

МОВА R ЯК ІНСТРУМЕНТ ДЛЯ НОВІТНЬОГО АНАЛІЗУ ДАНИХ

R – мова програмування та середовище розробки для статистичного аналізу даних. Середовище містить у собі інтерпретатор мови та різного роду допоміжні інструменти для відображення результатів певних задач. RStudio існує в двох базових версіях: для настільного локального застосування та для встановлення на окремий віддалений сервер і доступ буде виконуватись через веб-сторінку. Також використовувати можливості R є можливим за допомогою спеціальних розширень під Visual Studio.

Під час роботи з R є можливість виконувати одразу багато інструкцій, що записані в окремому файлі. Хоча R і орієнтована на розв'язок і аналіз статистичних задач, вона може використовуватись для матричних обрахунків з порівняльною швидкістю до математичних пакетів GNU Octave або MATLAB. На даний час створено багато пакетів для статистичних обчислень, біоінформатики, оптимізації, кластерного та іншого аналізу.

Для візуалізації даних у середовищі наявні: двовимірні, тривимірні графіки; гістограми; діаграми (схеми Ганта); звіти. Для роботи з даними R підтримує безліч базових операцій.

Для R написано багато різних бібліотек для відображення результатів аналізу. Найбільш популярні: dplyr, ggplot2, ggthemes.

Бібліотека dplyr робить зручнішою роботу із датафреймами, ggplot2 – найпопулярніша бібліотека для візуалізації, а ggthemes – теми для неї. Для того, аби встановити бібліотеку, потрібно виконати команду `install.packages("<назва бібліотеки>")`, а для того, аби завантажити її функції в оперативну пам'ять – команду `library(<назва бібліотеки>)`. `install.packages` треба виконати один раз, `library` виконувати кожного разу перед початком роботи.

Розглянемо як R допомагає в аналізі даних на прикладі кластерного аналізу. Кластерний аналіз потребує здійснення таких послідовних кроків: проведення вибірки об'єктів для кластеризації; визначення множини ознак, за якими будуть оцінюватися відібрані об'єкти; оцінка міри подібності об'єктів; застосування кластерного аналізу для створення груп подібних об'єктів; перевірка достовірності результатів кластерного рішення. Кожен з цих кроків відіграє значну роль у практичному здійсненні аналізу.

В результаті аналізу сукупності вхідних даних створюються однорідні групи у такий спосіб, що об'єкти всередині цих груп подібні між собою за деяким критерієм, а об'єкти з різних груп відрізняються один від одного. Отже, після того як стало зрозуміло, що собою являє кластерний аналіз в цілому та основи розробки в середовищі R можна приступити до рішення задачі кластеризації на основі табличних даних, які зберігаються в середовищі MS SQL Server.

Середовище розробки R та допоміжні пакети дають можливість дуже просто підключитись до нашого серверу та вилучити необхідні дані для аналізу. Серед великої кількості існуючих методів кластеризації даних одним з найчастіше використовуваних є метод k-means. Даний метод відрізняється тим, що кількість кластерів з самого початку не відомо і визначається числом k. Характерними особливостями даного методу являються його простота в реалізації та модифікації. Проте методу також притаманні деякі недоліки – проблема збіжності, і, як наслідок, неможливості визначення часу, що необхідний для кластеризації даних. Роботу алгоритму можна умовно поділити на чотири основні етапи: визначення k центрів кластерів, визначення належності об'єктів кластерам, визначення центрів k-кластерів, порівняння.

Середовище розробки R та основні базові пакети надають можливість провести кластерний аналіз за допомогою алгоритму k-means, а графічні можливості візуалізують відображення результатів згідно принципам аналізу.

Результатом задачі кластеризації за допомогою алгоритму k-means є векторні набори даних, візуалізація яких графічно з вказівкою центрів k, які були знайдені в ході вирішення задачі відображають розподіл користувачів в залежності від характеристик аналізу та вказують для яких груп користувачів є можливість активно просувати сервіс та вирішувати задачі з прогнозування соціально-рекомендаційної складової для поліпшення проведення їхнього часу в мережі та ефективного виконання мережею поставлених цілей.

Також, вирішення даної задачі дає можливість дізнатись про розподіл цільової аудиторії, що може навести на думку впровадження додаткових функцій та методів мережі під конкретні групи користувачів. Загалом, результати мають стратегічні значення для розвитку продукту. Отже, мова та середовище R дає широкий спектр послуг, інструментів та пакетів для статистичного аналізу даних. Також, в вищенаведеному прикладі показується як можна підійти до середовища зі сторони Data Science та даних, які зберігаються на серверах.