Д.О. Павленко, магістр, гр. ПІ-45м, ФІКТ
Науковий керівник – к.т.н., доц. І.І. Сугоняк
Житомирський державний технологічний університет

АНАЛІЗ ДАНИХ ЗА ДОПОМОГОЮ МОВИ R

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 швидкодією до математичних пакетів GNU Octave або MATLAB. На даний час створено багато пакетів для статистичних обчислень, біоінформатики, оптимізації, кластерного та іншого аналізів.

Для візуалізації даних у середовищі наявні:

- двовимірні, тривимірні графіки;
- гістограми ;
- діаграми (схеми Ганта);
- звіти.

Для роботи з даними R підтримує безліч базових операцій, розглянемо простіші з них.

1. Арифметичні операції (+,-./.*.^). Також є в наявності елементарні функції: log( ), log10( ), exp( ), sin( ), cos( ), tan( ), sqrt( ), а також abs( ). Функція round(x,n) округлює число до n десяткових знаків після коми.
2. Логічні операції. Це операції: <, >, <= (менше або дорівнює); >= (більше або дорівнює); == (дорівнює); != (не дорівнює); & (переріз); | (об'єднання).
3. Статистичні операції:
   - mean(x) обчислює вибіркове середнє масиву x;
   - sd(x) обчислює вибіркове середньоквадратичне відхилення x;
   - var(x) обчислює вибіркову дисперсію масиву x;
   - summary(x) виводить елементи дескриптивної статистики масиву x: мінімальне
   - значення, максимальне, обидві квартилі, медіану і середнє.
   - range(x) повертає найбільше і найменше значення в x. Якщо нас цікавить різниця між
   - найбільшим і найменшим значеннями, можна скористатись функцією diff(range(x))
4. Вектор - це один з можливих типів даних. Створюємо вектор за допомогою операції c( ). Операції додавання, різниці, множення векторів відбуваються поелементно. Якщо ж вектори, що, наприклад, додаються, мають різні довжини, то коротший вектор «циклічно» продовжується до розміру довгого, і після цього проводиться додавання поелементно.
5. Послідовності. Команда seq( ) створює послідовність чисел. Її часто використовують при графічному аналізі. Три аргументи, які зазвичай використовують в команді: початкове значення, кінцеве і крок (приріст). Якщо ж приріст =1, то достатньо двох аргументів.
6. Списки. Список – це структура, тобто вектор, елементи якого можуть мати різні типи: числові, текстові і т.д. Елементом списку може бути інший список. Списки створюються за допомогою команди list().
7. Фрейми даних. Фрейм - найбільш широковживаний тип змінних в R, який використовується для зберігання даних. Фрейми складаються з різноманітних типів даних (числових, текстових, логічних і т.д.). Традиційно, стовпчики розглядаються як змінні, рядки містять характеристики об'єктів.

Для R написано багато різних бібліотек для відображення результатів аналізу.

Найбільш популярні:

- dplyr
- ggplot2 та ggthemes.
- 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.

Перед початком роботи із датафреймом треба зрозуміти його основні характеристики та характер даних. В цьому можуть допомогти функції:

- head(), tail() - показати відповідно початок та кінець датафрейму.
- str() - показати його структуру
- summary() - показати прості статистичні показники по кожному стовпчику
- view(), або клік на датафрейм у правому верхньому куті — переглянути датафрейм у Rstudio

Датафрейми мають дуже велике значення та розповсюдження в контексті обробки даних.

Розглянемо як R допомагає в аналізі даних на прикладі кластерного аналізу. Кластерний аналіз – це метод багатомірного статистичного дослідження, до якого належать збір даних, що містять інформацію про вибіркові об'єкти, та упорядкування їх в порівняно однорідні, схожі між собою групи.

Отже, сутність кластерного аналізу полягає у здійсненні класифікації об'єктів дослідження за допомогою численних обчислювальних процедур. В результаті цього утворюються "кластери" або групи дуже схожих об'єктів. На відміну від інших методів, цей вид аналізу дає можливість класифікувати об'єкти не за однією ознакою, а за декількома одночасно. Для цього вводяться відповідні показники, що характеризують певну міру близькості за всіма класифікаційними параметрами.

Мета кластерного аналізу полягає в пошуку наявних структур, що виражається в утворенні груп схожих між собою об'єктів – кластерів. Водночас його дія полягає й у привнесенні структури в досліджувані об'єкти. Це означає, що методи кластеризації необхідні для виявлення структури в даних, яку нелегко знайти при візуальному обстеженні або за допомогою експертів.

Особливо бурхливий розвиток кластерного аналізу відбувся у 60-х роках минулого століття. Передумовами цього були поява швидкісних комп'ютерів та визнання класифікацій фундаментальним методом наукових досліджень.

Основними завданнями кластерного аналізу є:

- розробка типології або класифікації досліджуваних об'єктів;
- дослідження та визначення прийнятних концептуальних схем групування об'єктів;
- висунення гіпотез на підставі результатів дослідження даних;
- перевірка гіпотез чи справді типи (групи), які були виділені певним чином, мають місце в наявних даних.

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.

При визначенні міри подібності об'єктів кластерного аналізу використовуються чотири види коефіцієнтів: коефіцієнти кореляції, показники віддалей, коефіцієнти асоціативності та ймовірносні, коефіцієнти подібності. Кожен з цих показників має свої переваги та недоліки, які попередньо потрібно врахувати. На практиці найбільшого розповсюдження у сфері соціальних та економічних наук здобули коефіцієнти кореляції та віддалей.

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. 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 модна умовно поділити на чотири основні етапи: визначення k центрів кластерів; визначення належності об’єктів кластерам; визначення центроїдів 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 відображують 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 проводження 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 аналізу даних.