Teorijos klausimai iš temos KLASTERIZAVIMAS (Klasterinės analizės teorija.pdf)

1. Kuo skiriasi klasterizacija nuo klasifikavimo?

Klasterizacija yra pagrindinė duomenų išgavimo užduotis ir pagrindinė technika atliekant statistinę duomenų analizę.

Klasifikavimas yra objektų priskyrimas tam tikroms tikslinėms grupėms, dar kitaip vadinamoms klasėmis.

2. Koks yra klasterinės analizės tikslas?

Klasterinės analizės tikslas – suskirstyti objektus taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.

3. Išvardykite 5 klasterinės analizės etapus;

• Pasirinkti klasterizuojamus objektus.

• Nuspręsti, pagal kokius požymius klasterizuosime.

• Pasirinkti kiekybinį matą, kuriuo matuosime objektų panašumą.

• Vienu ar kitu metodu suskirstyti objektus į klasterius.

• Peržiūrėti gautus rezultatus.

4. Pagrindiniai 3 klasterinės analizės elementai;

Tiriamieji objektai - objektai, kurie bus klasterizuojami.

Objektų požymių matavimai - kiekvienam tiriamam objektui yra priskirti tam tikri požymiai, kurie apibūdinami skaitinėmis reikšmėmis.

Panašumo matai - kiekybiniai matai, kuriais matuojamas objektų panašumas arba skirtingumas.

5. Kokios matavimų skalės duomenims skaičiuojami metriniai atstumo matai?

Metriniai atstumo matai naudojami tada, kai objektus charakterizuojantys požymiai matuojami pagal intervalų arba santykių skalę

6. Kada objektai panašesni: kai atstumo mato reikšmė didesnė, ar kai mažesnė?

Objektai yra panašesni, kai atstumo mato reikšmė yra mažesnė, nes metriniai atstumo matai yra naudojami skirtumų tarp objektų matavimui

7. Kokias sąlygas turi tenkinti skaitinė neneigiama funkcija $d(X, Y)$, kad ją būtų galima vadinti metrika?

1) Simetriškumo: $d(X,Y) = d(Y,X)$;
2) Trikampio nelygybės: $d(X,Y) \leq d(X,Z) + d(Y,Z)$;
3) Netapačių objektų atskiriamumo: jei $X \neq Y$, tai $d(X,Y) \neq 0$;
4) Tapačių objektų neatskiriamumo: jei $d(X,Y) = 0$, tai X ir Y identiški.

8. Kuo skiriasi hierarchiniai ir nehierarchiniai klasterizavimo metodai?

Hierarchinių metodų rezultatai nusako klasterių tarpusavio hierarchiją, t. y. visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos savo ruožtu dar mažesni ir t.t.

Nehierarchiniai metodai paprastai taikomi tada, kai iš anksto žinomas (pasirenkamas) klasterių skaičius ir norima tiriamus objektus klasterizuoti.

9. Užrašykite hierarchinių jungimo metodų klasterizavimo schemos 4 žingsnius.

1. Turime N klasterių po 1 objektą ir NxN simetrinę atstumų matricą dij , i,j = 1, N.

2. Pagal atstumų matricą nustatome du klasterius, tarp kurių atstumas yra mažiausia (kurie yra panašiausi). Tarkime, kad tai klasteriai U ir V.

3. Sujungiame klasterius U ir V. Naują klasterį pavadiname (UV). Tada atstumų matricą pakeičiame taip:

a) išbraukiame stulpelius ir eilutes, atitinkančius klasterius U ir V;

b) pridedame eilutę ir stulpelį su atstumais tarp (UV) ir likusiųjų klasterių.

4. Kartojame 2 ir 3 žingsnius (N — 1) kartų. Procesą baigiame, kai visi objektai yra viename

klasteryje.

10. Užrašykite nehierarchinio k-vidurkių metodo 3 pagrindinius klasterizavimo schemos žingsnius.
Pradinių klasterių formavimas: Objektai suskirstomi į k pradinių klasterių.

Atstumų skaičiavimas ir objektų priskyrimas: Paeiliui apskaičiuojamas kiekvieno objekto atstumas iki klasterių centrų (atstumas paprastai skaičiuojamas naudojantis Euklido metrika arba jos kvadratu). Objektas skiriamas į artimiausią klasterį. Klasterių centrai perskaičiuojami.

Iteracija iki stabilumo: 2 žingsnis kartojamas tol, kol perskirstymų daugiau nėra.

Teorijos klausimai iš temos STATISTINĖS HIPOTEZĖS (Statistinės hipotezės (2024-04-17)

1. Kas yra statistinė hipotezė? Pateikti apibrėžimą.

Statistine hipoteze vadinamas teiginys apie statistinių duomenų tikimybinį skirstinį arba apie tam tikro skirstinio parametrą.

2. Ką nurodo nulinė ir alternatyvioji hipotezės? – parašyti paaiškinimą.

Nulinė hipotezė yra statistinė hipotezė, kuri teigia, kad nėra statistiškai reikšmingo skirtumo ar ryšio tarp tiriamų duomenų. Tai hipotezė, kurią siekiama atmesti siekiant įrodyti, kad alternatyvioji hipotezė yra teisinga. Alternatyvioji hipotezė yra statistinė hipotezė, kuri teigia, kad yra statistiškai reikšmingas skirtumas ar ryšys tarp tiriamų duomenų. Tai hipotezė, kurią tyrėjas nori įrodyti kaip teisingą, o nulinę hipotezę atmeta kaip klaidingą.

3. Kas yra statistinis kriterijus? Parašyti apibrėžimą.

Taisyklė, pagal kurią iš imties duomenų darome išvadas apie hipotezės teisingumą vadinama statistiniu kriterijumi.

4. Paaiškinti kada padaroma I rūšies klaida, kada II rūšies klaida.

I rūšies klaida įvyksta, kai nulinė hipotezė (H_o) atmetama, nors ji yra teisinga. Tai reiškia, kad mes padarome išvadą, jog yra statistiškai reikšmingas skirtumas ar ryšys, nors iš tikrųjų tokio skirtumo ar ryšio nėra.

II rūšies klaida įvyksta, kai nulinė hipotezė (H_o) neatmetama, nors ji yra klaidinga. Tai reiškia, kad mes padarome išvadą, jog nėra statistiškai reikšmingo skirtumo ar ryšio, nors iš tikrųjų toks skirtumas ar ryšys egzistuoja.

5. Pagal pateiktą uždavinio struktūrą nurodyti, kuris statistinis kriterijus yra taikomas.

Priklauso ar imtys priklausomos ar nepriklausomos.

Paprasčiausias priklausomų imčių atvejis yra tų pačių objektų pakartotiniai matavimai.

Jeigu norite atsakyti į panašius klausimus:

• Ar vyrų ir žmonų šeimoje pajamos yra vienodos? Ar studentai geriau išlaiko rudens nei

pavasario sesijos egzaminus? Ar eismo intensyvumas kelyje A1 vasarą didesnis nei žiemą?

– taikomas Vilkoksono ženklų kriterijus dviems priklausomoms imtims.

• Ar eismo intensyvumas žiemą Vilniuje didesnis nei Kaune?

– taikomas Mano-Vitnio-Vilkoksono kriterijus dviems nepriklausomoms imtims.

• Ar kaimo, rajonų centrų ir didžiųjų miestų gyventojai būstui išlaikyti išleidžia vienodą sumą

pinigų; Ar dienos metu skirtinguose keliuose yra vienodas eismo intensyvumas?

– taikomas Kruskalo-Voliso ranginis kriterijus k > 2 nepriklausomoms imtims.

• Ar tame pačiame kelyje eismo intensyvumas skirtingomis valandomis vienodas?

– taikomas Frydmano kriterijus priklausomoms k > 2 imtims.

6. Ką parodo reikšmingumo lygmuo?

Reikšmingumo lygmuo (α) parodo I rūšies klaidos tikimybę, tai yra tikimybę atmesti teisingą nulinę hipotezę. Tai nustatytas slenkstis, pagal kurį vertinama p-reikšmė, ir dažniausiai pasirenkamas kaip 0.05. Jei $p < \alpha$, nulinė hipotezė atmetama; jei $p \geq \alpha$, nulinė hipotezė neatmetama

7. Paaiškinkite p-reikšmės naudojimą hipotezių išvadų formulavime.

Jeigu $p < \alpha$: Hipotezė H_o atmetama. Jeigu $p \geq \alpha$: Hipotezė H_o neatmetama. P-reikšmė yra mažiausias reikšmingumo lygmuo, su kuriuo teisinga H_o gali būti atmesta turimiems duomenims

8. Bus pateikti statistinių hipotezių pavyzdžiai, kai taikomas vienas iš statistinių kriterijų (t kriterijus, Chi_kvadrato kriterijus, Vilkoksono kriterijus, Kruskalo-Voliso kriterijus, ANOVA, Frydmano kriterijus) Jus reiks pagal nurodytą p-reikšmę atsakyti kokia yra hipotezės išvada.

• Vilkoksono ženklų kriterijus priklausomoms imtims – kriterijus skirtas hipotezei apie dviejų priklausomų (porinių) imčių skirstinių lygybę, tikrinti. Išvada: Jei p-reikšmė < 0.05, tai yra statistiškai reikšmingas skirtumas tarp grupių medianų.

• Mano-Vitnio-Vilkoksono kriterijus nepriklausomoms imtims – kriterijus skirtas hipotezei apie dviejų nepriklausomų imčių skirstinių lygybę, tikrinti.

• Kruskalo-Voliso ranginis kriterijus nepriklausomoms imtims - kriterijus skirtas hipotezei apie dviejų ar daugiau populiacijų skirstinių lygybę, esant nepriklausomoms imtims, tikrinti. Išvada: Jei p-reikšmė < 0.05, tai yra statistiškai reikšmingas skirtumas tarp bent vienos grupės ir kitų.

• Frydmano kriterijus priklausomoms imtims – kriterijus skirtas hipotezei apie k kintamųjų (k > 2) skirstinių lygybę tikrinti, kai imtys yra priklausomos. Išvada: Jei p-reikšmė < 0.05, tai yra statistiškai reikšmingas skirtumas tarp bent vienos grupės ir kitų.

• T kriterijus:

Išvada: Jei p-reikšmė < 0.05, tai yra statistiškai reikšmingas skirtumas tarp grupių vidurkių.

• Chi-kvadrato kriterijus:

Išvada: Jei p-reikšmė < 0.05, tai reiškia, kad turime pakankamai įrodymų atmesti nulines hipotezes ir priimti alternatyvią hipotezę apie priklausomybę.

• ANOVA (vienos krypties dispersijos analizė):

Išvada: Jei p-reikšmė < 0.05, tai yra statistiškai reikšmingas skirtumas tarp bent vienos grupės ir kitų.

Teorijos klausimai iš temos FAKTORINĖ ANALIZĖ (faktorinė_paskaita.pdf)

1**. Apibrėžkite faktorinės analizės uždavinį.**

Atsižvelgiant į kintamųjų tarpusavio koreliacijas, suskirstyti stebimus kintamuosius įgrupes, kurias vienija koks nors tiesiogiainestebimas (*latentinis*) faktorius. Koks tas faktorius, sprendžiame patys nagrinėdami grupes sudarančius kintamuosius.

2. Apibrėžkite faktorinės analizės tikslą.

minimaliai prarandant informacijos pakeisti stebimą reiškinį charakterizuojančių požymių aibę kelių faktorių rinkiniu.

3. Išvardykite faktorinės analizės pagrindinius 4 etapus.

1. tikriname, ar duomenys tinka faktorinei analizei;

2. faktorių skaičiaus nustatymas bei faktorių skaičiavimo metodo parinkimas;

3. faktorių sukimas ir interpretavimas;

4. faktorių reikšmių įverčių skaičiavimas.

4. Paaiškinkite kokie elementai sudaro faktorinės analizės modelį.

- Tarkime stebime ***k*** kintamųjų $X_1, X_2, \dots, X_k$.
- Modelis grindžiamas prielaida, kad kiekvieno kintamojo X_i elgesį sąlygoja ***m*** bendrųjų latentinių faktorių $F_1, F_2, \dots, F_m$ ir specifinis (charakteringasis) latentinis faktorius $\varepsilon_i, i = \overline{1,k}$.
- Bendrųjų faktorių yra mažiau nei kintamųjų $(m < k)$. Tarkim X_i, $i = \overline{1,k}$ nuo faktorių priklauso tiesiškai. Tada ***FA* matematinis modelis** užrašomas:

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \dots + \lambda_{1m}F_m + \varepsilon_1, \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \dots + \lambda_{2m}F_m + \varepsilon_2, \\ &\vdots \\ X_k &= \lambda_{k1}F_1 + \lambda_{k2}F_2 + \dots + \lambda_{km}F_m + \varepsilon_k. \end{aligned}$$

- Daugikliai $\lambda_{ij}, i = \overline{1,k}, j = \overline{1,m}$ – vadinami ***faktorių svoriais***.

5. Išvardykite 4 faktorinės analizės modelio prielaidas;

a) stebimi kintamieji pasiskirstę pagal normalųjį dėsnį, t. y. $X_i \sim N(\mu_i, \sigma_i^2)$;
b) bendrieji faktoriai F_j **nekoreliuoti** ir $DF_j = 1$;
c) charakteringieji faktoriai ε_i, **nekoreliuoti** ir $D\varepsilon_i = \tau_i$;
d) faktoriai F_j ir ε_i **nekoreliuoti**, čia $i = \overline{1,k}, j = \overline{1,m}$.

6. Nurodykite tris reikalavimus, kuriuos turi tenkinti duomenys, kad jie tiktų faktorinei analizei;

ChatGPT

Adekvati koreliacija tarp kintamųjų: Koreliacijos matricos patikrinimas, taip pat „Bartlett's Test of Sphericity“ ir „Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy“ gali būti naudojami įvertinti, ar duomenys tinkami faktorinei analizei. Aukštos KMO reikšmės (didesnės nei 0.6) rodo, kad faktorinė analizė gali būti tinkama.

Pakankamai didelis imties dydis: Faktorinei analizei atlikti reikia pakankamai didelio imties dydžio, kad būtų galima gauti patikimus rezultatus. Bendroji taisyklė yra turėti bent 5–10 stebėjimų vienam analizuojamam kintamajam, nors geriausia turėti didesnę imtį. Mažos imtys gali sukelti nestabilumą ir mažą faktorių išskyrimo patikimumą.

Normalumas ir išskirtinių verčių nebuvimas: Nors faktorinė analizė gali būti taikoma ir su ne normaliai pasiskirsčiusiais duomenimis, normalumo prielaida padeda gauti patikimesnius rezultatus.

7. Kokią išvadą galima padaryti atlikus *Bartlett'o sferiškumo* testą ir gavus žemiau pateiktus rezultatus?

```
> cortest.bartlett(cor(D),n = nrow(D))
$chisq
[1] 207.1133

$p.value
[1] 7.591189e-36

$df
[1] 15
```

Kadangi, p-reikšmė 7.591189e-36 < 0.05 ,tai galime atmesti nulinę hipotezę, ir sakyti ,kad sferiškumas yra.

8. Kokias išvadas galima padaryti gavus *Kaiserio-Meyerio-Olkino mato ir* $MSAi$ reikšmes pateiktas žemiau:

```
> KMO(D)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = D)
Overall MSA =  0.83
MSA for each item =
 mpg  cyl disp   hp drat   wt
0.85 0.84 0.82 0.83 0.87 0.79
```

0.83 > 0.5. Vadinasi bendras modelis yra tinkamas, kadangi atitinka reikalavimus.

Visi šie kintamieji yra tinkami faktorinei analizei, nes jų MSA reikšmės yra didesnės nei 0.5

9. Užrašykite: a) tiesines daugdaras, kurios vadinamos pagrindinėmis komponentėmis; b) sąlygas, kurias turi tenkinti šios tiesinės daugdaros.

- Taikant pagrindinių komponenčių analizę randamos tarpusavyje nekoreliuojančių kintamųjų $X_1, X_2, \dots, X_k$ ***tiesinės daugdaros*** (kombinacijos) $Y_1, Y_2, \dots, Y_k$, t.y.

$$\boldsymbol{Y_1} = \sum_{j=1}^{k} \alpha_{1j} X_j, \quad \boldsymbol{Y_2} = \sum_{j=1}^{k} \alpha_{2j} X_j, \dots, \boldsymbol{Y_k} = \sum_{j=1}^{k} \alpha_{kj} X_j ,$$

tenkinančios sąlygas:

1) $cov(\boldsymbol{Y_i}, \boldsymbol{Y_j}) = \boldsymbol{0}, \boldsymbol{i}, \boldsymbol{j} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{k}, \boldsymbol{i} \neq \boldsymbol{j};$

2) $DY_1 \geq DY_2 \geq \cdots \geq DY_k;$

3) $\sum_{i=1}^{k} DY_i = \sum_{i=1}^{k} DX_i = D.$

Šios tiesinės daugdaros vadinamos ***pagrindinėmis komponentėmis***.

10. Tarkim, pirmoji pagrindinė komponentė yra $Y1=\alpha 11X1+\cdots+\alpha 1kXk$. Kaip apskaičiuojami koeficientai $\alpha 11,\alpha 12,\ldots,\alpha 1k$?

koeficientai $\alpha 11,\alpha 12,\ldots,\alpha 1k$α11,α12,…,α1k apskaičiuojami naudojant kintamųjų kovariacijos arba koreliacijos matricą ir jos savivektorius.

11. Kiek procentų bendrosios dispersijos paaiškina pirmoji pagrindinė komponentė?

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.0463	1.0715	0.57737	0.39289	0.3533	0.22799
Proportion of Variance	0.6979	0.1913	0.05556	0.02573	0.0208	0.00866
Cumulative Proportion	0.6979	0.8892	0.94481	0.97054	0.9913	1.00000

69,79 %

12. Kiek procentų bendrosios dispersijos paaiškina dvi pirmosios pagrindinės komponentės, pagal žemiau pateiktus rezultatus?

```
> pk <- prcomp(X)
> summary(pk)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.0463	1.0715	0.57737	0.39289	0.3533	0.22799
Proportion of Variance	0.6979	0.1913	0.05556	0.02573	0.0208	0.00866
Cumulative Proportion	0.6979	0.8892	0.94481	0.97054	0.9913	1.00000

Kadangi duota, kad pirmosios dvi pagrindinės komponentės paaiškina 69.79% ir 19.13% bendrosios dispersijos atitinkamai, tai norint sužinoti, kiek procentų bendrosios dispersijos paaiškina šios dvi komponentės kartu, reikia juos tiesiog sudėti:

0.6979 + 0.1913 = 0.8892

0.6979+0.1913=0.8892

Taigi, pirmosios dvi pagrindinės komponentės paaiškina 88.92% bendrosios dispersijos.

13. Taikydami pagrindinių komponenčių metodą skaičiuojame koreliacijų matricą. Kaip

Taigi, pagal pagrindinių komponenčių metodą, koreliacijų matricos tikrinės reikšmės nurodo, kiek koreliacijos yra paaiškinta kiekvieno pagrindinio komponento, o tikriniai vektoriai nurodo, kaip kiekviena iš naujų kintamųjų (pagrindinių komponenčių) yra susijusi su pradinėmis kintamomis.