МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

МАШИННОЕ ОБУЧЕНИЕ КАК СОСТАВЛЯЮЩАЯ ЧАСТЬ ПОДГОТОВКИ ИТ-СПЕЦИАЛИСТОВ: РАЗРАБОТКА ЭЛЕКТРОННОГО КУРСА

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 01.04.02 Прикладная математика и информатика
факультета компьютерных наук и информационных технологий
Ищук Марии Алексеевны

Научный руководитель:
к.ф.-м.н. М. В. Огнева
подпись, дата

Зав. кафедрой:
к.ф.-м.н. М. В. Огнева
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

Актуальность темы. Машинное обучение является одним из направлений искусственного интеллекта. Основной принцип заключается в том, что компьютер получает данные и «обучается» на них. Системы машинного обучения позволяют быстро применять знания, полученные при обучении на больших наборах данных, что позволяет им преуспевать в таких задачах, как распознавание лиц, распознавание речи, распознавание объектов, перевод, и многих других.

На данный момент изучение машинного обучения является очень актуальным, в связи с развитием информационных технологий, в частности, социальных сетей, интернет-магазинов и других онлайн сервисов, растет потребность в анализе и обработке больших объемов информации и извлечение из нее нужных данных. Но на эту тему существует не так много открытых и доступных курсов. В большинстве своем, они представлены на английском языке [1, 2]. Один из известных курсов на русском языке — это курс К. В. Воронцова [3, 4].

При изучении машинного обучения необходимо использование какого-либо языка программирования. Наиболее приспособленным для этого является язык программирования Python, поэтому именно он был выбран для изучения в данном курсе. Если данный язык не изучался ранее, то имеет смысл начать обучение именно с него. При создании курса необходимо учитывать теоретическую и практическую подготовленность обучающихся. Создание курса, который подойдет для любого ученика заключается в предоставлении информации с самых основ. Таким образом, начиная с обучения языку программирования Python, к концу такого курса, слушатель будет уметь решать практические задачи машинного обучения.

Цель магистерской работы – создание электронного курса по машинному обучению в системе MOODLE.

Поставленная цель определила **следующие задачи**:

1. Анализ актуальности машинного обучения;

2. Обзор литературы;
3. Обзор направлений бакалавриатов и магистратур, связанных с машинным обучением;
4. Обзор существующих курсов по машинному обучению;
5. Обзор основных алгоритмов машинного обучения;
6. Разбор данных алгоритмов;
7. Составление плана курса;
8. Создание теоретической базы;
9. Подбор примеров прикладных задач;
10. Решение подобранных задач;
11. Разработка материалов для обучения и контроля обучения;
12. Наполнение электронного курса http://school.sgu.ru/enrol/instances.php?id=200.

Методологические основы машинного обучения представлены в работах Коэльо Л.П., Ричарта В. [12], 16. Дэви С., Арно М., Мохамеда А. [16], Рашки, Мирджалили [19], Келлехера, Мак-Нейми, д`Арси [20], Жерона [21], Бенгфорта, Билбро, Охеды [22], Флаха П. [29], Загоруйко Н.Г. [30], Айвазяна С. А., Енюкова И. С., Мешалкина Л. Д. [33].

Теоретическая значимость магистерской работы заключается в изучении литературы машинного обучения, а также обзоров существующих направлений бакалавриатов и магистратур и электронных курсов.

Практическая значимость магистерской работы заключается в создании электронного курса по машинному обучению в системе Moodle.

Структура и объём работы. Магистерская работа состоит из введения, 9 разделов, заключения, списка использованных источников и 5 приложений. Общий объем работы – 161 страница, из них 131 страница – основное содержание, включая 27 рисунков и 3 таблицы, список использованных источников информации – 35 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Машинное обучение в современном мире» посвящен рассмотрению области машинного обучения и актуальности обучения специалистов в данной сфере.

В наше время машинное обучение присутствует в жизни каждого человека, даже если он об этом не догадывается. Интернет используется практически каждым, а он, в свою очередь, содержит множество платформ и сервисов, которые используют машинное обучение. Оно все больше внедряется разработчиками, чтобы упростить работу пользователя и сделать продукт удобным и востребованным.

Одно из наиболее распространенных направлений применения технологий машинного обучения – это таргетированная реклама, которая выдается пользователю, основываясь на его поисковых запросах и посещенных ресурсах. Также, машинное обучение широко используется в финансовой сфере, например, для предсказания курса акций или изменения цен на определенные валютные единицы.

Очевидно, что сфер применения машинного обучения очень много и их количество продолжает возрастать с каждым днем. Соответственно, возрастает количество вакансий в этой области. В пример можно привести рост вакансий с упоминанием машинного обучения на портале hh.ru. Количество вакансий возросло в 27 раз с 2012 по 2019 год.

Поэтому можно с уверенностью сказать, что машинное обучение является очень востребованным на рынке, и необходимость обучения специалистов для дальнейшего развития в данной сфере актуальна.

Второй раздел «Обзор направлений бакалавриатов и магистратур» посвящен обзору направлений бакалавриатов и магистратур, посвященных машинному обучению, которые существуют на данный момент.

Так как машинное обучение является одной из самых больших областей для развития и изучения в наше время, необходимо предоставлять возможности для развития для тех людей, которые хотели бы стать частью

этой сферы. Поэтому, некоторые высшие учебные заведения открывают новые факультеты/направления или вводят специальные дисциплины, связанные с машинным обучением. В основном направления магистратуры представлены в вузах Москвы и Санкт-Петербурга.

Третий раздел «Обзор курсов по машинному обучению» посвящен обзору существующих курсов по машинному обучению.

Помимо университетов существует и другая возможность начать изучать машинное обучение и развиваться в этой сфере – это различные онлайн и офлайн курсы, которые отличаются друг от друга сроками обучения, предоставленными материалами и знаниями.

Школа анализа данных «Яндекса» - своего рода офлайн курс, который содержит в себе отделения «Анализ данных», «Компьютерные науки» и «Большие данные». Предоставляет возможность заочного обучения или посещения вечерних занятий несколько раз в неделю. Для поступления необходима хорошая математическая подготовка.

Также, существует множество онлайн курсов, посвященных машинному обучению.

Но, у всех этих курсов есть свои недостатки. Некоторые курсы содержат видеолекции и транскрипции к ним, что является затруднительным для некоторых людей, которым проще, например, изучать презентации или текстовые документы. Материалы курсов в большинстве случаев отсутствуют в открытом доступе. Также, многие курсы имеют возможность прохождения только на английском языке, что делает прохождение курса и понимание информации практически невозможным для тех людей, которые английский язык не изучали. Помимо этого, курсы рассчитаны на индивидуальное обучение и не предназначены для обучения группы. Для выполнения практических заданий необходимо знание языка программирования, который используется в курсе. Другие курсы, хоть и содержат в себе большие объемы информации и множество практических задач, имеют довольно высокую цену, которая порой превышает отметку в

сто тысяч рублей. Стоит заметить, что материалы для самостоятельного изучения по разным разделам данной области присутствуют в открытом доступе.

Четвертый раздел «Подготовка к изучению машинного обучения» посвящен рассмотрению теоретических основ, языков программирования и инструментов, которые используются при изучении машинного обучения.

Перед началом изучения курса по машинному обучению рекомендуются определенные знания, которые будут использоваться при решении практических заданий.

Первое, в чем желательно разбираться при изучении машинного обучения – это базовые знания математики, такие как: линейная алгебра, теория относительности, математическая статистика, алгоритмы.

Обычно, все онлайн курсы предполагают знание какого-либо языка программирования. Конечно, каждый язык специфичен, имеет свой синтаксис и свои особенности, но знание хотя бы одного из них поможет при работе с другими. При выборе языка программирования, который будет использоваться в машинном обучении, необходимо учитывать его производительность в связи с тем, что предполагается работа с большими объемами важных данных.

Перед тем, как приступить к решению задач необходимо настроить окружение. Обычно все необходимые программы и инструменты, которые должны быть предустановлены перед началом работы, перечислены в самом курсе.

При выборе таких языков программирования, как Java, C++, Ruby, R или Scala можно использовать следующие среды разработки: JetBrains Intelliji Idea, Microsoft Visual Studio, JetBrains RubyMine, Rgui, Eclipse, RStudio. При выборе R или Python обычно используется JetBrains PyCharm, Anaconda и Jupyter Notebook.

Пятый раздел «Обзор литературы по машинному обучению» посвящен обзору литературы по машинному обучению и содержит в себе описание источников.

Так как машинное обучение и искусственный интеллект являются очень актуальными темами в наше время, имеется большое количество литературы, связанной с ними. Перед созданием электронного курса были рассмотрены основные источники.

Книга [12] описывает в основном сам Python – динамичный язык программирования, который позволяет быстро производить разведочный анализ данных и экспериментировать с ними, именно поэтому он используется при написании алгоритмов машинного обучения.

В книге [16] описывается Data Science – совокупность понятий и методов, которые позволяют придать смысл и понятный вид огромным объемам данных.

В книге [18] рассматривается практический подход к сбору данных, их обработке и дальнейшему использованию.

В книге [19] авторы описывают алгоритмы машинного обучения и анализа данных, постепенно подводя к различным сложным темам анализа данных.

В книге [20] рассмотрены информационное обучение, обучение на основе сходства, вероятностное обучение и обучение на основе ошибок.

Книга [21] – введение в теорию и практику решения задач с помощью нейронных сетей. В ней описаны основные моменты, необходимые для построения эффективных приложений.

Книга [22] направлена на обучение применения методов машинного обучения для анализа текста в реальных задачах.

В статье [23] говорится об основах машинного обучения. Но, помимо этого, приведено множество практических примеров, формул и иллюстраций, показывающих работку тех или иных алгоритмов.

Статьи [24] и [25] – это описание основ машинного обучения. Они будут полезны в начале изучения машинного обучения.

В статье [26] описываются этапы сбора данных, особенности работы на каждой фазе, а также приведены примеры для каждого этапа.

Статьи [27]-[28] описывают перспективы машинного обучения, то, какую роль машинное обучение играет в реальной жизни, и какие есть особенности практического применения в ней.

Шестой раздел «Создание электронного курса в системе Moodle» посвящен описанию содержания курса, а также самой системы Moodle.

После изучения литературы по машинному обучению был составлен план курса, который состоит из трех составляющих: изучение языка программирования Python, изучение алгоритмов машинного обучения и практические задачи машинного обучения.

Рассмотрим самостоятельно разрабатываемый курс по машинному обучению, размещенный на портале school.sgu.ru.

Данный курс может использоваться как при дистанционном обучении, так и для поддержки очного обучения.

Цель курса – изучить теоретические основы машинного обучения и научиться применять теоретические знания при решении практических задач, связанных с машинным обучением, используя язык программирования Python.

Данный курс может быть использован в ВУЗе в рамках дисциплин и практик по данной теме, а отдельные его элементы - в рамках спецкурсов по информатике для школьников старших классов. Кроме того, студенты и школьники могут использовать его для самостоятельного обучения.

Освоивший данный курс, будет

знать

1. основы языка программирования Python;
2. основные алгоритмы машинного обучения, такие как:
    - Регрессия

- Алгоритмы кластеризации
- Деревья решений
- Метод опорных векторов
- Коллаборативная фильтрация
- Наивный байесовский метод

уметь

Решать практические задачи, связанные с машинным обучением, с использованием языка программирования Python;

владеть

1. Навыками программирования на языке Python;
2. Навыками применения инструментов языка Python для решения задач машинного обучения.

Moodle (Modular Object-Oriented Dynamic Learning Environment) — система управления курсами. Представляет собой свободное веб-приложение, предоставляющее возможность создавать сайты для онлайн-обучения.

Создаваемый электронный курс находится на сайте school.sgu.ru. Система имеет множество возможностей для обучения. Помимо обыкновенных текстовых лекций в ней можно создавать интерактивные элементы, которые позволяют контролировать процесс обучения и отслеживать изменения в знаниях учеников.

Седьмой раздел «Первый раздел. Изучение языка программирования Python» содержит в себе теоретическую базу для первой части электронного курса.

Содержание первой части электронного курса:

1. Возможности языка
2. Установка Python и библиотек, используемых в машинном обучении
3. Синтаксис языка
4. Типы данных в Python
5. Числа в Python 3

6. Операторы и циклы
7. Ключевые слова
8. Встроенные функции
9. Строки в Python
10. Списки, кортежи и словари
11. Множества
12. Функции
13. Исключения и их обработка
14. Работа с файлами
15. Библиотеки в Python

Восьмой раздел «Второй раздел. Машинное обучение» содержит в себе теоретическую базу для второй части электронного курса, которая включает в себя основы теории о машинном обучении, а также основных алгоритмах машинного обучения.

Содержание второй части электронного курса:

1. Введение в машинное обучение
2. Регрессия
   2.1. Задачи для самостоятельного решения
3. Деревья решений
   3.1. Задачи для самостоятельного решения
4. Метод опорных векторов
   4.1. Задачи для самостоятельного решения
5. Наивный байесовский метод
   5.1. Задачи для самостоятельного решения
6. Алгоритмы кластеризации
   6.1. Задачи для самостоятельного решения
7. Коллаборативная фильтрация
   7.1. Задачи для самостоятельного решения

Девятый раздел «Третий раздел. Практические задачи машинного обучения» содержит в себе задачи для третей части электронного курса.

Содержание третей части электронного курса:

1. Линии разломов и шкала Рихтера
2. Определение стоимости дома
3. Рекомендательные системы

Линии разломов и шкала Рихтера.

Используемые алгоритмы: метод опорных векторов.

Постановка задачи: имеется набор данных, содержащий в себе данные о землетрясении. Необходимо нарисовать линии разломов между долготой и широтой, а также реализовать предсказание шкалы Рихтера.

Определение стоимости дома.

Используемые алгоритмы: регрессия, кластеризация.

Постановка задачи: имеется несколько наборов данных, один из которых содержит в себе статистику о проданных домах и их цене. В каждом наборе в соответствие дому ставится список критериев (наличие гаража, площадь дома и пр.). Необходимо натренировать алгоритм на определение стоимости конкретного дома (набора домов).

Рекомендательные системы.

Используемые алгоритмы: коллаборативная фильтрация.

Постановка задачи: имеется несколько наборов данных, которые содержат в себе различную информацию о фильмах. Необходимо реализовать систему рекомендации фильмов для любого выбранного пользователя.

ЗАКЛЮЧЕНИЕ

В ходе магистерской работы была изучена такая предметная область, как машинное обучение. Для этого была изучена литература на данную тему, произведен обзор существующих направлений бакалавриатов и магистратур, посвященных машинному обучению и науке о данных, а также электронных курсов по машинному обучению.

Был создан электронный курс по машинному обучению в системе Moodle http://school.sgu.ru/enrol/instances.php?id=200. Для этого был составлен план электронного курса. После составления плана была подготовлена теоретическая база на следующие темы: изучения языка программирования Python, основные алгоритмы машинного обучения. Для закрепления информации были разработаны тесты и практические задания. Для второго раздела были подготовлены разборы задач с применением каждого из алгоритмов машинного обучения и подобраны аналогичные примеры задач для самостоятельной работы. Для третей части курса были подобраны более сложные задачи машинного обучения, которые приближены к реальным задачам, которые решают специалисты в сфере машинного обучения.

Разработанный курс был апробирован в рамках проведения занятий по дисциплине «Машинное обучение» со студентами 3 курса 341 группы факультета компьютерных наук и информационных технологий направления математическое обеспечение и администрирование информационных систем Саратовского государственного университета имени Н. Г. Чернышевского.

Отдельные части магистерской работы были опубликованы в сборнике материалов X Всероссийской научно-практической конференции «Информационные технологии в образовании» [35].

Основные источники информации: Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python [12], Дэви С., Арно М., Мохамед А. Основы Data Science и Big Data. Python и наука о данных [16], Рашка, Мирджалили: Python и машинное обучение: машинное и

глубокое обучение с использованием Python, scikit-learn и Ten [19], Келлехер, Мак-Нейми, д`Арси: Основы машинного обучения для аналитического прогнозирования. Алгоритмы, рабочие примеры [20], Флах П. Машинное обучение [29].

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Frank Kane – Data Science, Deep Learning and Machine Learning with Python. [Электронный ресурс]: https://www.udemy.com/data-science-and-machine-learning-with-python-hands-on
2. Andrew Ng Machine Learning. [Электронный ресурс]: https://www.coursera.org/learn/machine-learning
3. Воронцов, Соколов – Введение в машинное обучение. [Электронный ресурс]: https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie
4. Воронцов – видеолекции «Машинное обучение». [Электронный ресурс]: https://yandexdataschool.ru/edu-process/courses/machine-learning
5. Специализация Машинное обучение и анализ данных. [Электронный ресурс]: https://www.coursera.org/specializations/machine-learning-data-analysis
6. Введение в науку о данных (An Introduction to Data Science). [Электронный ресурс]: https://www.coursera.org/learn/vvedeniye-v-nauku-o-dannykh
7. Курс по Machine Learning. [Электронный ресурс]: https://skillfactory.ru/ml-programma-machine-learning-online
8. Профессия Data Scientist с 0 до PRO. [Электронный ресурс]: https://course.skillbox.ru/profession-data-scientist-2
9. Бьёрн Страуструп. Язык программирования C++ = The C++ Programming Language / Пер. с англ. — 3-е изд. — СПб.; М.: Невский диалект — Бином, 1999. — 991 с.
10. Д. Флэнаган, Ю. Мацумото. Язык программирования Ruby = The Ruby Programming Language / пер. с англ. Н. Вильчинский. — 1-е изд. — СПб.: Питер, 2011. — 496 с.
11. Хэдли Уикем, Гарретт Гроулмунд. Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование данных = R for Data Science: Visualize, Model, Transform, Tidy, and Import Data. — Вильямс, 2017. — 592 с.
12. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на

языке Python. – 2-е издание, пер. с англ. Слинкин А.А. - М.: ДМК Пресс, 2016. - 302 с.: ил.

13. Герберт Шилдт. Java. Полное руководство, 10-е издание = Java. The Complete Reference, 10th Edition. — М.: «Диалектика», 2018. — 1488 с.

14. Прокопец А. Конкурентное программирование на SCALA. — ДМК пресс, 2017. — 342 с.

15. I. H. Witten, E. Frank Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). — Morgan Kaufmann, 2005

16. Дэви С., Арно М., Мохамед А. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил. – (Серия «Библиотека программиста»)

17. Witten I.H., Frank E. Data mining: practical machine learning tools and techniques. – 2nd ed. p. cm. – (Morgan Kaufmann series in data management systems), 2005

18. Murphy K.P. Machine learning: a probabilistic perspective – The MIT Press Cambridge, Massachusetts London, England, 2012

19. Рашка, Мирджалили: Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и Ten. – М: Вильямс, 2019 – 656 с.

20. Келлехер, Мак-Нейми, д`Арси: Основы машинного обучения для аналитического прогнозирования. Алгоритмы, рабочие примеры. – М: Вилямс, 2019 – 656 с.

21. Орельен Жерон: Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. Концепции, инструменты и техники. – М: Вильямс, 2018 – 688 с.

22. Бенгфорт, Билбро, Охеда: Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки. – СПб.: Питер, 2019 – 368 с.

23. Stephanie Yee, Tony Chu. A visual introduction to machine learning, 2015

24. Pedro Domingos. A Few Useful Things to Know about Machine Learning,

2013

25. Philip Guo. Data Science Workflow: Overview and Challenges, October 30, 2013

26. D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young. Machine Learning: The High-Interest Credit Card of Technical Debt, 2014

27. John Forman. The perilous world of machine learning for fun and profit pipeline jungles and hidden feedback loops, 2016

28. Mark Lutz. Learning Python: Powerful Object-Oriented Programming. O'Reilly Media; 5 edition, 2013. – 1650 c.

29. Флах П. Машинное обучение. — М.: ДМК Пресс, 2015. — 400 с.

30. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

31. Zachary Chase Lipton. The High Cost of Maintaining Machine Learning Systems, 2015. Нейт Сильвер. Сигнал и Шум. Почему одни прогнозы сбываются, а другие — нет // Азбука-Аттикус, КоЛибри, 2015

32. Bernhard Schölkopf, Alexander J. Smola Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond. — MIT Press, Cambridge, MA, 2002

33. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983.

34. William M. Bolstad. Introduction to Bayesian Statistics, 2nd Edition // Wiley-Interscience; 2nd edition.

35. Информационные технологии в образовании. Материалы X Всероссийской научно-практической конференции. – Саратов: ООО Издательский центр "Наука", 2018. – 452 с.