

**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**Vardenis Pavardenis**

n variantas

Studijų modulio

**P160M101 DAUGIAMATĖ STATISTINĖS ANALIZĖ**

**3 laboratorinis darbas**

**KAUNAS, 2020**

# UŽDUOTIS Nr. 1. Hierarchinis klasterizavimas

1. Imties sudarymas. Atrinkite 12 miestų (miestelių, gyvenviečių), kurių pavadinimai prasideda ta pačia raide kaip ir jūsų pavardė. Sudarykite atstumų matricą naudodamiesi atstumais keliais (galite naudotis maps.lt, project-osrm.org, google.com/maps ar kitu jums patogiu žemėlapiu).
2. Pavaizduokite žemėlapį su pasirinktais miestais.
3. Atlikite hierarchinį imties klasterizavimą (naudokite R funkciją hclust()). Panaudokite keturis klasterizavimo metodus (keisdami parametrą method). Palyginkite gautas dendrogramas. Kuris iš gautų rezultatų labiausiai atitinką miestų grupąvimą lyginant su jūsų subjektyviu požiūriu. Atsakymą pakomentuokite.
4. Pasirinkę vieną iš hierarchinio klasterizavimo metodu ir naudodami ilgiausių dendrogramos šakų būdu parinkite klasterių kiekį. Pateikite miestų klasterius.

# UŽDUOTIS Nr. 2. Daugiamačių skalių metodas

1. Naudodami užduoties nr. 1. imtį (miestų atstumų lentelę), sudarykite ir pateikite daugiamačių skalių metodo rezultatus naudodami vieną metrinį ir vieną ne metrinį daugiamačių skalių metodą. Palyginkite gautus rezultatus su realiu žemlėlapio vaizdu.

**UŽDUOTIS NR. 3. GAUSO skirstinių MIŠINIo KLASTERIZAVIMAS**

1. Sugeneruokite duomenų imtis atitinkančias Gauso skirstinių mišinį nurodyą variantų lentelėje. Imties dydžiai – 100, 500, 3000 taškų. Generavimui galite naudotis bet kokiomis, savo pasirašytomis, ar paketuose pateikiamomis funkcijomis.

Variantai:

|  |  |
| --- | --- |
| 1. | Ambrutis Andrius |
| 2. | Babkauskas Dominykas |
| 3. | Baliūnaitė Gintarė |
| 4. | Dzidolikas Vilius |
| 5. | Jablonskaitė Kamilija |
| 6. | Kraujalis Vytautas |
| 7. | Levinskas Matas |
| 8. | Lugas Dovydas |
| 9. | Pacevičius Rytis |
| 10. | Petrauskaitė Vaiva |
| 11. | Sprainaitis Mantas |
| 12. | Uzdila Ernestas |

1. Įvertinkite mišinio paramterus naudodami mclust paketą. Kovariacinių matricų struktūros parinkimui naudokite BIC kriterijų. Paaiškinkte kokia klasterių kovariacinė struktūra buvo parinkta. Palyginkite ją su tikrosiomis reikšmėmis.

**UŽDUOTIS NR. 4. k-vidurkių klasterizavimas**

**Miestų duomenys.**

1. Pirmos užduoties duomenis papildykite pasirinktų miestų koordinatėmis.
2. Panaudokite k-vidurkių metodą miestams klasterizuoti į 2–5 klasterius. Pateikite gautus klasterius. Gautą rezultatą palyginkite su anksčiau gautais hierarchinio klasterizavimo rezultatais (jei skaidytume į 2–5 klasterius pagal hierarchinį klasterizavima).

**Gauso mišinių modelis.**

1. Panaudokite k-vidurkių metodą antros užduoties duomenims kai imties dydis yra 500, kai klasterių kiekis yra 2–5.
2. Vizualizuokite gautus klasterius (pvz., pasirinkdami dvi pirmąsias pagrindines komponentes).
3. Nubraižykite atstumų kvadratų sumos klasterių viduje (within cluster sum of squares). Panaudokita alkūnės metodą optimaliam klasterio kiekio parinkimui.
4. Panaudokite vidutinį silueto matą optimaliam klasterio kiekio parinkimui.

Užduoties ataskaita pateikiama per <http://moodle.ktu.lt/> Ataskaitoje reikia pateikti visas užduotis atitinkantį **pilną R kodą**, bei jo įvykdymo rezultatą (tekstinę ar grafinę), atsakyti į užduotyse nurodytus klausimus, aprašyti sunkumus su kuriais susidūrėte ir jų sprendimo būdus.

# SPRENDIMAS

...

Literatūra

1. mclust package homepage. <https://sites.stat.washington.edu/mclust/>
2. Luca Scrucca. A quick tour of mclust, 2020. <https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html>
3. Luca Scrucca, Michael Fop, T. Brendan Murphy and Adrian E. Raftery (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, Vol. 8, No. 1, p. 289 – 317. <https://journal.r-project.org/archive/2016/RJ-2016-021/RJ-2016-021.pdf>
4. K-means Cluster Analysis. UC Business Analytics R Programming Guide. https://uc-r.github.io/kmeans\_clustering
5. Kisung You (2019). *L2 distance between Gaussian Mixture Models*. <http://kyoustat.com/pdf/note004gmml2.pdf>