

М 92

**Мухамедиев Р.И., Амиргалиев Е.Н.**

Введение в машинное обучение: Учебник. – Алматы, 2022. – 288 с.

Рекомендован к изданию с присвоением грифа УМО РУМС 21 декабря 2021 года, протокол № 2.

Рецензенты:

Баракшин В.Б., д.т.н., профессор, НГУ, Россия, Новосибирск

Никульчев Е.В., д.т.н., профессор, МИРЭА, Россия, Москва

Маткаримов Б.Т., д.т.н., профессор, Назарбаев университет, Казахстан, Нурсултан

**Мухамедиев Равиль Ильгизович  
Амиргалиев Едилхан Несипханович**

## **ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ**

**УЧЕБНИК**

Рекомендован к изданию УМО РУМС

Машинное обучение – часть искусственного интеллекта, в рамках которого в последние годы получены впечатляющие результаты в обработке разных видов сигналов и данных. В настоящем пособии даются весьма краткие теоретические и относительно подробные практические сведения о применении отдельных алгоритмов классификации и регрессии. Для практического освоения материала достаточно базовых навыков работы с языком Python. При этом, освоение возможностей основных библиотек, таких как matplotlib, numpy, pandas, sklearn происходит в процессе решения задач. Используя полученные знания и навыки, студенты смогут решать широкий круг задач классификации, регрессии, анализировать влияние отдельных признаков на работу классификаторов и регрессионных моделей, снижать размерность данных, визуализировать результаты и оценивать качество моделей машинного обучения.

Учебник предназначен для студентов, магистрантов и докторантов по направлению подготовки «061 Информационно-коммуникационные технологии», желающих приобрести практические навыки по применению методов классификации, регрессии, снижения размерности данных и освоить базовые математические концепции, относящиеся к машинному обучению.

ISBN 978-601-08-1177-5

Работа состоялась при поддержке грантов Комитета науки Министерства образования и науки Республики Казахстан BR10965172, AP09259587, AP08856412, BR05236839, BR05236447, 2318/ГФ3, и гранта ERASMUS+ Advanced Centre for PhD Students and Young Researchers in Informatics, ACeSYRI, reg.no. 610166-EPP-1-2019-1-SK-EPPKA2-CBHE-JP

## Содержание

ПРЕДИСЛОВИЕ	9
ВВЕДЕНИЕ	11
ЧАСТЬ I. МАТЕМАТИЧЕСКИЕ МОДЕЛИ И ПРИКЛАДНЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ	14
1 ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ	14
1.1 Составные части искусственного интеллекта	14
1.2 Машинное обучение в задачах обработки данных	16
1.3 Программное обеспечение для решения задач машинного обучения	19
1.4 Схема настройки системы машинного обучения	21
1.5 Контрольные вопросы	23
2 КЛАССИЧЕСКИЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ	24
2.1 Формальное описание задач машинного обучения	24
2.2 Линейная регрессия одной переменной	26
2.3 Полиномиальная регрессия	31
2.4 Классификаторы. Логистическая регрессия	33
2.5 Контрольные вопросы	37
2.6 Искусственные нейронные сети	37
2.6.1 Вводные замечания	37
2.6.2 Математическое описание искусственной нейронной сети	39
2.6.3 Алгоритм обратного распространения ошибки	42
2.6.4 Пошаговый пример расчета алгоритма обратного распространения ошибки	44
2.6.5 Активационные функции	47
2.7 Контрольные вопросы	49
2.8 Пример простого классификатора	49
2.9 Алгоритм k ближайших соседей (k-Nearest Neighbor – k-NN)	52
2.10 Алгоритм опорных векторов	53
2.11 Статистические методы в машинном обучении. Наивный байесовский вывод	55
2.11.1 Теорема Байеса и ее применение в машинном обучении	55
2.11.2 Алгоритм Naïve Bayes	57
2.11.3 Положительные и отрицательные свойства Naïve Bayes	59
2.11.4 Приложения наивного байесовского алгоритма	60
2.12 Композиции алгоритмов машинного обучения. Бустинг	61
2.13 Снижение размерности данных. Метод главных компонент	63
2.14 Контрольные вопросы	66
3 ОЦЕНКА КАЧЕСТВА МЕТОДОВ ML	67
3.1 Метрики оценки качества классификации	68
3.2 Матрица путаницы (Confusion matrix) и фиктивный классификатор (Dummy Classifier)	71
3.3 Подбор параметров по сетке	74
3.4 Кривые Precision-Recall и ROC	76
3.5 Микро- и макросреднее (Micro- and macro-average)	81
3.6 Перекрестная оценка модели (Cross-validation)	83
3.7 Показатели оценки качества регрессии	85
3.8 «Обучаемость» алгоритмов	85

3.9 Контрольные вопросы	88
4 ПРЕДОБРАБОТКА ДАННЫХ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ	90
4.1 Оценка набора данных	90
4.2 Входные переменные и целевая колонка	93
4.3 Обработка отсутствующих и аномальных значений	94
4.4 Поиск и преобразование дубликатов	96
4.5 Кодирование значений	96
4.6 Взаимная корреляция исходных значений	98
4.7 Нормализация данных	100
4.8 Специальные методы предобработки данных	103
4.9 Контрольные вопросы	104
5 МАШИННОЕ ОБУЧЕНИЕ В ЗАДАЧАХ С БОЛЬШИМ ОБЪЕМОМ ДАННЫХ	105
6 ГЛУБОКОЕ ОБУЧЕНИЕ	109
6.1 Вводные замечания	109
6.2 Пример. Реализация нейронной сети прямого распространения в пакете TensorFlow-Keras	111
6.3 Рекуррентные нейронные сети	118
6.4 Сверточные нейронные сети	119
6.4.1 Задачи компьютерного зрения и сверточные сети	119
6.4.2 Сверточный фильтр	122
6.4.3 Эксперименты со сверточными фильтрами	126
6.4.4 Параметры сверточных фильтров	128
6.4.5 Pooling (объединение)	130
6.4.6 Архитектуры сверточных сетей	130
6.4.7 Применение сверточных сетей для распознавания лиц	132
6.4.8 Пример. Реализация сверточной сети для распознавания изображений с использованием TensorFlow и GPU	139
6.5 Заключение к разделу «Глубокое обучение»	146
6.6 Контрольные вопросы	146
7 ИНТЕРПРЕТАЦИЯ ЧЕРНЫХ ЯЩИКОВ МАШИННОГО ОБУЧЕНИЯ.	148
ЧАСТЬ II. ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ. ЛАБОРАТОРНЫЙ ПРАКТИКУМ.	154
8 МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ	154
9 ЛАБОРАТОРНЫЙ ПРАКТИКУМ	156
9.1 Реализация линейной и полиномиальной регрессии в Python-numpy	156
9.1.1 Постановка задачи	156
9.1.2 Пример решения	158
9.1.3 Задача 1. Расчет параметров полиномиальной регрессии второго порядка	161
9.1.4 Задача 2. Матричный способ расчета параметров полиномиальной регрессии	162
9.2 Реализация алгоритма логистической регрессии в Python-numpy	163
9.2.1 Постановка задачи	163
9.2.2 Пример решения	165

9.2.3	Задача 1. Классификация объектов описываемых тремя свойствами	167
9.2.4	Задача 2. Расчет параметров точности классификатора	167
9.3	Предварительная обработка текстов	168
9.3.1	Постановка задачи	168
9.3.2	Задача. Разработка функции для предобработки твитов	169
9.4	Применение алгоритма Naïve Bayes для предсказания	176
9.4.1	Постановка задачи	176
9.4.2	Задача 1. Использование байесовского выражения для предсказания вероятности игры	176
9.4.3	Задача 2. Предсказание с помощью алгоритма Naïve Bayes	177
9.5	Реализация линейной регрессии в Python с применением sklearn	180
9.5.1	Постановка задачи	180
9.5.2	Пример решения	180
9.5.3	Задача 1. Окрашивание обучающего и тестового множеств	183
9.5.4	Задача 2*. Генерация данных для экспериментов	183
9.5.5	Задача 3**. Расчет параметров линейной регрессии	184
9.6	Реализация полиномиальной регрессии в Python	186
9.6.1	Постановка задачи	186
9.6.2	Пример решения	187
9.6.3	Задача 1. Визуализация обучающего и тестового множеств	189
9.6.4	Задача 2*. Построение регрессионных кривых и расчет показателей точности	189
9.6.5	Задача 3**. Расчет минимальной степени полиномиальной регрессии	190
9.6.6	Задача 4*. Применение RandomForestRegressor	192
9.7	Реализация классификатора на базе логистической регрессии	194
9.7.1	Постановка задачи	194
9.7.2	Пример решения	195
9.7.3	Задача 1*. Оценка качества работы логистической регрессии	196
9.7.4	Задача 2*. Классификатор для набора данных iris	198
9.8	Реализация классификатора на базе алгоритма ближайших соседей	199
9.8.1	Постановка задачи	199
9.8.2	Пример решения	200
9.8.3	Задача 1*. Оценка качества работы классификатора kNN	201
9.8.4	Задача 2. Визуализация результатов работы классификатора kNN	202
9.8.5	Задача 3. Настройка классификатора kNN	203
9.8.6	Задача 4*. Использование kNN для классификации данных breast_cancer	203
9.9	Метод опорных векторов - Support vector machines (SVM)	205
9.9.1	Информация о методе	205
9.9.2	Постановка задачи	206
9.9.3	Пример решения	207
9.9.4	Задача 0. Оценка точности классификатора SVC.	210
9.9.5	Задача 1*. Настройка классификатора SVM.	210
9.10	Реализация многослойной нейронной сети прямого распространения	212

9.10.1	Постановка задачи	212
9.10.2	Пример решения	213
9.10.3	Задача 1*. Определение точности классификации.	214
9.10.4	Задача 2. Визуализация результатов классификации.	215
9.10.5	Задача 3. Настройка параметров классификатора MLP.	216
9.10.6	Задача 4*. Классификация набора данных breast_cancer с помощью MLP.	217
9.11	Метод главных компонент	218
9.11.1	Постановка задачи	218
9.11.2	Пример решения	219
9.11.3	Задача 1. Оценка снижения вариативности.	222
9.11.4	Задача 2**. Снижение размерности набора данных iris.	223
9.11.5	Задача 3**. Снижение размерности набора breast_cancer	223
9.12	«Серебряная пуля» машинного обучения - метод XGBoost	225
9.12.1	Постановка задачи	225
9.12.2	Данные и результат работы	227
9.12.3	Задача 0. Многопоточная работа XGBoost.	229
9.12.4	Задача 1. Сравнение классификаторов при применении нормализации данных.	229
9.12.5	Задача 2**. Классификация большого набора данных.	229
9.13	Предобработка табличных данных	230
9.13.1	Постановка задачи	230
9.13.2	Задачи	231
9.14	Представление слов в векторном пространстве (Word2Vec)	232
9.14.1	Постановка задачи	232
9.14.2	Задача 1. Отображение слов в двухмерном пространстве.	232
9.14.3	Задача 2. Отображение слов в трехмерном пространстве.	234
9.15	Классификация данных с применением нескольких классификаторов	235
9.15.1	Постановка задачи	235
9.15.2	Задача 1. Настройка классификатора SVC	238
9.15.3	Задача 2. Сравнение классификаторов	238
9.16	Метод LIME (Local Interpretable Model-agnostic)	239
9.16.1	Постановка задачи	239
9.16.2	Пример решения	241
9.16.3	Задача 1. Оценка влияния параметров в наборе данных breast_cancer.	243
9.16.4	Задача 2*. Оценка влияния параметров в наборе данных iris.	244
9.17	Применение сверточных сетей для распознавания лиц	245
9.17.1	Постановка задачи	245
9.17.2	Пример применения сверточных сетей для идентификации персоны	245
9.17.3	Задача 1. Найти знаменитость похожую на вас.	248
9.17.4	Задача 2. Ограничение доступа на основе распознавания лица.	248
9.17.5	Задача 2. Найти имя человека по фотографии	249
10	ПРОЕКТ ПО СОЗДАНИЮ КЛАССИФИКАТОРА ЛИТОЛОГИЧЕСКИХ ТИПОВ.	250

10.1	Постановка задачи	250
10.2	Задача I	252
10.3	Исходные данные	252
10.3.1	Общее описание исходных данных	252
10.3.2	Синтетические данные	253
10.4	Задача II	254
10.5	Загрузка данных и пример построения классификатора	254
10.5.1	Описание программы GTSTxtReader	255
10.6	Задача 1	262
10.7	Задача 2	262
11	ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ ВЫЧИСЛЕНИЯ НА ОСНОВЕ NUMBA	263
	ЗАКЛЮЧЕНИЕ	270
	БЛАГОДАРНОСТИ	270
	ПРИЛОЖЕНИЕ 1. ОПЕРАЦИИ ЛИНЕЙНОЙ АЛГЕБРЫ	271
	Матрицы	271
	Векторы и скаляры	271
	Нулевая и единичная матрицы	272
	Операции с матрицами	272
	Умножение матриц и векторов	272
	Транспонирование матриц	274
	ПРИЛОЖЕНИЕ 2. ТАКСОНОМИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ	275
	ПРИЛОЖЕНИЕ 3. ДОПОЛНИТЕЛЬНЫЕ МЕТОДЫ ПРЕДОБРАБОТКИ ДАННЫХ	278
	Алгоритм поиска аномалий на основе нормального распределения	278
	Применение преобразования Фурье и вейвлет-анализа	280
	ПРИЛОЖЕНИЕ 4. ФИЗИЧЕСКИЕ ПРИНЦИПЫ ПОЛУЧЕНИЯ КАРОТАЖНЫХ ДАННЫХ	285