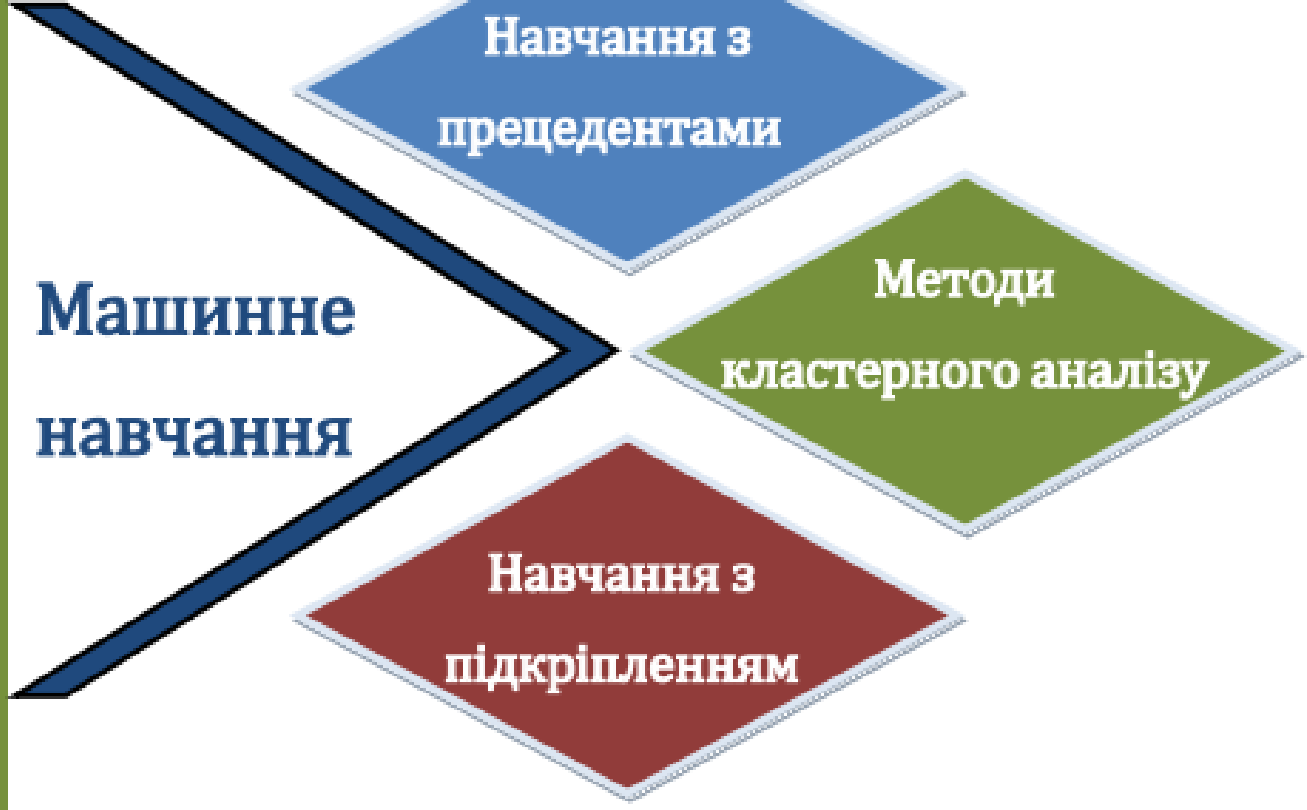

ЧАСТИНА 2

Методи кластерного аналізу

Антонюк С.В., Горбатенко М.Ю.,
Кириченко О.Л., Малик І.В.

Частина 2

Міністерство освіти і науки України
Чернівецький національний університет
імені Юрія Федьковича

Антонюк С.В., Горбатенко М.Ю., Кириченко О.Л., Малик І.В.

МАШИННЕ НАВЧАННЯ

Методи кластерного аналізу

Навчальний посібник

Чернівці
2023

УДК 004.8:519.7] (075.8)
М 382

Друкується за ухвалою Вченої ради Чернівецького національного університету імені Юрія Федьковича (протокол № 12 від 01.12.2021)

Рецензенти:
Сливка-Тилищак Ганна Іванівна, доктор фіз.-мат. наук, професор (Ужгородський національний університет);
Василик Ольга Іванівна, доктор фіз.-мат. наук, доцент (Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»)

Машинне навчання. Методи кластерного аналізу: Навчальний посібник. / Антонюк С.В., Горбатенко М.Ю., Кириченко О.Л., Малик І.В. Чернівці : Чернів. нац. ун-т ім. Ю.Федьковича, 2023, 208 с.

Пропоноване видання містить теоретичний матеріал та модельні приклади з дисципліни «Методи та засоби кластерного аналізу». Подано основні методи і описано моделі машинного навчання без вчителя. Наведено велику кількість модельних прикладів, які дозволяють краще усвідомити теоретичний матеріал, а також сформовано завдання і методичні вказівки для лабораторних робіт по курсу. Навчальний посібник призначено для студентів технічних і фізико-математичних спеціальностей, які володіють базовими знаннями з математичного аналізу, теорії ймовірностей, математичної статистики, програмування. Для студентів навчально-наукового інституту фізико-технічних і комп'ютерних наук ЧНУ, а також студентів інших факультетів, які навчаються за спеціальностями «Комп'ютерні науки»

УДК 004.8 (519.7)

Зміст

Вступ

У даному підручнику буде розглянуто основні методи кластеризації або навчання без вчителя (unsupervised learning). Слід зауважити, що на відміну від навчання зі вчителем (supervised learning) в кластерному аналізі з'являються нові задачі, що пов'язані з визначенням оптимальної кількості кластерів k_{opt}, які не розглядаються в інших задачах машинного навчання. На відміну від задач класифікації, неможливо побудувати однозначної міри відповідності між знайденими кластерами та "реальними групами", оскільки даних груп (вчителя) не існує. У зв'язку з цим, в кластерному аналізі відсутні перевірки на основі точності та k-fold валідація.

Перший розділ присвячено деяким математичним методам, які будуть безпосередньо розглядатися при пошуку оптимальних параметрів в задачах кластеризації. Основна увага в цьому розділі буде приділена пошуку наближених розв'язків оптимізаційних задач та моделювання значень випадкових величин на основі заданого розподілу (щільності $f(x)$).

У другому розділі підручника розглянемо основні поняття кластерного аналізу, а саме, кластеру, матриці відстаней, матриці подібності, оптимального числа кластерів, сумішей, нечі-

ткої кластеризації. Також, увага буде приділено кластеризації нечислових даних (бінарних, номінальних та порядкових змінних). Поряд із цим, у цьому розділі будуть розглянуті основні методи кластеризації неструктурованих даних - кластеризація текстів та графів. Зауважимо, що методи які використовуються для кластеризації неструктурованих даних, відрізняються від класичних методів кластеризації числових даних оскільки неможливо однозначно визначити відстань чи подібність між об'ктами.

Четвертий розділ присвячений розгляду кластеризації у випадку часткової наявності вчителя, тобто за умови наявності y_i для деякого відсотку даних. Задача кластеризації в цьому випадку тісно переплетена із задачами класифікації, що і буде обговорено у приорітетах розвязання задач із частковим навчанням.

Термінологічний показчик

Список літератури

[1] Charu C. Aggarwal and Chandan K. Reddy, editors. *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[2] B. Everitt, S. Landau, M. Leese, D. Stahl, and an O'Reilly Media Company Safari. *Cluster Analysis, 5th Edition*. John Wiley & Sons, 2011.

[3] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[4] Jon M. Kleinberg. An impossibility theorem for clustering. pages 463–470, 2002.

[5] O. Kyrychenko. Information technologies for statistical cluster analysis of information in complex networks. *Computer Systems and Information Technologies*, (4):47–51, 2022.

[6] Thomas Reutterer and Daniel Dan. Cluster Analysis in Marketing Research. In Christian Homburg, Martin Klarmann, and Arnd Vomberg, editors, *Handbook of Market Research*, Springer Books, page 345. Springer, June 2022.

[7] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, 1987.

[8] Остапов С.Е. Кириченко О.Л., Малик І.В. Стохастичні моделі в задачах штучного інтелекту. *Вісник Київського національного університету імені Тараса Шевченка. Серія фізико-математичні науки*, (2):53–57, 2021.

[9] Остапов С.Е. Кириченко О.Л., Малик І.В. Аналіз кластерної структури Інтернет-мереж на основі випадкових матриць. *Проблеми керування та інформатики*, (1):37–46, 2022.

Антонюк Світлана Володимирівна;

Горбатенко Микола Юрійович;

Кириченко Оксана Леонідівна;

Малик Ігор Володимирович

Навчальне видання

МАШИННЕ НАВЧАННЯ

Методи кластерного аналізу

Навчальний посібник

Відповідальний за випуск Я.М. Дрінь

Літературний редактор О.В. Колодій

Підписано до друку 28.06.2023. Формат 60 х 84/16.
Папір офсетний. Друк різографічний. Ум.-друк. арк. 4,8.
Обл.-вид.арк.4,0. Тираж 50. Зам. Н-222.
Друкарня Чернівецького національного університету
58012, Чернівці, вул. Коцюбинського, 2

Свідоцтво суб'єкта видавничої справи ДК №891 від 08.04.2002 р.