



# ТЕОРЕТИЧЕСКИЙ МИНИМУМ ПО BIG DATA

ВСЁ, ЧТО НУЖНО ЗНАТЬ О БОЛЬШИХ ДАННЫХ



# Краткое содержание

|  |     |
|--|-----|
| Предисловие.....   | 12  |
| Введение .....   | 16  |
| Почему Data Science?.....  | 18  |
| <b>Глава 1.</b> Об основах без лишних слов .....                         | 21  |
| <b>Глава 2.</b> Кластеризация методом k-средних.....                     | 39  |
| <b>Глава 3.</b> Метод главных компонент .....                            | 51  |
| <b>Глава 4.</b> Ассоциативные правила .....                              | 65  |
| <b>Глава 5.</b> Анализ социальных сетей .....                            | 77  |
| <b>Глава 6.</b> Регрессионный анализ.....                                | 93  |
| <b>Глава 7.</b> Метод k-ближайших соседей и обнаружение<br>аномалий..... | 107 |
| <b>Глава 8.</b> Метод опорных векторов .....                             | 117 |
| <b>Глава 9.</b> Дерево решений.....                                      | 127 |
| <b>Глава 10.</b> Случайные леса.....                                     | 137 |
| <b>Глава 11.</b> Нейронные сети .....                                    | 149 |
| <b>Глава 12.</b> A/B-тестирование и многорукие бандиты.....              | 167 |
| Приложения .....   | 179 |
| Глоссарий .....  | 188 |
| Литература и ссылки на источники.....                                    | 199 |
| Об авторах .....   | 204 |

# Оглавление

|   |           |
|---|-----------|
| <b>Предисловие .....</b>                        | <b>12</b> |
| От издательства.....                            | 15        |
| <b>Введение .....</b>                           | <b>16</b> |
| <b>Почему Data Science? .....</b>               | <b>18</b> |
| <b>Глава 1. Об основах без лишних слов.....</b> | <b>21</b> |
| 1.1. Подготовка данных.....                     | 22        |
| Формат данных.....                              | 23        |
| Типы переменных .....                           | 24        |
| Выбор переменных .....                          | 25        |
| Конструирование признаков .....                 | 25        |
| Неполные данные .....                           | 26        |
| 1.2. Выбор алгоритма.....                       | 27        |
| Обучение без учителя.....                       | 28        |
| Обучение с учителем .....                       | 29        |
| Обучение с подкреплением .....                  | 30        |
| Другие факторы.....                             | 31        |

---

|   |           |
|---|-----------|
| 1.3. Настройка параметров .....                       | 31        |
| 1.4. Оценка результатов .....                         | 33        |
| Метрики классификации .....                           | 34        |
| Метрика регрессии .....                               | 35        |
| Валидация.....  | 36        |
| 1.5. Краткие итоги .....                              | 38        |
| <b>Глава 2. Кластеризация методом k-средних .....</b> | <b>39</b> |
| 2.1. Поиск кластеров клиентов.....                    | 40        |
| 2.2. Пример: профили кинозрителей .....               | 41        |
| 2.3. Определение кластеров.....                       | 42        |
| Сколько кластеров существует? .....                   | 44        |
| Что включают кластеры? .....                          | 46        |
| 2.4. Ограничения .....                                | 48        |
| 2.5. Краткие итоги.....                               | 49        |
| <b>Глава 3. Метод главных компонент .....</b>         | <b>51</b> |
| 3.1. Изучение пищевой ценности .....                  | 52        |
| 3.2. Главные компоненты.....                          | 53        |
| 3.3. Пример: анализ пищевых групп.....                | 56        |
| 3.4. Ограничения .....                                | 61        |
| 3.5. Краткие итоги.....                               | 64        |
| <b>Глава 4. Ассоциативные правила .....</b>           | <b>65</b> |
| 4.1. Поиск покупательских шаблонов.....               | 66        |
| 4.2. Поддержка, достоверность и лифт .....            | 67        |

|  |            |
|--|------------|
| 4.3. Пример: ведение продуктовых продаж .....                                  | 69         |
| 4.4. Принцип Arğiögi .....   | 72         |
| Поиск товарных наборов с высокой поддержкой.....                               | 73         |
| Поиск товарных правил с высокой<br>достоверностью или лифтом.....              | 74         |
| 4.5. Ограничения .....   | 75         |
| 4.6. Краткие итоги.....  | 76         |
| <b>Глава 5. Анализ социальных сетей .....</b>                                  | <b>77</b>  |
| 5.1. Составление схемы отношений.....  | 78         |
| 5.2. Пример: геополитика в торговле оружием.....                               | 80         |
| 5.3. Лувенский метод .....   | 84         |
| 5.4. Алгоритм PageRank .....   | 86         |
| 5.5. Ограничения .....   | 90         |
| 5.6. Краткие итоги .....   | 91         |
| <b>Глава 6. Регрессионный анализ .....</b>                                     | <b>93</b>  |
| 6.1. Выведение линии тренда.....   | 94         |
| 6.2. Пример: предсказание цен на дома .....                                    | 95         |
| 6.3. Градиентный спуск .....   | 98         |
| 6.4. Коэффициенты регрессии.....   | 101        |
| 6.5. Коэффициенты корреляции.....  | 102        |
| 6.6. Ограничения .....   | 104        |
| 6.7. Краткие итоги.....  | 106        |
| <b>Глава 7. Метод k-ближайших соседей и обнаружение<br/>    аномалий .....</b> | <b>107</b> |
| 7.1. Пищевая экспертиза.....   | 108        |

---

|   |            |
|---|------------|
| 7.2. Яблоко от яблони недалеко падает .....                       | 109        |
| 7.3. Пример: истинные различия в вине .....                       | 111        |
| 7.4. Обнаружение аномалий.....                                    | 113        |
| 7.5. Ограничения .....  | 114        |
| 7.6. Краткие итоги.....   | 115        |
| <b>Глава 8. Метод опорных векторов.....</b>                       | <b>117</b> |
| 8.1 «Нет» или «о, нет!»?.....                                     | 118        |
| 8.2. Пример: обнаружение сердечно-сосудистых<br>заболеваний ..... | 118        |
| 8.3. Построение оптимальной границы.....                          | 120        |
| 8.4. Ограничения .....  | 124        |
| 8.5. Краткие итоги.....   | 125        |
| <b>Глава 9. Дерево решений.....</b>                               | <b>127</b> |
| 9.1. Прогноз выживания в катастрофе .....                         | 128        |
| 9.2. Пример: спасение с тонущего «Титаника» .....                 | 128        |
| 9.3. Создание дерева решений .....                                | 131        |
| 9.4. Ограничения .....  | 133        |
| 9.5. Краткие итоги.....   | 135        |
| <b>Глава 10. Случайные леса.....</b>                              | <b>137</b> |
| 10.1. Мудрость толпы .....  | 138        |
| 10.2. Пример: предсказание криминальной<br>активности.....        | 139        |
| 10.3. Ансамбли .....  | 144        |
| 10.4. Бэггинг.....  | 145        |

10.5. Ограничения..... 147  
10.6. Краткие итоги ..... 148

**Глава 11. Нейронные сети ..... 149**

11.1. Создание мозга ..... 150  
11.2. Пример: распознавание рукописных цифр..... 152  
11.3. Компоненты нейронной сети..... 156  
11.4. Правила активации ..... 159  
11.5. Ограничения..... 161  
11.6. Краткие итоги ..... 165

**Глава 12. А/В-тестирование и многорукие бандиты ..... 167**

12.1. Основы А/В-тестирования..... 168  
12.2. Ограничения А/В-тестирования..... 169  
12.3. Стратегия снижения эpsilon..... 169  
12.4. Пример: многорукие бандиты ..... 171  
12.5. Забавный факт: ставка на победителя ..... 174  
12.6. Ограничения стратегии снижения эpsilon ..... 175  
12.7. Краткие итоги ..... 176

**Приложения ..... 179**

Приложение А. Обзор алгоритмов обучения  
    без учителя ..... 180  
Приложение В. Обзор алгоритмов обучения  
    с учителем ..... 181  
Приложение С. Список параметров настройки ..... 182

---

|   |            |
|---|------------|
| Приложение D. Другие метрики оценки.....      | 183        |
| Метрики классификации .....                   | 183        |
| Метрики регрессии.....                        | 186        |
| <b>Глоссарий .....</b>                        | <b>188</b> |
| <b>Литература и ссылки на источники .....</b> | <b>199</b> |
| Источники на английском языке .....           | 199        |
| Литература на русском языке.....              | 202        |
| <b>Об авторах.....</b>                        | <b>204</b> |