УДК 004.85

МАШИННОЕ ОБУЧЕНИЕ КАК ИНСТРУМЕНТ АНАЛИЗА ДАННЫХ

[1]Кондратёнок Е. В., [2]Макареня С. Н.
[1]Белорусский национальный технический университет, Минск, Беларусь, elena_kondr@tut.by
[2]Белорусский национальный технический университет, Минск, Беларусь makarenya@bntu.by

Машинное обучение считается одним из самых больших достижений со времен микрочипа и является самой динамической и прогрессивной формой искусственного интеллекта. «Машинное обучение – это наука о том, как заставить компьютеры учиться и действовать так, как это делают люди, и совершенствовать свое обучение с течением времени автономно, предоставляя им данные и информацию в форме наблюдений и взаимодействий в реальном мире». – Дэн Фагелла [1].

Машинное обучение представляет собой подраздел искусственного интеллекта, стоящий на стыке таких дисциплин, как математика, статистика, теория вероятностей, теория графов и изучающий алгоритмы, способные самостоятельно обучаться на основе опыта. В процессе машинного обучения алгоритмы учатся поиску закономерностей и корреляций в больших наборах данных, а также принятию оптимальных решений и созданию прогнозов на основе этого анализа. Модели машинного обучения улучшаются по мере использования и становятся точнее по мере роста объема доступных данных.

Успехи в машинном обучении во многом обусловлены достижениями в области программного и аппаратного обеспечения и возросшей вычислительной мощности для запуска алгоритмов и специальных ресурсов для параллельной работы программ. Ведущими ИТ-компаниями разработаны сложные и мощные алгоритмы машинного обучения. На определенных данных тренируют алгоритмы, а затем используют для нахождения решения путем комплексного использования статистических данных, из которых выводятся закономерности и на основе которых делаются прогнозы.

Без явного программирования на языках высокого уровня, таких как Java и C++ машинное обучение позволяет машине учиться на большом количестве примеров, опыте и практике, выявляет закономерности и использует их, чтобы прогнозировать характеристики новых данных. Вместо написания кода, данные передаются в общий алгоритм, и алгоритм строит логику на основе этих данных.

При традиционном программировании есть данные и правила, выраженные на языке программирования, которые и составляют основную часть кода. Правила преобразуют данные и дают ответ.

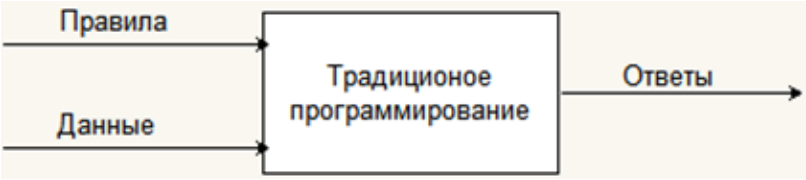

Рисунок 1 – Традиционное программирование

Целью машинного обучения является – создание точной математической модели, которая на основе выборочных данных известных как обучающая выборка, вырабатывает правила прогнозирования и принятия решений.

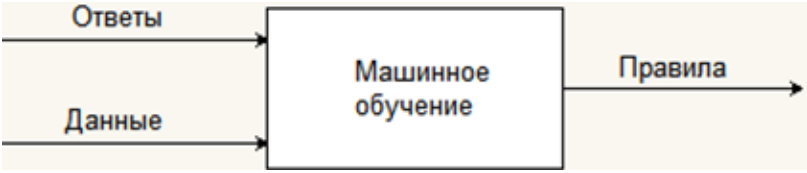

Рисунок 2 – Машинное обучение

Процесс машинного обучения включает следующие этапы:

Постановка задачи.

1. Сбор данных. Этап включает в себя сбор всех соответствующих данных из различных источников. Данные не должны содержать ошибок и должны быть релевантными. Способ сбора данных зависит от решаемой задачи.

2. Обработка данных. Это процесс очистки и преобразования необработанных данных в определенный формат. Нерелевантные

данные необходимо удалить. Пригодность данных к использованию и достоверность результатов прямым образом зависит от их правильной подготовки. После очищения и преобразования в определенный формат данные анализируются. В зависимости от размера набора данных выделяется обучающая выборка. Обучающая выборка делится на две группы: на первой обучается алгоритм, а вторая для оценки работы алгоритма – тестовая выборка.

3. Обучение алгоритма (моделирование). Под моделированием понимается использование алгоритма машинного обучения для поиска информации в собранных данных. На этом этапе происходит на поиск математической функции, которая точно выполнит указанную задачу. Обучение зависит от типа используемой модели. В простой линейной модели обучением является построение линий; для алгоритма случайного леса необходимо построить дерево принятия решений. Корректировка алгоритма происходит при изменении ответов. Алгоритм использует только часть данных, обрабатывает их, замеряет эффективность обработки и автоматически регулирует свои параметры до тех пор, пока не сможет последовательно производить желаемый результат с достаточной достоверностью. Эффективность алгоритма оценивается на тестовой выборке. Дополнительная корректировка алгоритма производится при необходимости. Работа алгоритма на тестовой выборке позволяет предотвратить переобучение. Это явление, при котором алгоритм хорошо работает только на обучающей выборке. После этих действий модель готова.

4. Развертывание. Когда скорость и точность работы модели приемлемы, модель должна быть развернута в реальной системе.

Этот процесс циклический. Можно начать проект со сбора данных, смоделировать их. Понять, что собранных данных недостаточно и вернуться к сбору данных. Смоделировать данные снова, найти хорошую модель, развернуть ее, обнаружить, что она не работает, создать другую модель, развернуть и ее, обнаружить, что она тоже не работает, и вернуться к сбору данных. Модели адаптивно улучшают свою производительность по мере увеличения количества образцов данных для изучения. Стандартная методология машинного обучения (Cross Industry Standard Process for Data Mining, CRISP-DM) представлена на рисунке 3.

Существует три основных типа машинного обучения:

1. Обучение с учителем;
2. Обучение без учителя;
3. Обучение с подкреплением.

Рисунок 3 – Стандартная методология машинного обучения (Cross Industry Standard Process for Data Mining, CRISP-DM)

Обучение с учителем. Данный тип машинного обучения посвящен решению задач, в которых данные это множество объектов или ситуаций и множество ответов или откликов. Зависимость между ответами и объектами неизвестна. Конечная совокупность «объект-ответ» называется обучающей выборкой. На основе обучающей выборки строится алгоритм, способный найти зависимости и дать точный ответ для нового набора объектов без ответов.

Два основных применения обучения с учителем: классификация и регрессия.

В задачах классификации множество ответов (меток класса) конечно. Цель классификации состоит в предсказании категориальных меток классов (дискретных, неупорядоченных значений, членства в группах) новых данных на основе обучающих данных.

В задачах регрессии ответами являются действительные числа или вектора.

Пример линейной регрессии: подбирается прямая линия с учетом x и y, которая с некоторыми критериями (например, среднеквадратичное расстояние) минимизирует расстояние между точками обучающей выборки и подобранной линией. Ответ на

тестовых данных предсказывается, опираясь на перехваченный и изученный наклон линии.

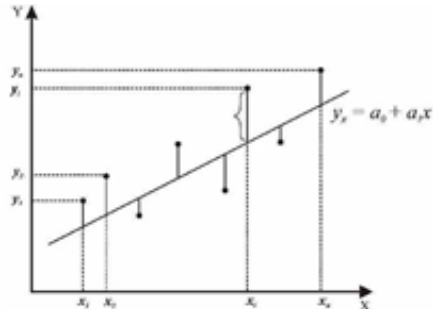

Рисунок 4 – Пример линейной регрессии

Обучение без учителя. Данный тип машинного обучения посвящен решению задач обработки данных, в которых известны только описания множества объектов (обучающая выборка без ответов). Требуется найти взаимосвязи, зависимости и закономерности между объектами.

Существует две основные задачи обучения: кластеризация и уменьшение размерности.

В задачах кластеризации обучающая выборка разбивается на непересекающиеся подмножества (кластеры). Каждый кластер состоит из похожих объектов, а объекты разных кластеров существенно отличаются. При решении задач классификации и регрессии кластеризация позволяет упростить обработку данных и принятия решений. После разбиения к каждому кластеру применяется свой метод анализа.

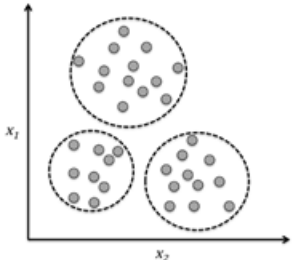

Рисунок 5 – Пример кластеризации

В категории уменьшения размерности работают с обучающей выборкой высокой размерности. Это является проблемой для вычислительной производительности алгоритмов машинного обучения. В связи с этим задача состоит в том, чтобы представить эти данные в пространстве меньшей размерности, минимизировав потери информации. Обучающую выборку сокращают, оставив по одному наиболее типичному представителю от каждого кластера.

Обучение с подкреплением – тип машинного обучения, предполагающий обучение на практике. Обучение без учителя и обучение с учителем предполагают пассивную передачу входных данных и обнаружение в них структур. Для обучения с подкреплением используются агенты обучения, которые обеспечивают активное принятие решений и обучение на собственных результатах.

Существует много алгоритмов машинного обучения, на основе которых строится модель. Выбор алгоритма зависит от характеристик набора данных, таких как объем, структура и качество. Кроме этого на выбор алгоритма влияет желаемый результат, требуемая точность предсказания и время, необходимое для обучения модели.

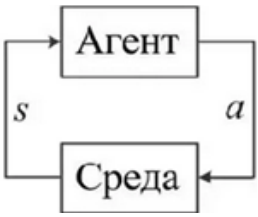

Рисунок 6 – Обучение с подкреплением

СПИСОК ЛИТЕРАТУРЫ

1. Towards data science [Электронный ресурс] / Machine Learning Introduction: A Comprehensive Guide – Режим доступа: https://towardsdatascience.com/machine-learning-introduction-a-comprehensive-guide-af6712cf68a3 / – Дата доступа: 15.11.2021.

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

3. Бессмертный И. А. Интеллектуальные системы, 2018 .

4. Машинное обучение, нейронные сети, искусственный интеллект [Электронный ресурс] / Введение в машинное обучение: полное руководство – Режим доступа: https://www.machinelearningmastery.ru/machine-learning-introduction-a-comprehensive-guide-af6712cf68a3/ – Дата доступа: 15.11.2021.