Опыт преподавания основ машинного обучения для бакалавров направления «Математическое обеспечение и администрирование информационных систем»

Казачкова А.А.[1], Лапшева Е.Е.[2]

[1]kazachkova.anna@gmail.com, [2]lapsheva@yandex.ru

Саратовский государственный университет имени Н.Г. Чернышевского

Статья посвящена содержанию и особенностям проведения практических занятий по дисциплине «Основы машинного обучения» у студентов направления «Математической обеспечение и администрирование информационных систем».

Ключевые слова: машинное обучение, python, jupyter notebook, google colab.

С большими данными (Big Data) и машинным обучением современный человек сталкивается все чаще. При обращении в банк или общении в соцсетях алгоритмы машинного обучения помогают сделать это взаимодействие удобнее, эффективнее и безопаснее. Машинное обучение и связанные с ним технологии быстро развиваются. Встает вопрос о подготовке специалистов, готовых создавать такие технологии и делать их более быстрыми, адаптивными к изменяющимся условиям современной жизни.

С 2019-2020 учебного года в программе бакалавриата по направлению «Математическое обеспечение и администрирование информационных систем» введена дисциплина, которая называлась «Основы машинного обучения», а затем – «Машинное обучение и анализ данных». Это трёхсеместровый курс, включающий в себя 432 часа, из них 188 аудиторных. Введение этой дисциплины в программу бакалавриата, к сожалению, совпало с пандемией COVID-19, что наложило отпечаток на методику проведения практических занятий.

В процессе подготовки заданий и материалов были рассмотрены различные общедоступные ресурсы: учебные курсы, научно-популярные статьи, подборки алгоритмов и наборов данных. Их количество с каждым годом растёт, однако не все из них могут быть рекомендованы для использования в качестве учебного или дополнительного материала. Все эти ресурсы можно разделить на две группы: первые рассматривают машинное обучения, отталкиваясь от подробного математического описания алгоритмов и моделей, а вторые, напротив, описывают способы применения готовых библиотек. Каждый из этих подходов имеет свои достоинства и слабые стороны. Поэтому важной и сложной задачей было найти сбалансированный вариант, позволяющий не перегружая студентов теоретическими выкладками и сосредоточившись на практической составляющей, тем не менее, продемонстрировать суть, принципы работы и особенности изучаемых алгоритмов.

В качестве среды для работы и подготовки отчётности был выбран облачный ресурс Google Colab [1], который основан на браузерной среде разработки Jupyter Notebook [2]. Данная среда имеет интерактивный интерфейс для языка Питон, в котором исполняемые блоки кода сочетаются с результатами их работы и текстовыми пояснениями в формате Markdown. Причём результатами работы могут быть не только выводимые в консоль текстовые и числовые данные, но также и графики, диаграммы и иллюстрации. Выбор Google Colab обоснован несколькими причинами: доступность среды как из университетских аудиторий, так и с домашних устройств, отсутствие необходимости установки не только самого языка программирования, но и библиотек, требующих объёмов дискового пространства и времени на установку, возможность запуска обучения разработанных моделей не только на CPU, но и на GPU серверов Google. Готовые решения студенты загружают в специально созданные задания курса на портале school.sgu.ru (LMS Moodle [3]).

В качестве языка программирования был выбран Питон 3 [4] как общепризнанный первоклассный инструмент для научных вычислений, включая анализ и визуализацию больших наборов данных, что обеспечивается большой и активно развивающейся экосистемой модулей, в том числе созданных сторонними разработчиками, начиная с numpy, pandas, matplotlib, plotly, scikit-learn [5]. На знакомство с основами синтаксиса и базовыми типами языка отводится 14 часов, сюда включается освоение базовых алгоритмических конструкций и практики применения коллекций, функций и возможностей стандартных библиотек. Данный раздел поддерживается автоматической системой проверки решений на портале school.sgu.ru [6]. В настоящий момент ведётся работа по расширению пула задач,

ориентированных на применение библиотеки numpy. Индивидуальные задания на использование matplotlib и pandas проверяются преподавателями вручную. Первый семестр завершается выполнением творческого задания, предполагающего самостоятельный поиск набора данных, его обработку и анализ с использованием изученных библиотек и алгоритмов.

Задания второго семестра имеют исследовательский характер. Первое задание заключается в подтверждении или опровержении двух гипотез относительно данных обезличенных результатов ОГЭ по информатике. В ходе выполнения задания осуществляется повторение возможностей модуля pandas. Второе задание также не предполагает знания специальных алгоритмов машинного обучения и состоит в создании собственного алгоритма предсказания оценок ОГЭ из предыдущего задания. В ходе выполнения данного задания студенты получают представление о релевантности данных, их полноте, сложности определения зависимостей. Третье задание связано с другим набором данных, где требуется провести разбиение на два класса и на большее количество классов по каким-либо самостоятельно выбранным характеристикам и их комбинациям. Третье задание напрямую связано с четвёртым, в котором происходит знакомство с методом ближайших соседей, метрикой близости, нормализацией. В данном задании студенты применяют изученный метод сначала на своих данных, затем на разбиении одногруппника и, наконец, на данных из задания. Тема пятого задания – деревья решений. Задание предполагает не только знакомство с этим методом, в том числе визуализацию и расшифровку результатов, но и сравнение его с предыдущим, определения плюсов и особенностей каждого из них. Заключительное задание этого семестра связано с задачей регрессии.

По результатам проведённых занятий был сделан ряд наблюдений и выводов:

– наибольшие трудности у студентов вызывает интерпретация результатов исследований и формулировка общих выводов;

– отмечается систематическое несоблюдение студентами требований к оформлению кода ввиду работы в среде jupyter, что приводит к проблемам оформления кода в бакалаврских дипломах;

– можно отметить трудности поиска данных для учебных исследований, т.к. общедоступные данные хорошо изучены и готовые модели представлены в сети в открытом доступе;

– курс требует хорошей математической базы и готовности к самостоятельной работе и поиску информации.

Курс постоянно обновляется и дополняется. В данный момент разрабатывается блок заданий третьего семестра курса, посвящённого нейросетям.

Список литературы

[1] Google Colaboratory [Электронный ресурс]. URL: https://colab.research.google.com/ (дата обращения 27.09.2021)

[2] Jupyter Notebook [Электронный ресурс]. URL: https://jupyter.org/ (дата обращения 27.09.2021)

[3] *Кудрина Е.В., Лапшева Е.Е., Огнева М.В.* Развитие образовательного портала обучения алгоритмизации и программирования саратовского государственного университета. / В сборнике: Информационные технологии в общем образовании ("ИТО-Саратов-2009"). Сборник трудов участников конференции. 2009. С. 162-165.
[4] Python documentation [Электронный ресурс]. URL: https://docs.python.org/3/ (дата обращения 27.09.2021)
[5] *Вандер П.Дж.* Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018
[6] *Казачкова А.А.* Тестирование и отладка программ при использовании автоматической проверки решений / В сборнике: Компьютерные науки и информационные технологии. Материалы Международной научной конференции. 2012. С. 131-133.