**ВВЕДЕНИЕ**

Выполнение лабораторной работы по дисциплине предполагает:

1. ознакомление с теоретическим материалом по теме работы,
2. выполнение работы согласно выданному индивидуальному заданию,
3. оформление отчета в Word и размещение его вместе с файлами, содержащими код на языке программирования и набор данных, подвергающихся исследованию описанным методом в системе Moodle по расписанию;
4. защита выполненной работы в аудитории.

При приёме защиты лабораторной работы преподаватель проверяет:

1) наличие исходного кода программы, реализующего требуемый метод анализа данных;

2) правильность работы программы для заданных наборов данных;

3) понимание студентом сути проведённых исследований и правильность сделанных выводов;

4) знание студентом ответов на контрольные вопросы;

5) наличие отчёта, оформленного в соответствии с требованиями к учебным текстовым документам.

Отчёт о выполнении лабораторной работы должен содержать следующие разделы:

1. Титульный лист с указанием названия и автора работы, а также преподавателя, проверяющего работу.

2. Задание на лабораторную работу.

3. Краткое описание теоретической части работы, включающее некоторые формулы и утверждения, лежащие в основе исследуемого метода.

4. Описание исходных данных для работы, их природы и происхождения.

5. Полученные результаты, включая их графическое представление в форме таблиц, диаграмм и рисунков.

6 Заключение, включающее выводы, сделанные по результатам работы.

## Лаб.р.1 Знакомство со средой для интеллектуального анализа данных R.

***Цель***: Ознакомиться с интерфейсом RStudio, научиться работать в режиме консоли и путем написания скриптов, а также подключать внешние пакеты, изучить основные методы обработки статистических данных.

***Теоретические сведения***

Начало работы и получение справочной информации [2, c.19-34].

Для установки среды R и Rstudio на свой домашний компьютер воспользуйтесь инструкцией по ссылке, приведенной в сноске[[1]](#footnote-1).

Директория – это место (папка), где находится ваш «проект». То есть там лежит скрипт, данные, картинки и прочее.

Зачастую, у каждого проекта своя директория, поэтому приходится часто их менять. Есть несколько способов это сделать.

1 Использовать функцию setwd("~/Desktop/R"), где в кавычках можно прописать путь к директории.

2 На панели R нажать кнопку Session и в выпадающем списке выбрать Set Working Directory. После этого можно выбрать нужную папку.

3 В нижнем правом окошке есть вся файловая система компьютера, где так же удобно можно установить нужную директорию. Нажав кнопку More, можно перейти в текущую директорию или установить новую.

С целью дальнейшего упрощения работы рекомендуется заранее установить некоторые основные пакеты с помощью функции install.packages набрав команду: install.packages(«название пакета»), или, войдя во вкладку Packages/Install Packages, затем подключить нужный пакет. Для работы чаще всего используются следующие пакеты:

psych – описательная статистика;

ggplot2 – графика;

lmtest – тестирование гипотез при построении линейных моделей;

MASS – поиск подходящих распределений;

dplyr – вычислительные операции с данными.

После загрузки всех необходимых пакетов нужно узнать в какой директории вы находитесь. Это можно сделать с помощью функцию getwd().

Импорт данных – это загрузка в среду R различных данных для последующей работы с ними. Важно понимать откуда мы берём данные. Обычно источником импорта выступают: обычные текстовые файлы (.csv, .txt),   
Excel-файлы (.xlsx, .xls), базы данных (SQL), интернет, статистические пакеты (SPSS, SAS, STATA).

Импорт файлов с расширениями csv, txt осуществляется с помощью функции read.csv(). Формат csv самый распространенный формат хранения данных в мире для анализа данных. Его можно получить, например, сохранив Excel-файл в формате csv (разделитель – запятая).

Проверка того, что все данные импортировались нормально, осуществляется с помощью функции str().

Экспорт данных – это сохранение данных с требуемым расширением на свой компьютер. Для экспорта данных применяют функции: write.csv(df, "table\_car.csv") или write.xlsx(df, "table\_car.xlsx") или write.html(df, "table\_car.html"), где df – имя фрейма данных в среде R.

***Задание***

1 Инсталлировать пакеты для ввода данных в различных форматах.

2 Изучить возможность получать справочную информацию по необходимым пакетам.

3 Создать свою директорию.

4 Загрузить данные для своего варианта, полученные у преподавателя в переменную.

5 Получить справочную информацию по своим данным, просмотреть их содержимое.

6 Проверить, есть ли среди данных пропуски, удалить с помощью соответствующей команды.

7 Изучить типы данных.

8 Выполнить экспорт очищенных данных в MS Excel.

9 Составить отчет.

***Контрольные вопросы***

1 Какие типы данных используются в среде R?

2 Как начать работу в R-Studio?

3 Как получить справочную информацию по пакетам и встроенным в среду R датасетам?

4 Как импортировать в среду R файлы с различными расширениями?

5 Как экспортировать из среды R файлы с различными расширениями?

## Лаб.р.2 Разведывательный анализ данных средствами R

***Цель***: провести простой анализ данных, состоящий из описательной статистики и визуализации.

***Теоретические сведения***

Разведочный анализ данных основан на построении визуализаций и вычислении характеристик описательной статистики, представленных в таблице 1.

Таблица 1 – Параметры описательной статистики

|  |  |
| --- | --- |
| Функция | Описание |
| mean() | Среднее значение |
| median() | Медиана |
| var() | Дисперсия |
| sd() | Стандартное отклонение |
| min() | Минимальное значение |
| max() | Максимальное значение |
| quantile() | Квантили (по умолчанию рассчитывается максимальное и минимальное значения, а также квартили) |
| summary() | Сводка по параметрам описательной статистики для всех переменных набора |

Среди данных могут быть пропуски, которые не позволят рассчитать статистические параметры для столбцов количественных переменных с пропусками без специальных настроек.

Функция is.na() возвращает набор данных (вектор, таблицу и т. п.) заполненный значениями true (если значение отсутствует) и false (при его наличии). Так как true маркируется 1, а false маркируется нулем, то функция sum(), примененная к такому преобразованному набору данных, дает число пропусков.

Для построения графиков используют высокоуровневые функции, доступные без подключения специализированных пакетов для построения визуализаций (таблица 2).

Таблица 2 – Функции для построения графиков

|  |  |
| --- | --- |
| Функция | Описание |
| plot() | Общая функция для построения графиков. В зависимости от передаваемых данных и настройки параметров, определяемая ей визуализация может быть и точечным графиком, и столбчатой диаграммой |
| boxplot() | График «ящик с усами» |
| hist() | Гистограмма |
| barplot() | Столбчатая диаграмма |
| scatterplot() | Диаграмма рассеяния (точечная) |
| pie() | Круговая диаграмма |

Для построения гистограммы с помощью библиотеки ggplot2 воспользоваться [3].

Команда Glimpse(f) (данная функция из ранее уставленного пакета «dplyr») позволяет посмотреть краткие сведения о наборе данных. Вывод на экран описательной статистики осуществляется также с помощью функции «describe», относящейся к пакету «psych».

***Задание***

1 Импортировать данные, полученные у преподавателя. согласно своему варианту.

2 Создать новую переменную-вектор, в которой будут 1, если значение в исходном векторе больше среднего, и -1, если значение переменной меньше среднего, и 0, если значение равно среднему c команды ifelse.

3 Вывести описательную статистику.

4 Построить графики абсолютных частот и плотности распределения с помощью базового пакета и библиотеки ggplot2.

5 Оформить отчет.

***Контрольные вопросы***

1 Как создать количественную переменную в среде R?

2 Как создать вектор качественных переменных в среде R?

3 Как рассчитать показатели описательной статистики?

4 Как построить гистограмму и график плотности распределения в среде R?

5 Какие типы данных вам известны?

6 Что такое генеральная совокупность? Что такое выборочная совокупность?

7 Что такое распределение случайной величины?

8 Методы предобработки данных (удаление дубликатов, обработка пропусков, выявление аномальных наблюдений)

9 Как визуализировать распределение количественных данных в R?

10 Как визуализировать распределение категориальных данных в R?

11 Перечислите числовые характеристики выборочной совокупности:

12 Какие параметры описывают нормальное распределение?

13 Что такое медиана и процентили, квантили, мода?

14 Что такое выбросы? Как определить их наличие в данных визуально в R?

## Лаб.3 Проверка гипотез

***Цель***: научиться формулировать статистические гипотезы и оценивать вероятности ошибок принятия гипотезы, проверять статистические гипотезы согласно алгоритму, делать выводы по результатам статистической проверки.

***Теоретические сведения***

Для освоения навыков проверки статистических гипотез в среде R воспользоваться материалом [3, с.231-246, 4, с.151-176].

***Алгоритм проверки статистических гипотез***

1 По выборочным данным формулируют основную *H0* и альтернативную *H1* гипотезы.

2 Задают уровень значимости (0,05 или 0,01).

3 В зависимости от *H0* определяют статистический критерий , имеющий известное распределение.

4 По выборке и формуле критерия рассчитывают наблюдаемое значение критерия .

5 В зависимости от вида *H1* определяют вид критической области и критические точки по соответствующим таблицам для распределения критерия .

6 По результатам проверки принадлежности к критической области делают вывод о принятии или отклонении гипотезы *H0*. Формулируют общий вывод, исходя из поставленной задачи.

В зависимости от имеющейся у исследователя информации, гипотезы классифицируются следующим образом(таблица 3-4).

Таблица 3 – Функции для тестирования параметрических гипотез

|  |  |
| --- | --- |
| Функция | Описание |
| Одновыборочные критерии о равенстве числовому параметру | |
| t. test (*x*,) | математического ожидания |
| prop.test(*x*,) |  |
| критерии о равенстве числовых характеристик | |
| var.test (*x,y*) | дисперсий двух групп |
| t. test (*x*,*y*) | математических ожиданий двух групп |
| prop.test(*x*, *y*) | долей двух групп |
| aov() | математических ожиданий в нескольких группах |
| bartlett.test () | дисперсий в нескольких группах |

Таблица 4 – Функции для тестирования непараметрических гипотез

|  |  |
| --- | --- |
| Функция | Описание |
| проверка гипотезы о законе распределения | |
| shapiro.test(*x*) | о согласии с нормальным законом распределения |
| критерии об однородности | |
| wilcox.test(*x*,*y*,paired=*F*) | критерий Манна-Уитни для двух независимых выборок |
| wilcox.test(*x*,*y*,paired=T) | критерий Вилкоксона для двух зависимых выборок |
| kruskal.test() | критерий Краслелла Уоллеса о равенстве распределений в нескольких группах. |

***Задание***

1 Получить данные у преподавателя согласно своему варианту.

2 Описать исходную совокупность с помощью показателей дескриптивной статистики

3 Построить гистограммы для количественных признаков и ящик с усами для исходного набора данных и выдвинуть предположение о функции распределения

4 Проверить гипотезу о нормальном распределении с помощью различных критериев.

5 Протестировать гипотезу об однородности и о различии в типичном значении выбранного признака выборки согласно варианту.

6 Оформить отчет.

***Контрольные вопросы***

1. Что такое нулевая гипотеза?
2. Что называется уровнем значимости (*P-value*)?
3. Какие параметры определяют нормальное распределение?
4. Какая нулевая гипотеза может быть протестирована на основании гистограммы?
5. Как проверяется гипотеза о соответствии выборке нормальному распределению в R?
6. Приведите пример независимых выборок.
7. Приведите пример зависимых выборок.
8. Какая нулевая гипотеза может быть протестирована на основании boxplot()?

10 Зачем проверять гипотезу о равенстве дисперсий в двух выборках, извлеченных из нормально распределенной генеральной совокупности? (*F*-критерий)?

11 Как протестировать гипотезу об однородности двух выборок в зависимости от законов распределения генеральных совокупностей, из которых они были извлечены ( *T*-критерий, Манна- Уитни, Вилкоксона)?

## Лаб.4 Регрессия как задача машинного обучения с учителем

**Цель:** изучение возможности построения уравнения регрессии для прогнозирования непрерывного результативного признака.

***Теоретические сведения***

Уравнение регрессии – функция, позволяющая по величине изменения одного корреллируемого признака определить среднюю величину другого признака. Выделим основные этапы регрессионного анализа.

*Первый этап*: Предположение. На этом этапе происходит выбор формы связи между переменными (модель).

*Второй этап*: Параметризация – происходит оценка значений параметра в выбранной формуле статистической связи. Форма связи (функция) линейная, нелинейная.

*Третий этап*: Проверка надёжности полученных оценок. На этом этапе осуществляются следующие тесты: F-тест (проверка статистической значимости выбранной формы связи), t-тест (проверка статистической значимости найденных числовых значений параметра). В результате анализа статистических данных, выбора и построения модели последовательно выполняются все три этапа.

В регрессионном анализе в машинном обучении для оценки качества предсказаний применяется несколько метрик.

1. Средняя абсолютная ошибка (Mean Absolute Error, MAE) –представляет собой среднее абсолютное значение разности между фактическими и предсказанными значениями.

2. Среднеквадратичная ошибка (Mean Squared Error, MSE) – измеряет среднее значение квадрата разности между фактическими и предсказанными значениями.

3. Корень из среднеквадратичной ошибки (Root Mean Squared Error, RMSE) –представляет собой квадратный корень из MSE, предоставляет интерпретируемую метрику, измеряющую среднее абсолютное отклонение модели.

4. Коэффициент детерминации (R-squared, R2) –измеряет долю дисперсии зависимой переменной, которая объясняется моделью, показывает, насколько хорошо модель соответствует данным. Значения R2 лежат в диапазоне от 0 до 1, где 1 указывает на идеальное соответствие.

***Задание***

1. *Сбор данных*. Собрать набор данных, который содержит как обучающие, так и тестовые данные. Обучающие данные будут использоваться для обучения модели, а тестовые данные - для оценки ее производительности.

2. *Предварительная обработка данных*. Перед обучением модели данные должны быть предварительно обработаны. Это может включать в себя удаление выбросов, заполнение пропущенных значений, масштабирование данных и т.д.

3. *Выбор модели.* Выберите модель регрессии, которая лучше всего подходит для вашего набора данных. Некоторые из популярных моделей регрессии включают линейную регрессию, полиномиальную регрессию, регрессию на основе деревьев и т.д.

4. *Обучение модели*. Используйте выбранную модель для обучения на обучающих данных. Это может включать в себя настройку гиперпараметров модели и оптимизацию функции потерь.

5. *Оценка модели*. Оцените производительность модели на тестовых данных. Это может включать в себя расчет метрик, таких как среднеквадратичная ошибка (RMSE), средняя абсолютная ошибка (MAE) и коэффициент детерминации (R2).

6. *Анализ результатов*. Проанализируйте результаты и сделайте выводы о производительности модели. Если производительность модели неудовлетворительна, вы можете рассмотреть возможность использования другой модели или дополнительной предварительной обработки данных.

7. *Применение модели*. Если производительность модели удовлетворительна, вы можете использовать ее для предсказания значений для новых данных.

**Контрольные вопросы**

1.Что такое задача регрессии в контексте машинного обучения?

2 Какие основные отличия между задачами регрессии и классификации?

3 Назовите основные метрики, используемые для оценки качества моделей регрессии.

4 Какая метрика будет значима для задачи, где большие ошибки более критичны, и почему?

5 Что такое переобучение и недообучение в контексте моделей регрессии?

6. Какие методы можно применить для борьбы с переобучением модели регрессии?

7 Какую информацию о датасете может передать среднеквадратичная ошибка (MSE) модели регрессии?

8 Что означает значение коэффициента детерминации равное 1.0?

9 Почему важен процесс отбора признаков в моделях регрессии?

10 Какие методы отбора признаков могут быть применены для моделей регрессии и какие у них преимущества?

## Лаб.5 Методы кластеризации

***Цель***: выработка практических навыков проведения многомернй классификации методами кластерного анализа средствами R и последующего анализа результатов.

***Теоретические сведения***

Для освоения навыков выполнения кластерного анализа в среде R воспользоваться материалом [5, с.51-71].

**Постановка задачи кластеризации** Пусть имеется выборка *U* из *n* объектов наблюдения, каждый из которых характеризуется m числовыми признаками *X*j. Известно, что каждый из этих объектов на самом деле относится к одному из *k* классов, причём признаки объектов из одного класса не слишком сильно различаются, а признаки объектов из разных классов различаются более существенно. Мера различия между признаками объектов может быть определена как расстояние *ρ*(*x*, *y*) между векторами признаков в соответствующем признаковом пространстве.

**Алгоритм k внутригрупповых средних.**

Идея алгоритма *k* внутригрупповых средних (англ. *k-means*) заключается в последовательном пересчёте внутригрупповых средних. Пусть вначале для каждого из *k* кластеров имеется некоторое начальное внутригрупповое среднее . Если внутригрупповые средние заданы, то каждый кластер *U*l очевидным образом определяются, как множество объектов наблюдения, вектора признаков для которых ближе к центру, чем к центрам других кластеров. После этого внутригрупповое среднее можно пересчитать, после чего снова пересчитать кластеры, пока средние не перестанут меняться.

В языке *R* для кластеризации методом *k* внутригрупповых средних используется функция kmeans() из пакета stats*.* Она позволяет производить кластеризацию данных различными методами.

Ниже приведена сигнатура этой функции.

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE) ,

где

x – матрица данных, каждая строка которой является вектором признаков очередного объекта наблюдения,

centers – количество кластеров k или набор начальных центров кластеров. Если набор начальных центров не задан, то в качестве начальных центров выбираются k случайных объектов наблюдения.

iter.max – максимальное количество итераций алгоритма.

nstart – количество наборов случайных начальных центров кластеров в случае, если центры выбираются случайно.

algorithm – реализация алгоритма k внутригрупповых средних, с помощью которой будет производиться кластеризация.

Функция возвращает объект класса kmeans, который имеет методы print и fitted. Этот объект представляет собой список, содержащий следующие элементы:

cluster – вектор из n целых положительных чисел, обозначающих номер кластера соответствующего объекта наблюдения;

centers – матрица, строки которой представляют собой центры соответствующих кластеров;

totss – сумма квадратов расстояний от объектов наблюдения до соответствующих центров кластеров;

withinss – вектор внутрикластерных сумм квадратов расстояний между объектами одного кластера для каждого кластера;

tot.withinss – общая сумма внутрикластерных сумм квадратов расстояний между объектами, то есть сумма элементов вектора withinss;

betweenss – междукластерная сумма квадратов расстояний, то есть разница между totss и tot.withinss;

size – количество объектов наблюдения, попавших в каждый из кластеров; iter – количество внешних итераций, совершённых в ходе работы алгоритма. ifault – код ошибки, возникшей в ходе работы алгоритма.

Доступ к этим значениям можно получить путём записи их названия в двойных квадратных скобках справа от переменной, содержащей модель, либо через знак доллара. Например, «a[["cluster"]]» или «a$cluster».

**Иерархическая кластеризация**

Часто задача кластеризации, описанная в разделе 3.1, возникает в несколько более сложной постановке: точное число кластеров k может быть не известно. В этом случае использование алгоритма k внутригрупповых средних не представляется возможным. Конечно, можно просто перебирать значения k и для каждого из них использовать этот алгоритм, но сравнение среднеквадратических расстояний до центров кластеров для различных количеств кластеров лишено смысла.

Вместо этого используются алгоритмы иерархической кластеризации (англ. hierarchical clustering), которые разбивают множество объектов наблюдения на кластеры, которые в свою очередь также разбиваются на кластеры и так далее. Получившийся граф, вершинами которого являются кластеры, а дуги направлены от кластеров более высокого уровня к соответствующим кластерам более низкого уровня, называется дендрограммой.

Построенная дендрограмма содержит информацию о близости между отдельными подмножествами объектов наблюдения, что позволяет подбирать число кластеров и производить более тщательный кластерный анализ в отдельных случаях. Иерархическая кластеризация, как и обыкновенная, производится на основе расстояний ρ(x, y) между объектами наблюдения из заданной выборки. Эти расстояния также могут быть определены по-разному. На их основе строятся более сложные показатели качества кластеризации, которые допускают сравнение для различного числа кластеров. Далее с помощью различных методов, основанных на последовательном объединении и разбиении имеющихся кластеров, строится дендрограмма.

Существует ряд различных методов иерархической кластеризации:

1. Метод Уорда (англ. Ward’s method).

2. Метод одиночной связи (англ. single linkage).

3. Метод полной связи (англ. complete linkage).

4. Метод средней связи (англ. pair-group method using arithmetic averages).

5. Метод Мак-Куитти (англ. McQuitty’s method).

6. Метод медиан.

***Задание***

Входные данные: n объектов, каждый из которых характеризуется двумя числовыми признаками {*xi*} и {*yi*}. Требуется исследовать работу алгоритмов кластеризации объектов наблюдения по двум признакам. Для каждого набора данных требуется выполнить следующие задания.

1 Провести кластеризацию объектов наблюдения с помощью алгоритма *k* внутригрупповых средних. Выбрать оптимальное количество кластеров, исходя из критерия минимизации внутригрупповых дисперсий и максимизации расстояния между центрами кластеров.

2 Графически изобразить на плоскости разбиения объектов наблюдения в соответствии с кластерами и в соответствии с классами *ci*. Также отметить центры каждого кластера. Количество кластеров должно соответствовать количеству классов.

3 Для разбиения на кластеры вычислить сумму квадратов расстояний от каждого объекта наблюдения до центра соответствующего кластера.

4 Провести кластеризацию исходных данных иерархическим способом . Сравнить результаты кластеризации на тестовом множестве.

Все описанные задания требуется выполнить для двух наборов данных:

1) Смоделированные независимые случайные векторы (*X*, *Y*), *n1* из которых относятся к первому классу, а *n2* – ко второму классу. Векторы, относящиеся к первому классу, распределены по гауссовскому закону с математическим ожиданием *a*1 и корреляционной матрицей *R1*, а векторы, относящиеся ко второму классу, – по гауссовскому закону с математическим ожиданием *a2* и корреляционной матрицей *R2*.

2) Реальные статистические данные из заданного набора (выдаются преподавателем).

Отчёт кроме прочих обязательных элементов должен включать:

1. изображения данных в виде точек на плоскости, причём данные из разных классов должны изображаться отличающимися друг от друга;
2. изображения результатов кластеризации данных в виде точек на плоскости, причём данные из разных кластеров должны изображаться отличающимися друг от друга, кроме того, требуется отметить центры кластеров;
3. значения суммы квадратов расстояний от каждого объекта наблюдения до центра соответствующего кластера.

***Варианты заданий на лабораторную работу***

Все описанные задания требуется выполнить для двух наборов данных.

1. В таблице 1 (Приложения A ) для каждого варианта задания1 приведены значения количеств объектов наблюдения для каждого класса (*n*1 и *n*2), значения векторов математических ожиданий для каждого класса (*a*1 и *a*2) и корреляционные матрицы для каждого класса (*R*1 и *R*2) для моделируемой выборки из гауссовских случайных векторов.

**2*.*** В таблице2 представлены варианты заданий для задания 2. Варианты реальных наборов данных, расположенных в папке DATA на странице курса в Moodle:

Таблица 2– Варианты заданий для кластерного анализа

|  |  |  |
| --- | --- | --- |
| Вариант | Название файла | Описание данных |
| 1 | 2 | 3 |
|  | Soybean (01-soybean.txt) | Первый признак: protein (столбец № 10),второй признак: oil (столбец № 11), класс: loc (столбец № 3) |
|  | Данные о новорожденных  (02-birth.txt) | Первый признак: Head (столбец № 6), второй признак: Chest (столбец № 7), класс: Sex (столбец № 2) |
|  | Потребление пищи в Дании в 1985 году (03-vitamina.txt) | Первый признак: Avit (столбец № 9), второй признак: Cvit (столбец № 20), класс: sex (столбец № 4) |
|  | Physical Growth of California Boys and Girls  (04-physical-growth.txt) | Первый признак:  WT2 (столбец № 3), второй признак: WT9 (столбец № 5),  класс: Sex (столбец № 2) |
|  | Успеваемость студентов  (05-trp-grades.txt) | Первый признак: L.att (столбец № 2), второй признак: P.att (столбец № 3),класс: Exam (столбец № 8) |
|  | Wine (06-wine.txt ).  Ссылка: http://archive.ics.uci.edu/ml/datasets/Wine | Первый признак: alcohol (столбец № 2), второй признак: color-intensity (столбец № 11), класс: cultivar (столбец № 1). |
|  | Forest Fires (07-forest-fires.txt) Ссылка: http://archive.ics.uci.edu/ml/datasets/Forest+Fires | Первый признак: ISI (столбец № 8), второй признак: wind (столбец № 11), класс: month (столбец № 3) |
|  | Wisconsin Prognostic Breast Cancer (08-wpbc.txt).  Ссылка: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) | Первый признак: Первый признак: perimeter (столбец № 6), второй признак: area (столбец № 7), класс: Outcome (столбец № 2) |
|  | Abalone Data (09-abalone.txt)  Ссылка:  http://archive.ics.uci.edu/ml/datasets/Abalone | Первый признак: Length (столбец № 2), второй признак: Diameter (столбец № 3), класс: Sex (столбец № 1)67 |
|  | 1985 Auto Imports Database (10-automobile.txt).  Ссылка:  http://archive.ics.uci.edu/ml/datasets/Automobile | Первый признак: length (столбец № 11), второй признак: width (столбец № 12), класс: num-of-doors (столбец № 6) |

Продолжение таблицы 2

|  |  |  |
| --- | --- | --- |
| 1 | 2 | 3 |
|  | Glass Identification Database  (11-glass.txt).  Ссылка: http://archive.ics.uci.edu/ml/datasets/Glass+Identification | Первый признак: RI (столбец № 2), второй признак: Al (столбец № 5), класс: Type of glass (столбец № 11) |
|  | Echocardiogram Data  (12-echocardiogram.txt).  Ссылка: http://archive.ics.uci.edu/ml/datasets/Echocardiogram | Первый признак: fractional-shortening (столбец № 5), второй признак: wall-motion-score (столбец № 8), класс: still-alive (столбец № 2) |
|  | Horse Colic Database  (13-horse-colic.txt ).  Ссылка: http://archive.ics.uci.edu/ml/datasets/Horse+Colic | Первый признак: rectal temperature (столбец № 4), второй признак: total protein (столбец № 20), класс: outcome (столбец № 23). |
|  | Pima Indians Diabetes Database  (14-pima-indians-diabetes.txt). Ссылка: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes | Первый признак: Body mass index (столбец № 6), второй признак: Diabetes pedigree function (столбец № 7), класс: Class variable (столбец № 9) |
|  | Auto-Mpg Data  (15-auto-mpg.txt ). Ссылка: http://archive.ics.uci.edu/ml/datasets/Auto+MPG | Первый признак: mpg (столбец № 1), второй признак: horsepower (столбец № 4), класс: cylinders (столбец № 2). |

***Контрольные вопросы***

1 В чём состоит задача кластеризации данных?

2 Какие существуют различные способы определения расстояния между объектами наблюдения по их признакам?

3 К какому классу сложности относится задача кластеризации в классической постановке?

4 Как работает классическая реализация алгоритма k внутригрупповых средних?

5 Что можно сказать о сходимости алгоритма k внутригрупповых средних?

6 Какую функцию минимизирует алгоритм k внутригрупповых средних?

7 Какие существуют альтернативные варианты реализации алгоритма k внутригрупповых средних?

8 Какие существуют методы автоматического выбора начальных центров кластеров для алгоритма k внутригрупповых средних?

10 Что такое иерархическая кластеризация?

## Лаб.6 Вероятностное обучение с помощью наивного баесовского классификатора

***Цель***: научиться создавать модель наивного байесовского классификатора

Для освоения навыков выполнения классификации с помощью алгоритма наивного Байеса в среде R воспользоваться материалом [6-8]

В основе байесовской классификации (**алгоритм Наивный Байес)** лежит гипотеза максимальной вероятности, т.е. объект *d*i считается принадлежащим классу *c*j, если при достигается наибольшая апостериорная вероятность .

По формуле Байеса, *P*(*c*j/*d*)=*P*(*c*j)*P*(*d*/*c*j)/P(d)≈P(cj)P(d/cj),

где

*P*(*d*|*c*j)) – вероятность встретить объект d среди объектов класса *c*j,

*P*(*c*j) и *P*(*d*) – априорные вероятности класса *c*j и объекта *d* (последняя, не влияет на выбор класса и может быть опущена).

В среде R расчеты выполняются обычно с использованием функций NaiveBayes() из пакета klaR или naiveBayes() из пакета e1071.

Визуализация трёхмерных данных в языке R может быть проведена с помощью функции plot3d из пакета rgl. Она принимает координаты точек в каждом из трёх пространственных измерений и изображает эти точки на графике. В полученном графическом окне график можно вращать и масштабировать. Сигнатура этой функции приведена ниже.

plot3d(x, y, z, xlab, ylab, zlab, type = "p", col, size, lwd, radius, add = FALSE, aspect = !add, ...),

где

x, y, z – векторы координат точек по каждому измерению;

xlab, ylab, zlab – названия координатных осей;

type – способ отображения точек на графике;

Поддерживаемые значения:

«p» для точек,

«s» для сфер,

«l» для линий,

«h» для отрезков, перпендикулярных плоскости z = 0,

и «n» для отсутствия графического представления;

col – вектор цветов для каждой точки;

size – размер изображаемых точек;

lwd – толщина линий для соответствующего способа представления точек;

radius – радиус сфер для соответствующего способа представления точек;

add – логическое значение, означающее, добавлять ли точки на уже имеющийся график, или же создать новый;

aspect – логическое значение, означающее, подбирать ли масштаб автоматически.

***Задание***

Входные данные: n объектов, каждый из которых характеризуется тремя числовыми признаками {*xi*} и {*yi*}, {*zi*},а также номером класса {*ci*}. Требуется исследовать работу алгоритма классификации объектов наблюдения по трем признакам. Для каждого набора данных требуется выполнить следующие задания.

1. Случайным образом разделить имеющуюся выборку в соотношении 85/15 на обучающую и контрольную выборку.

2. Произвести классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритмов классификации по наивному Байесу и методу k-ближайших соседей.

3. Изобразить объекты графически в трёхмерном пространстве. Для объектов разных классов и разных выборок следует использовать разные обозначения. Отдельно представить графики, на одном из которых объекты из контрольной выборки имеют свои настоящие классы, а на другом – классы, к которым их отнёс классификатор.

4. Оценить вероятность ошибочной классификации

Все описанные задания требуется выполнить для двух наборов данных.

В таблице 2 (Приложения A) представлены

1.Смоделированные независимые случайные векторы (*X*, *Y*, *Z*), *n*1 из которых относятся к первому классу, а *n*2 – ко второму классу. Векторы, относящиеся к первому классу, распределены по гауссовскому закону с математическим ожиданием *a*1 и корреляционной матрицей *R*1, а векторы, относящиеся ко второму классу, – по гауссовскому закону с математическим ожиданием *a*2 и корреляционной матрицей *R*2

2. Реальные статистические данные из заданного набора (выдаются преподавателем). Отчёт кроме прочих обязательных элементов должен включать:

1. изображения данных в виде точек на плоскости, причём данные из разных классов и из разных выборок должны изображаться отличающимися друг от друга;
2. изображения результатов классификации данных в виде точек на плоскости, причём данные из разных классов и из разных выборок должны изображаться отличающимися друг от друга;
3. значения оценок вероятности ошибочной классификации для каждого из случаев.

***Варианты заданий на лабораторную работу***

1. Смоделированные независимые случайные векторы (*X*, *Y*, *Z*), *n*1 из которых относятся к первому классу, а *n*2 – ко второму классу. Значения количеств объектов наблюдения для каждого класса (*n*1 и *n*2), значения векторов математических ожиданий для каждого класса (*a*1 и *a*2) и корреляционные матрицы для каждого класса (*R*1 и *R*2) для моделируемой выборки из гауссовских случайных векторов представлены в таблице 2 Приложения A.

2. Реальные статистические данные из заданного набора (выдаются преподавателем согласно варианту лабораторной работы №5, третий признак определяется из таблицы 3.

Таблица 3– Третий признак для решения задачи классификации.

|  |  |
| --- | --- |
| Вариант | Третий признак |
|  | size (столбец № 9) |
|  | Length (столбец № 5) |
|  | E\_bmr (столбец № 7) |
|  | WT18 (столбец № 9) |
|  | Tst (столбец № 6) |
|  | optical-density (столбец № 13) |
|  | temp (столбец № 9) |
|  | fractal dimension (столбец № 13) |
|  | Height (столбец № 4) |
|  | height (столбец № 13) |
|  | Fe (столбец № 10) |
|  | lvdd (столбец № 7) |
|  | packed cell volume (столбец № 19) |
|  | Age (столбец № 8) |
|  | acceleration (столбец № 6) |

***Контрольные вопросы.***

В чём состоит задача классификации?

2. Как вероятность ошибочной классификации оценивается по контрольной выборке? Каким свойствам отвечает её оценка?

3. Какую функцию минимизирует байесовский классификатор?

4. Как записывается формула Байеса? Что такое априорная и апостериорная вероятности и где они фигурируют в этой формуле?

5. В чём достоинства и недостатки байесовского классификатора? Почему он редко используется на практике?

6. Как работает байесовский классификатор для случая двух классов и одинаковых априорных вероятностей появления объектов?

7. Описать процесс генерации модельных данных из многомерных нормальных векторов с заданными векторами математического ожидания и ковариационной матрицей.

8. Подтвердить соответствие полученных данных условиям (проверить равенство выборочных средних и дисперсий математическим ожиданиям и элементам ковариационным матрицам ) с помощью проверки гипотез и визуализации.

## Лр.7 Классификация с помощью логистической регрессии

***Цель:*** научиться строить модель логистической регрессии и интерпретировать полученные результаты с точки зрения решения задачи классификации объектов.

***Теоретические сведения***

Для освоения навыков выполнения классификации в среде R воспользоваться материалом [5, с.51-71].

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится, так называемая, зависимая переменная *y*, принимающая лишь одно из двух значений – как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) – [вещественных](https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D1%89%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D1%87%D0%B8%D1%81%D0%BB%D0%BE) *x*1, *x*2, ..., *x*n, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

 Логистическая функция имеет вид

,

где Z – линейная комбинация независимых переменных.

Подбор коэффициентов регрессионного уравнения осуществляется на обучающей выборке c помощью метода максимального правдоподобия.

Ключевые термины для логистической регрессии:

Логит (logit) – функция, которая отображает вероятность принадлежности классу в интервал ±∞ (вместо интервала от 0 до 1).

Отношение шансов (odds) – Отношение "успеха" (1) к "неуспеху" (0).

Логарифм отношения шансов (log odds) – результат в преобразованной модели (в виде линейной комбинации), который отображается назад в вероятность.

Реализация на языке R представлена в следующем фрагменте:

# чтение данных в формате CSV

rdata <- read.csv("input.csv", sep = ',', header = FALSE)

# оценивание модели

model = glm(formula = target ~ x + y + z, data = rdata, family = binomial)

# вывод результатов оценки модели

print(summary(model))

***Задание***

Повести классификацию объектов методом логистической регрессии ( использовать наборы данных, лабораторной работе № 6).

***Контрольные вопросы.***

1. Что такое логистическая регрессия?

2. Какие задачи решает логистическая регрессия?

3 В чем отличие логистической регрессии от линейной регрессии?

4. Как интерпретировать коэффициенты логистической регрессии?

5. Какие методы оценки качества модели вы знаете?

7. Как выбрать оптимальное количество признаков для модели?

8. Какие ограничения есть у логистической регрессии?

9. Какие существуют методы улучшения качества модели?

## Лр.8.ROC-анализ для сравнения методов классификации

***Цель:*** научиться выбирать лучший алгоритм для осуществления классификации объектов различными методами

***Теоретические сведения***

Когда речь заходит о сравнении методов классификации в R, важно учитывать несколько ключевых аспектов. Сравнение методов классификации в R осуществляется на основе нескольких факторов, включая производительность модели, удобство использования, способность к обобщению, требования к предварительной обработке данных и т. д. К широко используемым инструментам сравнения качества предсказания класса в R различными методами относятся: кросс-валидация для оценки производительности модели, а также сравнение их метрик качества, точность, полнота, F-мера или ROC-кривая. Необходимо также учитывать интерпретируемость моделей, а также их способность обобщения для новых данных.

Большой выбор методов классификации в R дает возможность подобрать подходящий метод в зависимости от конкретной задачи и особенностей данных.

Задача алгоритма классификации состоит в том, чтобы относить ранее неизвестные объекты к тому или иному классу.

Результатом классификации могут быть четыре варианта:

* истинно-положительный результат (true-positive, TP) – TP - предсказано True, в действительности True
* ложноположительный результат (false-positive, FP) – FP - предсказано True, в действительности False
* истинно-отрицательный результат (true-negative, TN) – предсказано False, в действительности False;
* ложноотрицательный результат (false-negative, FN) – предсказано False, в действительности True.

Для численного представления качества оценки в R используют матрицу ошибок(confusion matrix). Матрица ошибок в машинном обучении(таблица 4) – это таблица, которая показывает количество правильных и неправильных предсказаний модели, в которой по строкам указываются реальные значения (истинные классы), а по столбцам — предсказанные значения (прогнозы модели)..

Таблица 4 – Матрица ошибок

|  |  |  |
| --- | --- | --- |
|  | **Predicted = 0** | **Predicted = 1** |
| Actual = 0 | True Negatives (TN) | False Positives (FP) |
| Actual = 1 | False Negatives (FN) | True Positives (TP) |

Три из основных метрик, которые широко используются для оценки качества работы классификационных моделей, включают точность (Precision), полноту (Recall) и F-меру (F1-score).

1. Точность измеряет, сколько из объектов, которые модель отнесла к положительному классу, действительно принадлежат к положительному классу. Она определяется как отношение числа верно классифицированных положительных объектов к общему количеству объектов, которые модель отнесла к положительному классу.

*Precision* = *TP*/(*TP* + *FP* )

2. Полнота измеряет, сколько из общего числа реальных положительных объектов модель успешно обнаружила. Она определяется как отношение числа верно классифицированных положительных объектов к общему количеству реальных положительных объектов.

*Recall* = *TP* /(*TP* + *FN* )

3. *F*-мера представляет собой сбалансированную метрику, которая учитывает и точность, и полноту. Она является гармоническим средним между точностью и полнотой.

*F1-score* = 2 ×(*Precision* × *Recall*)/(*Precision* + *Recall*)

Точность и полнота обычно работают в противоположных направлениях: увеличение точности может привести к снижению полноты и наоборот. *F*-мера учитывает обе эти метрики и представляет собой компромисс между точностью и полнотой. F-мера особенно полезна в случаях, когда классы несбалансированы, т.е. когда количество объектов в одном классе существенно превышает количество объектов в другом.

Эти метрики помогают в оценке качества работы классификационных моделей и помогают понять, насколько хорошо модель способна правильно классифицировать положительные и отрицательные объекты.

ROC-анализ (Receiver Operating Characteristic) — это метод оценки качества модели классификации, основанный на анализе кривой зависимости истинно положительных результатов от ложноположительных. ROC-анализ используется для сравнения методов классификации и определения оптимального порога для принятия решений в задаче логистической регрессии.

Для оценки качества логистической модели строится матрица ошибок, рассчитываются коэффициенты чувствительности и специфичности, которые варьируются в зависимости от порога отсечения.

Результатом логистической регрессии является вероятность того что событие произойдет, но чтобы сделать предсказание нам необходимо определить пороговое (threshold) значение t, такое что:

*P*(*y*=1)>=*t*, событие произошло,

*P*(*y*=1)<*t*, событие не произошло.

При высоком пороговом значении модель редко будет предсказывать положительный результат (только при высокой вероятности P(y=1)). И обратно, при низком пороговом значении модель будет чаще предсказывать положительный исход и реже отрицательный. Матрица ошибок также может использоваться для настройки модели и выбора оптимального порога принятия решений.

ROC-кривая строится на графике, где по оси абсцисс откладываются значения ложноположительных результатов, а по оси ординат — истинно положительные результаты. Каждой точке на кривой соответствует определенный порог принятия решения. Чем ближе кривая к верхнему левому углу, тем лучше модель классификации. Отсюда можно получить два значения, которые помогут нам определить какие ошибки делает модель:

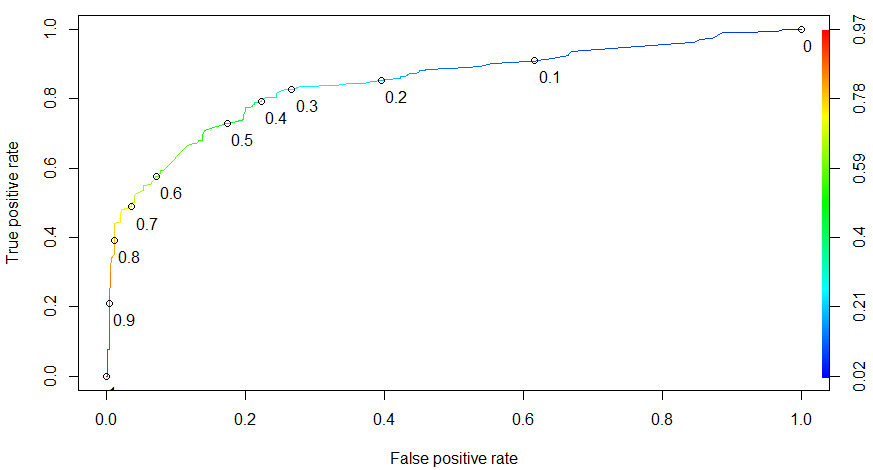
Чувствительность –доля верно предсказанных позитивных исходов (True positive rate)

*Sensitivity*=*TP*/(*TP*+*FN*)

Специфичность – доля верно предсказанных негативных исходов (True negative rate).

*Specificity*=*TN*/(*TN*+*FP*)

Модель с более высоким пороговым значением будет иметь более высокую чувствительность и низкую специфичность. Модель с низким пороговым значением наоборот. Если предпочтений нет можно оставить пороговое значение *t*=0,5. Подобрать необходимое пороговое значение можно с помощью ROC-кривой (рисунок 1).



True negative rate указывается на оси абсцисс, True positive rate на оси ординат. Кривая показывает соотношения этих величин при разных значения пороговой величины.

Построить ROC-кривую в R можно следующим набором команд (используется встроенный набор данных iris).

# Загрузка необходимых библиотек

library(pROC)

library(ggplot2)

# Создание примера модели логистической регрессии и ROC-кривой

# Имитация обучающего и тестового набора данных (используем встроенный набор данных iris)

data(iris)

set.seed(123)

train\_indices <- sample(1:nrow(iris), nrow(iris)\*0.7) # 70% данных для обучения

train\_data <- iris[train\_indices, ]

test\_data <- iris[-train\_indices, ]

# Создание модели логистической регрессии

logit\_model <- glm(Species ~ ., data = train\_data, family = "binomial")

# Построение ROC-кривой

roc\_curve <- roc(test\_data$Species, predict(logit\_model, newdata = test\_data, type = "response")[, "versicolor"])

# Визуализация ROC-кривой

ggplot(as.data.frame(roc\_curve), aes(x = 1 - specificity, y = sensitivity)) +

geom\_line() +

geom\_abline(intercept = 0, slope = 1, linetype = "dashed") +

labs(title = "ROC Curve",

x = "False Positive Rate",

y = "True Positive Rate") +

theme\_minimal()

***Задание***

Провести сравнение результатов классификации методами, использованными в лабораторных 6, 7.

***Контрольные вопросы.***

1 Какие метрики используются для оценки качества предсказания?

2 Что такое матрица ошибок?

3 Что представляет собой ROC-кривая и для чего она используется в контексте оценки качества классификационных моделей?

2. Как интерпретировать ROC-кривую и какие параметры из неё могут быть полезны для оценки качества модели?

4 Какие значения на ROC-кривой указывают на хорошую предсказательную способность модели, а какие наоборот?

5 Как соотносятся ROC-кривая и метрики, такие как AUC-ROC (площадь под ROC-кривой) с точностью, полнотой и F-мерой?

6 Каковы преимущества использования ROC-анализа для сравнения методов классификации по сравнению с использованием только точности или полноты?

## Лаб.9 Классификация с использованием деревьев решений и правил

***Цель***: научиться применять алгоритмы построения деревьев решений, решающих правил, а также проводить сравнительный анализ классификационных моделей на их основе.

***Теоретические сведения***

Алгоритм CART (Classification and Regression Trees) представляет собой метод построения двоичных деревьев решений для задач классификации и регрессии.

1. Основная идея:

Алгоритм CART используется для построения деревьев решений в задачах классификации и регрессии.

Цель алгоритма: разбить данные на подмножества таким образом, чтобы в каждом подмножестве объекты были как можно более однородными с точки зрения целевой переменной.

2. Процесс построения дерева:

На каждом шаге алгоритм выбирает переменную и значение, чтобы разделить данные на два подмножества. Выбор происходит на основе критерия неоднородности, такого как критерий Джини для задач классификации или критерий наименьших квадратов для задач регрессии. Этот процесс повторяется рекурсивно для каждого подмножества, пока не выполнится условие останова.

3. Условие останова:

Дерево строится до тех пор, пока не будет достигнуто определенное количество объектов в листовых узлах, или до тех пор, пока нет возможности провести дальнейшее разделение.

4. Преимущества:

Простота интерпретации полученной модели в виде дерева решений.

Устойчивость к выбросам и некорректным данным.

5. Ограничения:

Склонность к переобучению при недостаточном прунинге дерева.

Не всегда способен улавливать сложные зависимости в данных.

Алгоритм CART представляет собой мощный инструмент для построения простых и понятных моделей, особенно в случаях, когда интерпретация решений играет важную роль.

Пример использования алгоритм CART для классификации в R:

# Установка пакета rpart, если он не установлен

# install.packages("rpart")

# Загрузка необходимых библиотек

library(rpart)

library(rpart.plot) # Дополнительная библиотека для визуализации деревьев

# Создание и обучение модели с использованием алгоритма CART

# Предположим, что у вас есть данные в переменной "data", где последний столбец - это целевая переменная, а все предыдущие - признаки

model <- rpart(target ~ feature1 + feature2 + ..., data = data, method = "class")

# Визуализация построенного дерева решений

rpart.plot(model) # Это отобразит дерево решений, используя библиотеку rpart.plot

***Задание***

1.Повести классификацию объектов CART ( использовать наборы данных, лабораторной работе № 6), выполнив настройку параметров модели и оценить ее производительность на обучающем и тестовом наборах данных.

2. Провести сравнительный анализ различных моделей по метрикам точности, полноты, F-меры, AUC-ROC и другим метрикам.

3. Визуализировать полученных деревьев решений для наглядного представления классификационных правил.

4.Сформулировать выводы о производительности и применимости классификационных моделей на основе деревьев решений и решающих правил.

***Список литературы***

1 **Борисов, В. В.,** Экспертные системы: учебное пособие/ В. В. Борисов, А. В. Бобряков, А. Е. Мисник – Смоленск: Универсум, 2021. – 110 с

2 **Кабаков, Р**. R в действии. Анализ и визуализация данных на языке R : практическое руководство / Р. Кабаков ; пер. с англ. П. А. Волковой. - 2-е изд. - Москва : ДМК Пресс, 2023. - 590 с. - ISBN 978-5-89818-347-9. - Текст : электронный. - URL: https://znanium.com/catalog/product/2102634 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

3 **Лонг, Д.** R. Книга рецептов: проверенные рецепты для статистики, анализа и визуализации данных : практическое руководство / Д. Лонг, П. Титор ; пер. с англ. Д. А. Беликова. - Москва : ДМК Пресс, 2020. - 510 с. - ISBN 978-5-97060-835-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/1210661 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

4 **Мастицкий, С. Э**. Статистический анализ и визуализация данных с помощью R : практическое руководство / С. Э. Мастицкий, В. К. Шитиков. - 2-е изд. - Москва : ДМК Пресс, 2023. - 497 с. - ISBN 978-5-89818-601-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/2108480 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

5 **Гайдель, А.В.** Лабораторный практикум по курсу «Интеллектуальный анализ данных»: практикум / А.В. Гайдель, А.Г. Храмов. – Самара: Изд-во Самарского университета, 2019. – 104 с

6 [A Step By Step Guide To Implement Naive Bayes In R | Edureka](https://www.edureka.co/blog/naive-bayes-in-r/#Practical%20Implementation%20of%20Naive%20Bayes%20In%20R)

7 <https://rpubs.com/aksyuk/MM_Lab-02>

8 Классификация, регрессия и другие алгоритмы Data Mining с использованием R Шитиков В. К., Мастицкий С. Э. <https://ranalytics.github.io/data-mining/index.html>

2 **Кабаков, Р**. R в действии. Анализ и визуализация данных на языке R : практическое руководство / Р. Кабаков ; пер. с англ. П. А. Волковой. - 2-е изд. - Москва : ДМК Пресс, 2023. - 590 с. - ISBN 978-5-89818-347-9. - Текст : электронный. - URL: https://znanium.com/catalog/product/2102634 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

3 **Лонг, Д.** R. Книга рецептов: проверенные рецепты для статистики, анализа и визуализации данных : практическое руководство / Д. Лонг, П. Титор ; пер. с англ. Д. А. Беликова. - Москва : ДМК Пресс, 2020. - 510 с. - ISBN 978-5-97060-835-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/1210661 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

4 **Мастицкий, С. Э**. Статистический анализ и визуализация данных с помощью R : практическое руководство / С. Э. Мастицкий, В. К. Шитиков. - 2-е изд. - Москва : ДМК Пресс, 2023. - 497 с. - ISBN 978-5-89818-601-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/2108480 (дата обращения: 08.01.2024). – Режим доступа: по подписке.

5 **Гайдель, А.В.** Лабораторный практикум по курсу «Интеллектуальный анализ данных»: практикум / А.В. Гайдель, А.Г. Храмов. – Самара: Изд-во Самарского университета, 2019. – 104 с

6 [A Step By Step Guide To Implement Naive Bayes In R | Edureka](https://www.edureka.co/blog/naive-bayes-in-r/#Practical%20Implementation%20of%20Naive%20Bayes%20In%20R)

7 <https://rpubs.com/aksyuk/MM_Lab-02>

8 Классификация, регрессия и другие алгоритмы Data Mining с использованием R Шитиков В. К., Мастицкий С. Э. https://ranalytics.github.io/data-mining/index.html

4. https://github.com/enikolaev/MMO – Репозиторий с примерами кода из лабораторных работ. 5. https://archive.ics.uci.edu/ml/index.html – Репозиторий наборов данных для машинного обучения (Центр машинного обучения и интеллектуальных систем). 6. https://www.kaggle.com – Портал и система проведения соревнований по проблемам анализа данных. 7. https://www.mockaroo.com – Сайт для генерации наборов данных.

1. <https://bdemeshev.github.io/installation/r/R_installation.html> [↑](#footnote-ref-1)