1. Мета та завдання навчальної дисципліни

Метою викладання дисципліни є надання студентам знань з мови програмування R, яка призначена для статистичного аналізу, прогнозування та візуалізації даних. Провідні університети миру, аналітики компаній, банків та дослідницьких центрів використовують R при проведенні наукових розрахунків і розв'язанні задач прогнозної аналітики.

Завдання дисципліни полягає в засвоєнні студентами основ програмування в R, імпорту та експорту даних, очищення та перетворення даних, візуалізації даних, статистичного аналізу даних, побудови прогнозних моделей за допомогою алгоритмів машинного навчання.

У результаті вивчення дисципліни фахівець повинен

з н а т и : структури даних; умовні, логічні та циклічні конструкції; математичні, статистичні, текстові функції; синтаксис функції користувача; функції сімейства apply; функції сімейства map; функції імпорту даних з текстових файлів з роздільниками, функції імпорту даних з Excel, функції імпорту даних зі статистичних програм SAS, STATA, SPSS, функції імпорту даних з баз даних, функції імпорт даних з веб-сторінок; основи перетворення даних за допомогою функцій пакетів tydir, dplyr, data.table; функції візуалізації даних за допомогою пакетів ggvis, ggplot2; означення незміщеної, спроможної, ефективної оцінки, довірчі інтервали для невідомих параметрів, означення статистичної гіпотези, критерія, помилок першого та другого роду, рівня значущості критерію, функції потужності критерію, основні параметричні та непараметричні критерії для перевірки гіпотез про параметри розподілів, основні поняття дисперсійного, кореляційного та регресійного аналізу, алгоритми машинного навчання.

Підготовлений фахівець повинен

в м і т и : створювати вектори, матриці, фактори, таблиці даних списки та проводити з ними операції, складати умовні, логічні, циклічні конструкції, користуватися вбудованими функціями та складати власні функції користувача, проводити операції з датою, часом, регулярними виразами, імпортувати дані з текстових файлів з роздільниками, з Excel, зі статистичних програм SAS, STATA, SPSS, з баз даних, з веб-сторінок, керувати даними, а саме: упорядковувати дані, перетворювати один тип даних в інший, працювати з пропущеними значеннями, викидами та помилками в даних, керувати даними за допомогою функцій пакетів dplyr, data.table, візуалізувати дані, проводити статистичний аналіз даних: знаходити вибіркове середнє, дисперсію, знаходити точкові та інтервальні оцінки невідомих параметрів, використовувати критерії для перевірки статистичних гіпотез, використовувати алгоритми машинного навчання для побудови прогнозних моделей, вміти оцінювати якість прогнозних моделей та налаштовувати оптимальні параметри моделей.

2. Програма навчальної дисципліни

ЗМІСТОВИЙ МОДУЛЬ 1. Структури даних.

ТЕМА 1. Знайомство з R
Установка R, RStudio. Робочий простір R. Арифметичні оператори (+, -, *, /, **, %%, %/%). Оператор присвоювання. Основні типи даних в R: numeric, integer, logical, character, complex.
ТЕМА 2. Вектори Vectors.
Створення вектора за допомогою функцій c(), vector(). Задання імен елементів вектора за допомогою функції names(). Арифметичні операції з векторами. Вибір елементів вектора.
ТЕМА 3. Матриці Matrices
Створення матриці за допомогою функції matrix(). Створення матриці за допомогою об'єднання векторів. Задання імен стовбців та рядків матриці за допомогою функцій rownames(), colnames(). Функції rowSums(), colSums(). Додавання стовбців та рядків матриці за допомогою функції cbind(), rbind(). Вибір елементів матриці. Арифметичні операції з матрицями.
ТЕМА 4. Категоріальні змінні (фактори) Factors.

Створення фактору за допомогою функції factor(). Номінальні та порядкові фактори. Функція levels().
ТЕМА 5. Таблиці даних Data Frames.
Створення таблиці даних за допомогою функції data.frame(). Функції head(), tail(), str(). Вибір елементів з таблиці. Функція subset(). Сортування даних таблиці за допомогою функції order().
ТЕМА 6. Списки Lists.
Створення списку за допомогою функції list(). Задання імен елементів списку. Вибір елементів зі списку. Додавання елементів у список.

ЗМІСТОВНИЙ МОДУЛЬ 2. Конструкції та функції

ТЕМА 1. Умовні та логічні конструкції.
Оператори відношення (==, !=, <, >, <=, >=). Логічні оператори (&, |, !, &&, ||). Умовні оператори (if, else, else if, switch).
ТЕМА 2. Циклічні конструкції.
Цикл while. Цикл for. Цикл repeat, break.
ТЕМА 3. Вбудовані функції та створення функції користувача.
Математичні, статистичні, текстові функції. Створення функції користувача за допомогою my_fun <- function(arg1, arg2){body}. Встановлення пакетів install.packeges(), завантаження пакетів search(), library(), require().
ТЕМА 4. Функції сімейства apply.
Функції apply(), lapply(), sapply(), tapply(), mapply(), vapply() як альтернатива циклам. Використання анонімних функцій разом з функціями сімейства apply.
ТЕМА 5. Дати, час.
Функції Sys.Date(), Sys.time(). Класи POSIXct, POSIXlt. Функції as.Date() та as.POSIXct() добування дати та часу з текстових виразів. Функція format() для перетворення дат в заданий формат. Арифметичні дії з датою та часом.
ТЕМА 6. Регулярні вирази.
Функції grep(), grepl(), sub(), gsub().
ТЕМА 7. Функції сімейства map.
Пакет purrr. Функції map(), map_dbl(), map_lgl(), map_int(), map_chr(), map2(), pmap(), invoke_map() як альтернатива циклам. Використання анонімних функцій разом з функціями сімейства map.

ЗМІСТОВНИЙ МОДУЛЬ 3. Імпорт даних в R

ТЕМА 1. Імпорт даних з текстових файлів з роздільниками
Функції read.table(), read.csv(), read.delim(), read.csv2(), read.delim2().
Пакет readr, функції read_delim(), read_csv(), read_tsv().
Пакет data.table, функція fread().
ТЕМА 2. Імпорт даних з Excel
Пакет readxl, функції excel_sheets(), read_excel(). Пакет gdata, функція read.xls().
Пакет XLConnect, функції loadWorkbook(), getSheets(), readWorksheet(), createSheet(), writeWorkSheet(), saveWorkbook().
ТЕМА 3. Імпорт даних зі статистичних програм SAS, STATA, SPSS
Пакет haven, функції read_sas(), read_stata(), read_dta(), read_spss(), read_por(), read_sav().
Пакет foreign, функції read.dta(), read.spss().
ТЕМА 4. Імпорт даних з баз даних
Пакет DBI, функції dbConnect(), dbListTable(), dbReadTable(), dbGetQuery(), dbFetch(), dbDisconnect().
ТЕМА 5. Імпорт даних з веб-сторінок
Функція download.file().

ЗМІСТОВНИЙ МОДУЛЬ 4. Основи керування даними

ТЕМА 1. Вивчення структури даних
Функції class(), dim(), names, str(), glimpse(), summary(), head(), tail().
ТЕМА 2. Упорядкування даних за допомогою пакету tydir
Пакет tydir. Функції gather(), spread(), separate(), unite()
ТЕМА 3. Перетворення одного типу даних в інший
Функції as.numeric(), as.character(), as.vector(), as.matrix(), as.data.frame(), as.factor(), as.logical().
ТЕМА 4. Перетворення календарних дат
Функції as.Date(), as.POSIXct(). Пакет lubridate, функції ymd(), hms(), ymd_hms().
ТЕМА 5. Перетворення текстових даних
Пакет stringr, функції str_trim(), str_pad(), str_detect(), str_replace(), tolower(), toupper().
ТЕМА 6. Пропущені значення
Знаходження пропущених значень, заміна або виключення пропущених значень. Функції is.na(), complete.cases(), na.omit().
ТЕМА 7. Викиди та помилки в даних
Знаходження викидів та помилок в даних, заміна або виключення викидів. Функції boxplot(), summary(), hist().
ТЕМА 8. Керування даними за допомогою пакету dplyr
Вибір змінних за допомогою функції select(). Створення нових змінних за допомогою функції mutate(). Вибір спостережень за допомогою функції filter(). Впорядкування спостережень за допомогою функції arrange(). Обчислення статистик за допомогою функції summarise(). Групування даних за допомогою функції group_by(). Оператор покрокових обчислень %>% (pipe operator).
ТЕМА 9. Керування даними за допомогою пакету data.table

ЗМІСТОВНИЙ МОДУЛЬ 5. Візуалізація даних

ТЕМА 1. Візуалізація даних за допомогою пакету ggvis
ТЕМА 2. Візуалізація даних за допомогою пакету ggplot2

ЗМІСТОВНИЙ МОДУЛЬ 6. Статистичний аналіз даних

ТЕМА 1. Вступ до статистичного аналізу даних.
Вибірка, статистики: моменти, асиметрія і ексцес, варіаційний ряд і порядкові статистики, емпіричний розподіл. Точкові оцінки та їхні властивості. Інтервальні оцінки, поняття довірчого інтервалу та рівня довіри. Довірчі інтервали для середнього та медіани. Перевірка статистичних гіпотез, основні поняття: рівень значущості, p-value, помилки I та II роду.
ТЕМА 2. Параметрична перевірка гіпотез
Критерії нормальності: критерій хі-квадрат Пірсона, критерій Шапіро-Уілка, критерій Колмогорова-Смирнова (Лілліефорса). Спрощена перевірка нормальності за асиметрією та ексцесом: критерій Харке-Бера. Критерії для перевірки гіпотези про середні: t- і z-критерії Стьюдента для однієї і двох вибірок, зв'язані вибірки. Критерії для перевірки гіпотези про дисперсії: критерії хі-квадрат і Фішера. Критерії для перевірки гіпотези про значення параметра розподілу Бернуллі: порівняння значення параметра із заданим, порівняння параметрів розподілів двох вибірок (випадки зв'язаних і незалежних вибірок). Довірчий інтервал для параметра розподілу Бернуллі: Вальда, Уілсона. Довірчі інтервали Уілсона для різниці параметрів двох вибірок.
ТЕМА 3. Непараметрична перевірка гіпотез
Критерії знаків: одновибірковий, для зв'язаних вибірок. Рангові критерії: критерій Уілкоксона-Манна-Уітні, критерій Уілкоксона двовибірковий, критерій Уілкоксона для зв'язаних вибірок.

Перестановочні критерії. Двовибіркові критерії згоди: Колмогорова-Смірнова, Крамера-фон Мізеса (Андерсона).
ТЕМА 4. Дисперсійний аналіз.
ТЕМА 5. Аналіз залежностей.
ТЕМА 6. Лінійний регресійний аналіз.

ЗМІСТОВНИЙ МОДУЛЬ 7. Машинне навчання

ТЕМА 1. Класифікація.
Дерева рішень Decision Trees, метод найближчих сусідів k-Nearest Neighbors, логістична регресія Logistic Regression, випадковий ліс Random Forest, градієнтний бустінг XGBoost.
ТЕМА 2. Прогнозування.
Проста лінійна регресія Simple Linear Regression, множинна лінійна регресія Multivariable Linear Regression, метод найближчих сусідів k-Nearest Neighbors.
ТЕМА 3. Кластерізація.
Методу k середніх Clustering with k-means, ієрархічна кластерізація Hierarchical Clustering.
ТЕМА 4. Оцінка якості прогнозних моделей, налаштування оптимальних параметрів.

3. Методичне забезпечення

1. Роберт И.Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил.
2. Сергей Мастицкий, Владимир Шитиков Статистический анализ и визуализация данных с помощью R – 2014. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com
3. Алексей Шипунов, Евгений Балдин и др. Наглядная статистика. Используем R! 2014. – 296 с.
4. У. W. Н. Венэбльз, Д. М. Смит и Рабочая группа разработки R Введение в R. Заметки по R: среда программирования для анализа данных и графики. Пер. с англ. – Москва, 2013. – 109 с.
5. Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. – М.: Издательство Российского университета дружбы народов, 2010. – 207 с.
6. Зарядов И.С. Статистический пакет R: теория вероятностей и математическая статистика. – М.: Издательство Российского университета дружбы народов, 2010. – 141с.

4. Рекомендована література

Базова

1. Lander, Jared P. R for Everyone: Advanced Analytics and Graphics – Pearson Education, Inc.? 2014. – 426 p.
2. Kabacoff Robert I. R in Action: Data analysis and graphics with R – Manning Publications Co., 2011. – 447 p.
3. Crawley, Michael J. The R book. – John Wiley & Sons, Ltd., 2013. – 975 p.

Допоміжна

1. Gareth James et al. An Introduction to Statistical Learning: with Applications in R – Springer, 2013. – 426 p.
2. Brett Lantz Machine Learning with R – Packt Publishing, 2013. – 375 p.
3. Yanchang Zhao R and Data Mining: Examples and Case Studies – Elsevier, 2012. – 150 p.
4. Graham Williams Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!) – Springer, 2011. – 374 p.

5. Hadley Wickham ggplot2: Elegant Graphics for Data Analysis (Use R!) – Springer, 2009. – 213 p.
6. Norman Matloff The Art of R Programming : A Tour of Statistical Software Design – No Starch Press, 2011. – 400 p.