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/