Лабораторная работа 1
Знакомство с языком R

1. Установить и запустить программу R.
2. Настроить рабочую директорию с помощью функции setwd().
3. Сгенерировать вектор X~U(0, 1) и построить гистограмму.
4. Сохранить результаты моделирования в текстовый файл (*.txt, *.csv), гистограмму в графический файл (*.jpg, *.svg) с использованием функций записи в файл.
5. Познакомиться с основными структурами данных numeric(), matrix(), data.frame(), list().
6. Познакомиться со справочной системой.
7. Установка и загрузка дополнительных пакетов с помощью функций install.packages() и library().

Литература

1. https://ru.stackoverflow.com/questions/506597/Книги и учебные ресурсы по языку R

Лабораторная работа 2
Стратегии работы с большими массивами данных

1. Сгенерировать большой массив данных и записать в один файл [1]. Установить пакет purrr. Записать массив данных по частям в несколько файлов [2]. Сформировать репрезентативная выборку ограниченного размера.
2. Выполнить загрузку данных с использованием различных стратегий [1]. Сделать выводы. Установить пакеты data.table, sqldf, ff.
3. Установить и загрузить библиотеки sqldf и nycflights13.Ознакомиться со структурой набора данных flights. Вычислить количество наблюдений для всех перевозчиков carrier в таблице flights. Отобразить в консоли значения полейdep_time,dep_delay,arr_time, carrier, tailnum из таблицы flights (первые и последние 5 строк). Вычислить среднее время задержки прибытия (mean_arr_delay) и отправления (mean_dep_delay) для различных перевозчиков (carrier) [3].
4. Сгенерировать data.frame с тремя столбцами и 100 строками. Преобразовать данные из широкого в длинный формат. Установить пакет reshape2 [4].

Литература

1. https://stackoverflow.com/questions/1727772/quickly-reading-very-large-tables-as-dataframes/
2. https://stackoverflow.com/questions/57047338/split-dataset-per-rows-into-smaller-files-in-r
3. https://www.tutorialspoint.com/big_data_analytics/introduction_to_sql.htm
4. Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R (2014).

Лабораторная работа 3
Подготовка исходных данных

1. Сгенерировать вектор (массив, таблица данных) и добавить в него элементы NA. Очистить данные с использованием функцииis.na() [1].
2. Сгенерировать таблицу данных с числовыми и текстовые столбцами. Очистить данные с функции complete.cases() [1].
2. Сгенерировать числовую таблицу данных с пропусками. С использованием функции preProcess из пакета caret заполнить пропуски предсказанными значениями (среднее, медиана) [2].
3. Сгенерировать два числовых набора данных, добавить в них выбросы. С использованием функции boxplot обнаружить выбросы и удалить их [3, 4].
4. Сгенерируйте таблицу данных, в которой дублируются строки. Удалите строки с использованием функций unique(), duplicated(). Сравните результаты [5].

5. Обработать пропуски в данных с использованием пакета mice [6].
6. Разобрать пример с мультиколлинеарностью [7].

Литература

1. http://datascientist.one/removing-na-values-r/
2. https://r-analytics.blogspot.com/2017/01/blog-post.html
3. http://datascientist.one/delete-outliers-with-boxplot-r/
4. https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/
5. https://stackoverflow.com/questions/13967063/remove-duplicated-rows
6. https://habr.com/ru/company/infopulse/blog/305692/
7. https://datascienceplus.com/multicollinearity-in-r/

Лабораторная работа 4
Обработка данных. Выбор признаков (Feature Selection)

1. Установить пакет CARET, выполнить команду names(getModelInfo()), ознакомиться со списком доступных методов выбора признаков. Выполните графический разведочный анализ данных с использование функции featurePlot() для набора данных из справочного файла пакета CARET:

```
x <- matrix(rnorm(50*5),ncol=5)
y <- factor(rep(c("A", "B"), 25))
```

Сохранить полученные графики в *.jpg файлы. Сделать выводы.

2. С использование функций из пакета Fselector [2] определить важность признаков для решения задачи классификации. Использовать набор data(iris). Сделать выводы.
3. Установите пакет Boruta и проведите выбор признаков для набора данных data("Ozone") [3, 4]. Построить график boxplot, сделать выводы.

Литература

1. https://topepo.github.io/caret/train-models-by-tag.html#implicit-feature-selection
2. https://miningthedetails.com/blog/r/fselector/
3. https://www.jstatsoft.org/article/view/v036i11/v36i11.pdf
4. https://www.datacamp.com/community/tutorials/feature-selection-R-boruta
5. https://habr.com/ru/post/264915/
6. http://ai.stanford.edu/~ronnyk/wrappersPrint.pdf

Лабораторная работа 5
Обработка данных. Выбор экземпляров (Instance Selection)

1. Выполните классификацию k-ближайших соседей с использованием функции knn() из пакета class на наборе данных iris [1]. Проведите нормализацию данных, разделите выборку на обучающую и тестовую. Оцените построенную модель с использованием функции CrossTable() из пакета gmodels. Постройте матрицу ошибок [2] и диагональную оценку качества прогноза (*diagonal mark quality prediction*).
2. Рассмотрите пример реализации метода опорных векторов с использованием функции svm() из пакета e1071. Постройте линейный классификатор для прогнозирования. Для подбора параметров модели выполните перекрестную проверку с делением исходной выборки на 10 равных частей (cross=10) [3, с.172].
3. Выполните расчет главных компонент с использованием пакета vegan() и его функции rda(). Постройте ординационную диаграмму методом PCA [3, с. 49] и сделайте выводы.

Литература

1. https://en.proft.me/2017/01/22/classification-using-k-nearest-neighbors-r/
2. https://habr.com/ru/company/ods/blog/328372/
3. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. – Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining
4. Olvera-López, José & Carrasco-Ochoa, Jesús & Martínez-Trinidad, José Francisco & Kittler, Josef. (2010). A review of instance selection methods. Artif. Intell. Rev. 34. 133-143. 10.1007/s10462-010-9165-y. https://mafiadoc.com/a-review-of-instance-selection-methods-soft-computing-and-_5b054f698ead0ed4758b4586.html
5. Top 10 algorithms in data mining http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

Лабораторная работа 6
Обработка данных. Дискретизация для классификации (Discretization)

1. С использованием функции discretize() из пакета arules выполните преобразование непрерывной переменной в категориальную [1] различными методами: «interval» (равная ширина интервала), «frequency» (равная частота), «cluster» (кластеризация) и «fixed» (категории задают границы интервалов). Используйте набор данных iris. Сделайте выводы.
2. С использованием пакета discretization выполните дискретизацию с использованием алгоритмов Chi2 и CAIM [2]. Используйте набор данных iris. Сравните результаты и сделайте выводы.
3.

Литература

1. http://finzi.psych.upenn.edu/library/arules/html/discretize.html
2. https://cran.r-project.org/web/packages/discretization/index.html

Лабораторная работа 7
Организация распределённых вычислений

1. Установите пакет sparklyr, установите Java Virtual Machine (JVM). Подключитесь к локальному Spark-кластеру. Загрузите таблицу flights из пакета nycflights13 в Spark-кластер [1]. Выполните запросы (задание 3, Лабораторная работа 2). Сравните результаты, сделайте выводы.
2. Настройте для использования Hadoop [2-5], подсчитайте количество слов в файле *.txt с использованием HDFS [3]. Файл сгенерировать самостоятельно.

3. Установите MongoDB [6, 7]. Подключите библиотеку mongolite. Выполните пример для набора iris с использование функции mongo() из видеоролика [7]. Сохраните код и сделайте выводы.

Литература

1. https://r-analytics.blogspot.com/2020/02/spark-r-connect.html
2. 4 Ways To Use R And Hadoop Together https://www.edureka.co/blog/4-ways-to-use-r-and-hadoop-together/
3. http://www.rdatamining.com/big-data/r-hadoop-setup-guide
4. https://github.com/jeffreybreen/hadoop-R
5. Video: Using R with Hadoop https://www.r-bloggers.com/video-using-r-with-hadoop/
6. https://data-flair.training/blogs/mongodb-tutorials-home/
7. Connect to MongoDB Database in R https://www.youtube.com/watch?v=JBEKJflNV2g
8. https://www.blue-granite.com/blog/using-hadoop-data-r-distributed-machine-learning
9. https://data-flair.training/blogs/r-hadoop-integration/

Лабораторная работа 8
Практические задачи
с использованием различных инструментов обработки больших данных

Выберите два любых кейса [1]. Опишите входные данные, стек моделей и технологий, которые можно использовать для решения выбранных кейсов. Приведите иллюстративные примеры с использованием R. Сделайте выводы.

Литература

1. https://data-flair.training/blogs/big-data-case-studies/
2. https://data-flair.training/blogs/big-data-use-cases-case-studies-hadoop-spark-flink/