ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.8

СОВРЕМЕННЫЕ ТЕХНОЛОГИИ МАШИННОГО ОБУЧЕНИЯ

БАРЩЕВСКИЙ Евгений Георгиевич
кандидат технических наук, профессор
ФГБОУ ВО «Государственный университет морского и речного флота
им. адмирала С.О. Макарова»
г. Санкт-Петербург, Россия

Актуальность работы обусловлена широким использованием машинного обучения в различных сферах науки, техники, человеческой деятельности. В статье рассматривается состав машинного обучения, категории машинного обучения, работа машинного обучения, программные инструменты машинного обучения.

Ключевые слова: машинное обучение, категория машинного обучения, кластеризация, классификация.

Введение (Introduction). Машинное обучение – это особый способ обучать компьютер решать определенные задачи без применения программирования. На международной конференции по искусственному интеллекту и анализу данных Artificial Intelligence Journey (AI Journey) президент по глобальным продажам, маркетингу и операциям Microsoft Жан-Филипп Куртуа сообщил, что пандемия COVID-19 форсировала интерес к использованию машинного обучения: 80% компаний уже внедряют его в свою деятельность, а 56% планируют увеличить объем инвестиций в эту сферу [1; 6].

Современное машинное обучение состоит из трех основных частей:

1. Алгоритмы, которые подсказывают компьютеру, какие источники требуется использовать, чтобы получить правильное решение задачи.

2. Наборы данных или датасеты. Это память машины, в которой находится информация о предыдущем опыте решения задачи.

3. Признаки – индивидуальные параметры.

Стоит более подробно рассмотреть, как работает машинное обучение.

Методы и материалы (Methods and Materials). Типы машинного обучения. Принято разделять все типы машинного обучения на три категории:

– с учителем (supervised learning);

– без учителя (unsupervised learning);

– с подкреплением (reinforcement learning).

На практике в настоящий момент реализованы три ключевых области машинного обучения:

Рекомендательные системы. Рекомендательные системы – это наиболее узнаваемая модель машинного обучения, из используемых сегодня. Вы видите сервисы или сайты, которые пытаются рекомендовать книги или фильмы, статьи, базируясь на ваших предыдущих действиях. Они пытаются выводить вкусы и предпочтения, и идентифицировать неизвестные предметы, которые представляют интерес. К рекомендательным относятся следующие системы:

Amazon.com это, возможно, наиболее известный сайт в электронной коммерции применивший рекомендации. Основываясь на покупках и активности на сайте, Amazon.com рекомендует книги и другие вещи, которые могут вызвать интерес.

Netflix также рекомендует DVD, которые могут быть интересны и предлагают приз в 1M$ для исследователей, которые могут улучшить качество их рекомендаций.

Социальные сети, такие как Фейсбук, используют варианты рекомендательных техник для выявления людей, наиболее вероятно подходящих под определение «еще не связанных друзей».

Кластеризация. Кластеризация менее очевидна, но оказывается в не менее известных упоминаниях. Как следует из названия, методы кластеризации пытаются группировать большие числа предметов вместе в кластеры, которые имеют общее сходство. Таким образом, устанавливают иерархию и порядок в больших или трудных для понимания множествах данных, и таким способом устанавливают интересные закономерности или делают набор данных более легким для понимания.

Google News группирует новостные статьи по названию, используя технику кластеризации.

Поисковые механизмы, такие как Clusty, так же группируют свои поисковые результаты.

Заказчики могут быть сгруппированы в сегменты (кластера) при помощи техники кластеризации, основанные на атрибутах: доход, местоположение, покупательские привычки.

Кластеризация помогает определять структуру и даже иерархию, в большой коллекции вещей, которую, может быть, даже сложно осмыслить. Предприятия могут использовать эту технику для определения скрытых групп среди пользователей, или разумной организации большой коллекции документов, или определения общих паттернов, использования для сайтов, используя их логи [2].

Классификация.

Модели классификации позволяют решать является или нет предмет частью определенной категории или есть ли у нее некоторый атрибут.

Yahoo!, Mail решают является или нет входящее сообщение спамом, основываясь на предшествующих письмах и сообщений на спам от пользователей, а также характеристиках самих писем.

Google's Picasa и другие приложения для управления фотографиями могут определять область изображения содержащую человеческое лицо.

Программа оптического распознавания символов классифицирует малые области отсканированного текста на отдельные символы.

Классификация помогает решить вопрос о том, соответствует ли новый кусок вводных данных или предмет предыдущим рассмотренным шаблонам; и она часто используется для классификации поведения или шаблона. Это может быть использовано для обнаружения подозрительной сетевой активности или мошенничества. А также для выяснения того, указывает ли на разочарование или на удовлетворение сообщение пользователя. Каждая из этих моделей работает лучше, когда снабжена большим количеством хороших входных данных. В некоторых случаях, эти методы должны не только работать на больших объемах данных, но должны получать результат быстро, и эти факторы делают масштабируемость главной задачей. Одна из основных причин использовать Mahout – это именно масштабируемость. Как неоднократно отмечается в книге, нет готового рецепта который можно взять и применить к типовой ситуации. Для каждого случая нужно пробовать различные алгоритмы и входные данные. Только поняв суть алгоритмов можно успешно применять библиотеку [7].

Для обучения используют различные программные инструменты. Наиболее продвинутыми считаются [4; 5; 7]:

– TensorFlow;
– Shogun;
– io;
– Rapid Miner;
– Google Cloud ML Engine;
– Amazon Machine Learning (AML);
– NET;
– Apache Mahout;
– Microsoft Azure ML;
– SberCloud ML Space.

Все они имеют свои особенности. Основные отличия в применяемых языках программирования и совместимости с определенными операционными системами [3]. Каждый инструмент заточен для решения узконаправленных задач. Стоит более подробно рассмотреть некоторые модели обучения поискового робота.

TensorFlow – открытая программная библиотека для машинного обучения, разработанная компанией Google для решения задач

построения и тренировки, которая позволяет обучать искусственный интеллект решению разных задач. Используется для охоты на новые планеты, предотвращения слепоты, помогая врачам сканировать диабетическую ретинопатию и спасения лесов, предупреждая власти о признаках незаконной вырубки леса. Это то, на чем строятся AlphaGo и Google Cloud Vision [7], io – доменная зона в IT-сфере. Появилась даже новая расшифровка аббревиатуры – Input/Output. А еще .IO может означать «Internet Organization». Словом, идеальный вариант для стартапов и медиа, сайтов на тему инноваций и технологий.

RapidMiner – это программная многопользовательская платформа, которая представляет собой интегрированную среду для обработки данных в больших информационных массивах, машинного обучения, текстовой аналитики и построения прогностических моделей, а также для решения иных задач Data Mining.

Сервис Google Cloud Search позволяет легко находить нужную для работы информацию с помощью ноутбука, мобильного телефона или планшета. Поиск выполняется по корпоративному контенту в сервисах Google Workspace или в сторонних источниках данных. предоставляемый компанией Google набор облачных служб, которые выполняются на той же самой инфраструктуре, которую Google использует для своих продуктов, предназначенных для конечных потребителей, таких как Google Search и YouTube.

Amazon Machine Learning (AML) – это зонтичный термин, объединяющий различные облачные платформы, решающие большинство инфраструктурных задач, включая предварительную обработку данных, обучение и оценку моделей с дальнейшим созданием прогнозов. AWS поддерживает вас на каждом этапе перехода к использованию машинного обучения с помощью самого универсального набора сервисов искусственного интеллекта и машинного обучения, инфраструктуры и ресурсов для внедрения.

Mahout это *opensource* библиотека для машинного обучения от Apache. Алгоритмы, которые библиотека реализует в совокупности можно назвать машинным обучением или коллективным интеллектом. Это может означать многое, но в настоящий момент это означает в первую очередь рекомендательные системы (коллаборативная фильтрация), кластеризацию и классификацию.

Mahout содержит ряд моделей и алгоритмов, многие все еще в разработке или экспериментальной фазе (алгоритмы). На этом раннем этапе жизни проекта, три ключевые темы наиболее заметны: рекомендательные системы (коллаборативная фильтрация), кластеризация и классификация. Это далеко не все что есть в Mahout, но эти темы наиболее заметные и зрелые.

В теории Mahout – это проект, открытый для реализации любого вида моделей машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. Создание умного сочинение в стиле Gmail с помощью языковой модели char ngram. – URL:https://towardsdatascience.com/gmail-style-smart-compose-using-char-n-gram-language-models-a73c09550447 (дата обращения: 05.04.2022).
2. 25 Google Search Statistics to Bookmark ASAP. – URL:https://blog.hubspot.com/marketing/google-search-statistics (дата обращения: 05.04.2022).
3. Google Mobile Blog. Voice Search arrives in 13 new languages. – URL:https://blog.google/products/search/voice-search-arrives-in-13-new-languages (дата обращения: 05.04.2022).
4. Marketing Artificial Intelligence. – URL:https://www.marketingaiinstitute.com/blog/how-search-engines-use-artificial-intelligence. (дата обращения: 05.04.2022).
5. Search Personalization Using Machine Learning. – URL:http://faculty.washington.edu/hemay/search_personalization.pdf (дата обращения: 05.04.2022).
6. *Whitby B.* Artificial Intelligence: A Beginner's Guide. London: Oneworld Publications, 2008. 192 p.
7. *Zhang L.* Sentiment Analysis and Opinion Mining / L. Zhang, B. Liu. – Boston: Springer, 2017. 905 p. DOI: 10.1007/978-1-4899-7687-1_907.

UDC 004.8

MODERN MACHINE LEARNING TECHNOLOGIES

BARSHCHEVSKY Evgeny Georgievich
Candidate of Sciences in Technology, Professor
State University of the Sea and River Fleet named after Admiral S.O. Makarov
St. Peterburg, Russia

The relevance of the work is due to the widespread use of machine learning in various fields of science, technology, and human activity. The article discusses the composition of machine learning, categories of machine learning, the work of machine learning, machine learning software tools.
Keywords: machine learning, machine learning category, clustering, classification.

СИСТЕМЫ КОНТРОЛЯ ВЕРСИЙ ДЛЯ ОБУЧЕНИЯ ПРОГРАММИРОВАНИЮ

БУНЬКИН Виктор Иванович
кандидат технических наук, доцент кафедры цифровой экономики
НОЧУ ВО «Московский финансово-промышленный университет «Синергия»
г. Москва, Россия

В статье приводится описание двух популярных средств контроля версий – СКВ git и облачного онлайн-сервиса GitHub. Дается краткая справка по данным системам и рекомендации по их использованию при изучении различных дисциплин, связанных с программированием.
Ключевые слова: система контроля версий, технологии программирования, *СКВ* git, GitHub.

При изучении программирования, а также других вопросов, связанных с бурно развивающимися информационными технологиями, стоит задача не только получить соответствующие знания, но и научиться навыкам работы с такими инструментальными средствами, которые в своей работе активно используют профессионалы.

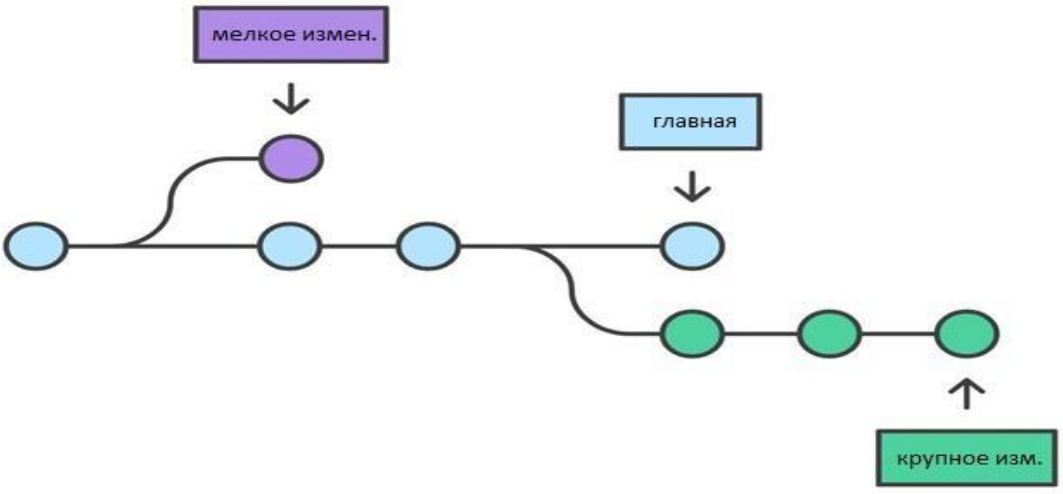

Рисунок 1. **Основные приемы работы с СКВ git**