathrm{TAIP}, X_2 = \mathrm{NE},\ X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(X_2 = \mathrm{NE}, X_1 = \mathrm{TAIP}, Y = ☺)} = \frac{2}{2} = 1.$$

Sudauginame ką tik apskaičiuotus visus tris narius ir galutinis tikėtinumo įvertis yra dvi trečiosios

$$\mathrm{P}((X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP}) | Y = ☺) = \frac{2}{3} \cdot 1 \cdot 1 = \frac{2}{3}.$$

Sudėtingiausia sąlyginė tikimybė apskaičiuota. Gautą rezultatą įrašome į apibendrintąją Bayeso teoremos formulę:

$$P\big(Y = ☺ \,\big| (X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})\big) = \frac{\frac{2}{3} \cdot \mathrm{P}(Y = ☺)}{\mathrm{P}(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})}.$$

Pereikime prie dviejų likusių apriorinių tikimybių skaičiavimo. Tikimybė, jog atsiliepimas yra teigiamas, yra $\mathrm{P}(Y = ☺) = \frac{3}{10}$, nes yra 3 eilutės su teigiamais atsiliepimais iš 10. Gautą rezultatą įrašome į apibendrintąją Bayeso teoremos formulę:

$$\mathrm{P}\big(Y = ☺ \,\big| (X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})\big) = \frac{\frac{2}{3} \cdot \frac{3}{10}}{\mathrm{P}(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})}.$$

Apskaičiuojame likusią tikimybę $\mathrm{P}(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})$. Šiuo atveju reikia išrinkti visas eilutes, kuriose informatyvumas yra TAIP, aiškumas yra NE, originalumas yra TAIP. Tokių eilučių turime net 6, ir todėl rezultatas yra $\frac{6}{10}$. Taip pat įrašome į apibendrintąją Bayeso teoremos formulę:

$$\mathrm{P}\big(Y = ☺ \,\big| (X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})\big) = \frac{\frac{2}{3} \cdot \frac{3}{10}}{\frac{6}{10}} = \frac{1}{3}.$$

Taigi gavome, jog skaitytojai knygą įvertins teigiamais su tikimybe $\frac{1}{3}$, kai informatyvumas TAIP, aiškumas NE, originalumas TAIP.

Smulkiau nedetalizuosime, bet lygiai tokiu pačiu principu apskaičiuojamos visos tikimybės, tik dabar atsiliepimas yra neigiamas. Apskaičiuota tikimybė, jog skaitytojai knygą įvertins neigiamai, yra

$$\mathrm{P}\big(Y = ☹ \,\big|(X_1 = \mathrm{TAIP}, X_2 = \mathrm{NE}, X_3 = \mathrm{TAIP})\big) = \frac{\frac{4}{7} \cdot \frac{7}{10}}{\frac{6}{10}} = \frac{2}{3}.$$

Galutinis Bayeso klasifikatoriaus siūlomas sprendimas yra randamas maksimizavimo būdu. Iš visų gautų tikimybių kiekvienai klasei atskirai randama didžiausia reikšmė. Kadangi $\frac{1}{3} < \frac{2}{3}$, tai Bayeso klasifikatorius prognozuoja neigiamą atsiliepimą.

Pratęskime pavyzdį. Kokia būtų Bayeso klasifikatoriaus prognozė, jei informatyvumas yra TAIP, aiškumas yra TAIP, o originalumas yra NE? Vėlgi formalizuojame klausimą užrašydami aposteriorinės tikimybės formule $\mathrm{P}(Y = ☺ \,|(X_1 = \mathrm{TAIP}, X_2 = \mathrm{TAIP}, X_3 = \mathrm{NE}))$. Atitinkamai užrašome apibendrintos Bayeso teoremos formulę:

$$\mathrm{P}\big(Y = ☺ \,\big|(X_1 = \mathrm{TAIP}, X_2 = \mathrm{TAIP}, X_3 = \mathrm{NE})\big) =$$
$$= \frac{\mathrm{P}((X_1 = \mathrm{TAIP}, X_2 = \mathrm{TAIP}, X_3 = \mathrm{NE})|Y = ☺) \cdot \mathrm{P}(Y = ☺)}{\mathrm{P}(X_1 = \mathrm{TAIP}, X_2 = \mathrm{TAIP}, X_3 = \mathrm{NE})}.$$

Pradėkime nuo sudėtingiausios tikimybės – tikėtinumo $\mathrm{P}((X_1 = \mathrm{TAIP}, X_2 = \mathrm{TAIP}, X_3 = \mathrm{NE})|Y = ☺)$ skaičiavimo. Skleidžiame tikėtinumo formulę grandinėlės principu:

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

Vėlgi sandaugoje turime tris dauginamuosius. Pirmasis jų $\mathrm{P}(X_1 = \mathrm{TAIP}|Y = ☺)$ – sąlyginė tikimybė, jog informatyvumas (X_1) yra TAIP su sąlyga, jog atsiliepimas X_1 yra teigiamas. Suskaičiavę eilutes duomenų matricoje analogiškai kaip buvo pademonstruota praeitame pavyzdyje, gauname

$$\mathrm{P}(X_1 = \mathrm{TAIP}|Y = ☺) = \frac{\mathrm{P}(X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(Y = ☺)} = \frac{2}{3}.$$

Skaičiuodami antrąją tikėtinumo formulėje esančią tikimybę gauname 0, nes neradome nei vienos tokios eilutės pagal skaitiklyje esančią sąlygą $(X_2 = \mathrm{TAIP},\ X_1 = \mathrm{TAIP}, Y = ☺)$, t.y.

$$\mathrm{P}(X_2 = \mathrm{TAIP}|(X_1 = \mathrm{TAIP}, Y = ☺)) = \frac{\mathrm{P}(X_2 = \mathrm{TAIP},\ X_1 = \mathrm{TAIP}, Y = ☺)}{\mathrm{P}(X_1 = \mathrm{TAIP}, Y = ☺)} = \frac{0}{2} = 0.$$

Skaičiuojant trečiąją paskutiniąją tikėtinumo formulėje esančią tikimybę gauname neapibrėžtumą – dalybą iš nulio:

$$\mathrm{P}(X_3 = \mathrm{NE}|(X_2 = \mathrm{TAIP}, X_1 = \mathrm{TAIP}, Y = ☺)) =$$

$$\frac{\mathrm{P}(X_3=\mathrm{NE}, X_2=\mathrm{TAIP}, X_1=\mathrm{TAIP}, Y=☺)}{\mathrm{P}(X_2=\mathrm{TAIP}, X_1=\mathrm{TAIP}, Y=☺)} = \frac{0}{0}.$$

ir veiksmas yra negalimas. Tai vadinamas ***„nulinės tikimybės“ problema***. Bet ar tai reiškia, jog Bayeso klasifikatoriaus sudaryti negalime? Atsakymas – galima, įvedant tam tikras prielaidas.

Apibendrinkime pavyzdį:

- pateiktame pavyzdyje buvo pademonstruota, kaip skaičiuojant gali būti gaunamos sąlyginių tikimybių reikšmės lygios nuliui.
- Kadangi skaičiuojant tikėtinumą šios sąlyginės tikimybės yra dauginamos, tai aposteriorinė (modelio prognozės) tikimybė tampa lygi nuliui, jei nors viena sąlyginės tikimybės reikšmė yra nulinė. Šiai problemai spręsti yra galimi keli būdai.

Pirmasis būdas – vadinamasis ***adityvusis glodinimas*** (angl. *Additive Smoothing*). Taikomas būtent tuomet, kai susiduriama su „nulinės tikimybės“ problema. Adityviojo glodinimo metodas taikytinas kategoriniams duomenims yra vadinamas ***Laplaso glodinimu*** (angl. *Laplace Smoothing)* arba bendriau Lidstono glodinimu (angl. *Lidstone Smoothing*), kurio idėja įvesti **pseudotikimybes vietoj nulinių reikšmių**. Tokia modifikacija yra realizuota daugelyje programinių įrankių, siūlančių Bayeso modelį.

Kitas būdas – priimti prielaidą, jog ***požymiai yra tarpusavyje nepriklausomi***. Požymių nepriklausomumo prielaida leidžia išvengti griežtų ir daug ribojančių sąlygų tikrinimo skaičiuojant sąlygines tikimybes. Šiuo atveju Bayeso klasifikatorius yra vadinamas ***naiviuoju*** (angl. *naive*). Naiviojo Bayeso klasifikatoriaus mokymo demonstracinis pavyzdys pateiktas kitoje dalyje.