Pradinė duomenų analizė ir jų transformacijos

Kategorinių kintamųjų transformavimas

Matematinis aparatas labiausiai "pritaikytas" dirbti su skaičiais. Todėl, turėdami kategorinį kintamąjį X, įgyjantį k $(k \geq 1)$ skirtingų reikšmių, galime galvoti, kad

$$X \in \{1, 2, \ldots, k\}.$$

Dvi reikšmes įgyjantį kintamąjį galima koduoti binariniu kodu : 0 ir 1 .

Kartais k -reikšmis kategorinis kintamasis X transformuojamas į vadinamąjį fiktyvų (angl. dummy) kintamąjį X', kurio reikšmės yra k - ženkliai binariniai žodžiai, sudaryti iš $k-1$ nulio ir vieno vieneto. Pavyzdžiui

X_1	X_1'
saulėta	100
debesuota	010
lietinga	001

Skaitinio atributo reikšmių standartizavimas

Tarkime, kad $x_1, x_2, \ldots, x_n$ yra imtyje stebėtos skaitinio kintamojo X reikšmės.

Skaitinio atributo reikšmių standartizavimas

Tarkime, kad $x_1, x_2, \ldots, x_n$ yra imtyje stebėtos skaitinio kintamojo X reikšmės.

Dešimtainis standartizavimas.

Apibrėžkime sveiką neneigiamą skaičių K:

$$K = \min \left\{ k \,:\, k \geq 0\,,\; 10^{-k} \max_{1 \leq i \leq n} |x_i| \leq 1 \right\}.$$

Tada

$$x_i' = x_i\, 10^{-K} \in [-1;\, 1]\,.$$

Skaitinio atributo reikšmių standartizavimas

Tarkime, kad $x_1, x_2, \ldots, x_n$ yra imtyje stebėtos skaitinio kintamojo X reikšmės.

Min-max standartizavimas.

Tegul

$$m = \min_{1 \le i \le n} x_i \text{ ir } M = \max_{1 \le i \le n} x_i \,.$$

Tada

$$x_i' = \frac{x_i - m}{M - m} \in [0;\, 1] \,,$$

arba

$$x_i' = \frac{2x_i - M - m}{M - m} \in [-1;\, 1] \,.$$

Skaitinio atributo reikšmių standartizavimas

Tarkime, kad $x_1, x_2, \ldots, x_n$ yra imtyje stebėtos skaitinio kintamojo X reikšmės.

***z* - standartizavimas**. Tai labiausiai paplitęs standartizavimas, kuris reiškia vadinamųjų z reikšmių skaičiavimą. Tarkime, kad

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

Tuomet z reikšmės skaičiuojamos pagal formulę

$$z_i = \frac{x_i - \bar{x}}{s}\,.$$

Pastebėsime, kad z reikšmių vidurkis visada lygus 0, o standartinis nuokrypis visada lygus 1 :
$\bar{z} = 0,\ s_z = 1\,.$

Tolydžiųjų kintamųjų diskretizavimas

Tarkime, kad X yra vienas iš tolydžių nepriklausomų kintamųjų, o Y - priklausomas kategorinis kintamasis, turintis k kategorijų. Galime laikyti, kad $Y \in \{1, 2, \ldots, k\}$, o imties įrašai yra $\{(x_i, y_i)\ ,\ i = 1, 2, \ldots, N\}$. Diskretizuosime kintamąjį X , transformuodami jį į kategorinį kintamąjį, turintį ne daugiau kaip k_X kategorijų $(2 \leq k_X < N)$.

Pirmiausiai visą kintamojo X leistinų reikšmių intervalą skaidome į nesikertančius intervalus taip, kad skirtingos reikšmės x_i priklausytų skirtingiems intervalams. Tarkime, kad I_1 ir I_2 yra bet kurie du gretimi dalinimo intervalai. Toliau, pasirinkę reikšmingumo lygmenį $0 < \alpha < 1$, tikriname ar statistiškai reikšmingas skirtumas tarp kintamojo Y skirstinių, kai X priklauso intervalams I_1 ir I_2. Tikrinimui naudosime χ^2 kriterijų.

χ^2 kriterijus

Porinė dažnių lentelė

$X\backslash Y$	1	2	...	k	Σ
I_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
I_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

Kriterijaus funkcija (kitaip dar vadinama kriterijaus statistika) yra

$$\chi^2 = \sum_{m=1}^{2}\sum_{j=1}^{k} \frac{(n_{mj} - E_{mj})^2}{E_{mj}}\ , \qquad E_{mj} = \frac{n_{m.}\, n_{.j}}{n}\ .$$

E_{mj} - tikėtinieji dažniai. Jei dažnių lentelėje kuris nors iš $n_{m.}$ ar $n_{.j}$ lygus 0, laikysime, kad E_{mj} lygus kokiam nors mažam teigiamam skaičiui, tarkime $E_{mj} = 0,1$.
Laisvės laipsnių skaičius lygus $(2-1)(k-1) = k-1$.

χ^2 kriterijus

Jei gautoji χ^2 reikšmė yra mažesnė už χ^2 skirstinio su $k-1$ laisvės laipsniu α lygmens kritinę reikšmę $\chi^2_\alpha(k-1)$, tai galime tvirtinti, kad kintamojo Y skirstinys nepriklauso nuo to kuriam iš intervalų I_1 ar I_2 priklauso X reikšmės. Todėl intervalus I_1 ir I_2 galime sujungti ir vienetu sumažinti turimų intervalų skaičių.

Intervalų jungimo procedūra

Tegul I_1 ir I_2 yra gretimi intervalai, kuriems χ^2 reikšmė buvo mažiausia, tarkime $\chi^2 = u$. Toliau nagrinėjame du galimus atvejus.

a) Jei
$u < \chi^2_\alpha(k-1)$,
tai intervalus I_1 ir I_2 sujungiame ir vėl skaičiuojame χ^2 reikšmes tik jau mažesniam intervalų skaičiui. Tai kartojame tol, kol intervalų skaičius taps ne didesnis už pageidaujamą kategorijų kiekį k_X. Pastebėsime, kad kiekviename žingsnyje (aišku, išskyrus pirmąjį) reikės perskaičiuoti ne daugiau, kaip dvi χ^2 reikšmes.

b) Kuriame nors žingsnyje gauname, kad
$u \geq \chi^2_\alpha(k-1)$
ir turimas intervalų skaičius vis dar didesnis už k_X. Tada, sumažinę reikšmingumo lygmenį α , padidiname $\chi^2_\alpha(k-1)$ ir tęsiame procedūrą (žr. punktą a))

Trūkstamosios reikšmės

Kartais kai kuriuose įrašuose nėra vieno ar kelių kintamųjų reikšmių.
Tai vadinamosios trūkstamosios reikšmės (praleistieji stebėjimai).
Jei tokių įrašų dalis visoje imtyje nedidelė, galima juos paprasčiausiai išmesti arba, atliekant skaičiavimus, ignoruoti.
Tačiau kartais ir trūkstamos reikšmės (arba jų skaičius) yra informatyvios!

Trūkstamosios reikšmės

Kartais kai kuriuose įrašuose nėra vieno ar kelių kintamųjų reikšmių.
Tai vadinamosios trūkstamosios reikšmės (praleistieji stebėjimai).
Jei tokių įrašų dalis visoje imtyje nedidelė, galima juos paprasčiausiai išmesti arba, atliekant skaičiavimus, ignoruoti.
Tačiau kartais ir trūkstamos reikšmės (arba jų skaičius) yra informatyvios!
Galimi ir tokie variantai

1. Visas trūkstamas vieno kintamojo reikšmes keičiame viena konstanta. Šios konstantos reikšmė labai priklauso tiek nuo tiriamų duomenų prigimties, tiek nuo tyrimo metodų.
2. Visas trūkstamas vieno skaitinio kintamojo reikšmes keičiame jo vidurkiu.
3. Jei įrašai jau yra kokiu nors būdu klasifikuoti, tai trūkstamas vieno skaitinio kintamojo reikšmes keičiame jo vidutine reikšme toje klasėje, kuriai priklauso nagrinėjamas įrašas.

Išskirtys

Išskirtys - tai mažai tikėtinos arba net visiškai neįmanomos kai kurių kintamųjų reikšmės. Pavyzdžiui, nutarę visų skaitinių kintamųjų trūkstamas reikšmes pakeisti 0, galime "atrasti", kad kai kurie piliečiai visai nieko nesveria arba jų svoris normalus, bet ūgis lygus 0.

- Kategorinio kintamojo išskirtimi paprastai laikoma reikšmė, nepriklausanti jo leistinų reikšmių aibei.
- Dažniausiai išskirtimis laikomos tos skaitinio kintamojo reikšmės x_i, kurių z reikšmės absoliučiuoju didumu didesnės už 3 : $|z_i| > 3$. Toks apibrėžimas grindžiamas tikimybių teorijoje gerai žinoma Čebyšovo nelygybe. Beje, kai kurie autoriai siūlo laikyti "įtartinomis" jau tas reikšmes x_i, kurioms $|z_i| > 2$.

Objektų artumo matai

Apibrėžimas. *Atstumu tarp dviejų objektų **a** ir **b** vadinsime neneigiamą skaitinę funkciją $d(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek skirtingi **a** ir **b**. Objektai yra panašūs, jei atstumas tarp jų mažas.*

Apibrėžimas. *Dviejų objektų **a** ir **b** panašumo koeficientu vadinsime skaitinę funkciją $s(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek panašūs **a** ir **b**. Kuo didesnis panašumo koeficientas, tuo panašesniais vadinami objektai.*

Apibrėžimas. *Atstumas $d(\boldsymbol{a}, \boldsymbol{b})$ yra metrika, jei jis tenkina sąlygas:*

1) simetriškumo: $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$;

2) trikampio nelygybės: $d(\boldsymbol{a}, \boldsymbol{b}) \leq d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{c}, \boldsymbol{b})$;

3) netapačių objektų atskiriamumo:

$$d(\boldsymbol{a}, \boldsymbol{b}) = 0 \iff \boldsymbol{a} = \boldsymbol{b} .$$

Objektų artumo matai

Apibrėžimas. *Atstumu tarp dviejų objektų **a** ir **b** vadinsime neneigiamą skaitinę funkciją $d(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek skirtingi **a** ir **b**. Objektai yra panašūs, jei atstumas tarp jų mažas.*

Apibrėžimas. *Dviejų objektų **a** ir **b** panašumo koeficientu vadinsime skaitinę funkciją $s(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek panašūs **a** ir **b**. Kuo didesnis panašumo koeficientas, tuo panašesniais vadinami objektai.*

Apibrėžimas. *Atstumas $d(\boldsymbol{a}, \boldsymbol{b})$ yra metrika, jei jis tenkina sąlygas:*

1) simetriškumo: $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$;

2) trikampio nelygybės: $d(\boldsymbol{a}, \boldsymbol{b}) \leq d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{c}, \boldsymbol{b})$;

3) netapačių objektų atskiriamumo:

$$d(\boldsymbol{a}, \boldsymbol{b}) = 0 \iff \boldsymbol{a} = \boldsymbol{b} .$$

Objektų artumo matai

Apibrėžimas. *Atstumu tarp dviejų objektų **a** ir **b** vadinsime neneigiamą skaitinę funkciją $d(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek skirtingi **a** ir **b**. Objektai yra panašūs, jei atstumas tarp jų mažas.*

Apibrėžimas. *Dviejų objektų **a** ir **b** panašumo koeficientu vadinsime skaitinę funkciją $s(\boldsymbol{a}, \boldsymbol{b})$, nusakančią kiek panašūs **a** ir **b**. Kuo didesnis panašumo koeficientas, tuo panašesniais vadinami objektai.*

Apibrėžimas. *Atstumas $d(\boldsymbol{a}, \boldsymbol{b})$ yra metrika, jei jis tenkina sąlygas:*

1) simetriškumo: $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$;

2) trikampio nelygybės: $d(\boldsymbol{a}, \boldsymbol{b}) \leq d(\boldsymbol{a}, \boldsymbol{c}) + d(\boldsymbol{c}, \boldsymbol{b})$;

3) netapačių objektų atskiriamumo:

$$d(\boldsymbol{a}, \boldsymbol{b}) = 0 \iff \boldsymbol{a} = \boldsymbol{b} .$$

Vieno atributo įrašų artumas

Atributo tipas	Atstumas	Panašumo koeficientas
Vardinis	$d(u,v) = \begin{cases} 0\,, & \text{jei } u = v\,, \\ 1\,, & \text{jei } u \neq v\,. \end{cases}$	$s(u,v) = 1 - d(u,v)$
Ranginis (reikšmės- $0, 1, 2, \ldots, M$)	$d(u,v) = \dfrac{\lvert u - v \rvert}{M}$	$s(u,v) = 1 - d(u,v)$
Skaitinis	$d(u,v) = \lvert u - v \rvert$	$s(u,v) = \dfrac{1}{1 + d(u,v)}$
		$s(u,v) = \mathrm{e}^{-d(u,v)}$
		$s(u,v) = 1 - \dfrac{d(u,v) - d_{\min}}{d_{\max} - d_{\min}}$

Minkovskio metrika

Kai objektą nusako k skaitinių atributų, tenka matuoti atstumus tarp vektorių

$$\mathbf{u} = (u_1, u_2, \ldots, u_k) \text{ ir } \mathbf{v} = (v_1, v_2, \ldots, v_k)\,.$$

Dažniausiai naudojami Minkovskio metrikos

$$d(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{k} |u_i - v_i|^q \right)^{1/q}, \quad q \geq 1\,,$$

atskiri atvejai.

Minkovskio metrikos atskiri atvejai

- $q = 1$. *Manheteno (blokinis) atstumas*

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{k} |u_i - v_i| .$$

Kai **u** ir **v** yra binariniai vektoriai, ši metrika dar vadinama *Hamingo atstumu* ir reiškia nesutampančių bitų skaičių.

Minkovskio metrikos atskiri atvejai

- $q = 1$. *Manheteno (blokinis) atstumas*

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{k} |u_i - v_i| \, .$$

Kai **u** ir **v** yra binariniai vektoriai, ši metrika dar vadinama *Hamingo atstumu* ir reiškia nesutampančių bitų skaičių.

- $q = 2$. *Euklido metrika*

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{k} |u_i - v_i|^2} \, .$$

Minkovskio metrikos atskiri atvejai

- $q = 1$. *Manheteno (blokinis) atstumas*

$$d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{k} |u_i - v_i| .$$

Kai **u** ir **v** yra binariniai vektoriai, ši metrika dar vadinama *Hamingo atstumu* ir reiškia nesutampančių bitų skaičių.

- $q = 2$. *Euklido metrika*

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{k} |u_i - v_i|^2} .$$

- $q = \infty$. *Čebyšovo (maksimumo) atstumas*

$$d(\mathbf{u}, \mathbf{v}) = \max_i |u_i - v_i| .$$

Mahalanobio atstumas

$$d_M(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{u}-\mathbf{v})\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\mathbf{v})^T}\,,$$

čia $\boldsymbol{\Sigma}^{-1}$ yra atributų kovariacijų matricos $\boldsymbol{\Sigma} = (\sigma_{ij})_{k\times k}$ atvirkštinė matrica. Priminsime, kad i - tojo ir j - atributų kovariacija apibrėžiama taip:

$$\sigma_{ij} = \operatorname{cov}(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{n-1}\sum_{l=1}^{n}(x_{li}-\bar{x}_i)(x_{lj}-\bar{x}_j)\,,$$

čia $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})^T$ i - tojo atributo reikšmių stulpelis,

$$\bar{x}_i = \frac{1}{n}\sum_{l=1}^{n} x_{li}\,, \quad i = 1,2,\ldots,k\,.$$

Kai kovariacijų matrica yra vienetinė, Mahalanobio atstumas sutampa su Euklido metrika.

Nesutapimų metrika

Atstumas tarp diskrečiųjų vektorių paprastai reiškiamas nesutapimų metrika

$$d_{\Delta}(\mathbf{u}, \mathbf{v}) = \frac{1}{k} \sum_{\substack{i=1 \\ u_i \neq v_i}}^{k} 1 \,.$$

Vektorių panašumo matai

Vektorių skaliarinę sandaugą ir vektoriaus ilgį žymėsime

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{k} u_i v_i \,, \quad \|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} \,.$$

Be to, pažymėsime $\bar{\mathbf{v}}$ - vidurkių vektorių, sudarytą iš k vienodų koordinačių $\bar{v}$

$$\bar{\mathbf{v}} = (\bar{v}, \bar{v}, \ldots, \bar{v}) \,, \quad \bar{v} = \frac{1}{k} \sum_{i=1}^{k} v_i \,.$$

Vektorių panašumo matai

Pavadinimas	$s(\mathbf{u}, \mathbf{v})$ formulė
Suderinamumo	$1 - d_{\triangle}(\mathbf{u}, \mathbf{v})$
Žakardo	$\dfrac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \mathbf{u} \cdot \mathbf{v}}$
Vektorių kampo kosinusas	$\dfrac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$
Koreliacijos	$\dfrac{(\mathbf{u} - \bar{\mathbf{u}}) \cdot (\mathbf{v} - \bar{\mathbf{v}})}{\|\mathbf{u} - \bar{\mathbf{u}}\| \, \|\mathbf{v} - \bar{\mathbf{v}}\|}$

Binarinių vektorių panašumas

Jei **u** ir **v** yra binariniai vektoriai, tai 0 ir 1 išsidėstymą jų koordinatėse nusako dažnių lentelė

$u_i \backslash v_i$	0	1
0	k_{00}	k_{01}
1	k_{10}	k_{11}

Čia k_{lm} reiškia koordinačių, tenkinančių sąlygas $u_i = l\,,\ v_i = m\,,$ skaičių. Todėl suderinamumo koeficientas

$$s_{\text{sud}}(\mathbf{u}, \mathbf{v}) = 1 - d_{\Delta}(\mathbf{u}, \mathbf{v}) = \frac{k_{00} + k_{11}}{k_{00} + k_{01} + k_{10} + k_{11}}\,.$$

Skaičiuojant Žakardo panašumo koeficientą $s_J(\mathbf{u}, \mathbf{v})$, sutampantys nuliai nėra svarbūs

$$s_J(\mathbf{u}, \mathbf{v}) = \frac{k_{11}}{k_{01} + k_{10} + k_{11}}\,.$$

Koreliacijos koeficientai

Kartais jais remiantis vertinamas objektų (įrašų) panašumas. Statistikoje įprastas **u** ir **v** koreliacijos koeficiento apibrėžimas yra

$$r(\mathbf{u},\mathbf{v}) = \frac{\mathrm{cov}(\mathbf{u},\mathbf{v})}{s_{\mathbf{u}} s_{\mathbf{v}}} = \frac{(\mathbf{u}-\bar{\mathbf{u}})\cdot(\mathbf{v}-\bar{\mathbf{v}})}{\|\mathbf{u}-\bar{\mathbf{u}}\|\,\|\mathbf{v}-\bar{\mathbf{v}}\|}\,.$$

Koreliacijos koeficientas visada priklauso intervalui $[-1;\ 1]$.

Pastaba. Jei $r(\mathbf{u},\mathbf{v}) = 0$, tai dar nereiškia, kad tarp vektorių nėra jokios priklausomybės. Pavyzdžiui, imkime vektorius

$$\begin{aligned} \mathbf{u} &= (-2,-1,0,1,2)\,, \\ \mathbf{v} &= (4,1,0,1,4)\,, \end{aligned}$$

kurių koordinatės susietos lygybe $v_i = u_i^2$. Tačiau, nežiūrint to, $r(\mathbf{u},\mathbf{v}) = 0$. Kitaip sakant, koreliacijos koeficiento lygybė nuliui rodo, kad nėra tiesinės priklausomybės.

Heterogeniškų vektorių artumo matai

1. Pasirenkame i -tojo atributo atstumą $d_i(\mathbf{u}, \mathbf{v}) = d(u_i, v_i) \in [0;\ 1]$.

Heterogeniškų vektorių artumo matai

1. Pasirenkame i -tojo atributo atstumą $d_i(\mathbf{u}, \mathbf{v}) = d(u_i, v_i) \in [0;\ 1]$.
2. Apibrėžiame i -tojo atributo asimetrijos ir trūkstamų reikšmių indikatorių δ_i . Jis lygus 0, jei i - tasis atributas asimetrinis ir $u_i = v_i = 0$ arba kai kuriame nors vektoriuje trūksta i - tosios koordinatės. Kitais atvejais $\delta_i = 1$.

Heterogeniškų vektorių artumo matai

1. Pasirenkame i -tojo atributo atstumą $d_i(\mathbf{u}, \mathbf{v}) = d(u_i, v_i) \in [0;\ 1]$.
2. Apibrėžiame i -tojo atributo asimetrijos ir trūkstamų reikšmių indikatorių δ_i . Jis lygus 0, jei i - tasis atributas asimetrinis ir $u_i = v_i = 0$ arba kai kuriame nors vektoriuje trūksta i - tosios koordinatės. Kitais atvejais $\delta_i = 1$.
3. Pasirenkame i -tojo atributo svorį $w_i \geq 0$ taip, kad visų svorių suma būtų lygi 1, t.y. $w_1 + w_2 + ... + w_k = 1.$

Heterogeniškų vektorių artumo matai

1. Pasirenkame i -tojo atributo atstumą $d_i(\mathbf{u}, \mathbf{v}) = d(u_i, v_i) \in [0;\ 1]$.
2. Apibrėžiame i -tojo atributo asimetrijos ir trūkstamų reikšmių indikatorių δ_i . Jis lygus 0, jei i - tasis atributas asimetrinis ir $u_i = v_i = 0$ arba kai kuriame nors vektoriuje trūksta i - tosios koordinatės. Kitais atvejais $\delta_i = 1$.
3. Pasirenkame i -tojo atributo svorį $w_i \geq 0$ taip, kad visų svorių suma būtų lygi 1, t.y. $w_1 + w_2 + ... + w_k = 1.$
4. Randame atstumą tarp vektorių **u** ir **v**

$$d(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{k} \delta_i \right)^{-1} \sum_{i=1}^{k} w_i \delta_i d_i(\mathbf{u}, \mathbf{v})$$