УДК 04 **Информационные технологии**

***Матвеев Андрей Николаевич**, студент, каф. МОЭВМ, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург*

e-mail: andrewsamsunguser@gmail.com

ИНСТРУМЕНТЫ ПОСТРОЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Аннотация: В статье рассматриваются инструменты построения моделей машинного обучения. Приводятся формальная задача машинного обучения, классификация методов и моделей машинного обучения по способу обучения и алгоритм работы моделей. Отмечается, что результат работы моделей машинного обучения зависит не только от входного набора данных, но и от конфигурации. Рассматриваются особенности и возможности автоматизированного машинного обучения, приводятся основные методы оптимизации гиперпараметров.

Ключевые слова: машинное обучение, модели машинного обучения, алгоритмы машинного обучения, автоматизированное машинное обучение, оптимизация гиперпараметров.

Annotation: The article discusses tools for building machine learning models. A formal task of machine learning, a classification of methods and models of machine learning according to the method of learning, and an algorithm for the operation of models are given. It is noted that the result of machine learning models depends not only on the input data set, but also on the configuration. The features and capabilities of automated machine learning are considered, the main methods for optimizing hyperparameters are given.

Keywords: machine learning, machine learning models, machine learning

algorithms, automated machine learning, hyperparameter optimization.

Машинное обучение (далее – ML) представляет собой раздел науки об искусственном интеллекте, изучающий способы решения прикладных задач посредством обнаружения связей и закономерностей в наборе исходных массивов данных при помощи алгоритмов и моделей, основанных на статистических и вычислительных методах [1]. Главными особенностями ML являются отсутствие чёткой инструкции, которой следует разработанная математическая модель, использование больших, обычно неструктурированных данных и преобладание принципа чёрного ящика в процессе работы модели. Модели ML используются для оптимизации широкого круга процессов, что делает актуальным исследование особенностей их построения.

Целью работы является изучение инструментов построения моделей ML. Для её достижения были использованы аналитический, синтетический, индуктивный и дедуктивный методы обработки тематических исследований, научных публикаций и релевантных литературных источников.

Формально задачу ML можно представить следующим образом [2]. Пусть есть множество X объектов и множество Y ответов. Предполагается наличие неизвестной функциональной зависимости $f : X \to Y$ между объектами и ответами. Известна совокупность S пар вида (объект, ответ), называемая обучающей выборкой:

$$S = \{(x_i, y_{x_i} = f(x_i)) \in X \times Y | i = 1, \dots, l\}.$$

Требуется найти приближенный вид f при помощи построения такой аппроксимирующей функции $aS : X \to Y$, что $\forall\, x \in X,\, aS(x) \approx f(x)$.

Все методы ML можно классифицировать по способу обучения [3]. Основными группами методов являются обучение с учителем (методы классификации) и без учителя (методы кластерного анализа). В случае обучения с учителем для каждого объекта из обучающей выборки задаётся пара «прецедент-решение», и задача сводится к определению класса объекта по его признакам. При обучении без учителя обучающая выборка не содержит заранее

известных ответов для набора прецедентов, и от модели требуется самостоятельно объединить данные в кластер, обобщающий эти данные.

К моделям классификации относятся деревья решений, нейронные сети, решающие правила, алгоритмы покрытия и другие [4]. Наиболее наглядной и легко интерпретируемой является модель на основе решающего дерева. Методы кластеризации делятся на иерархические и неиерархические. Методы иерархической кластеризации могут быть агломеративными, предполагающими последовательное объединение исходных объектов отдельных кластеров и уменьшение числа кластеров, и дивизимными, в которых в начале работы алгоритма все объекты принадлежат одному кластеру. Методы неиерархической кластеризации представляют собой итеративные методы разделения исходной выборки, при котором новые кластеры формируются до момента выполнения правила остановки.

Алгоритм работы любой модели ML можно разбить на следующие этапы [5]:

1. Сбор данных. На этом этапе собираются материалы для обучения и помещаются в единый источник – электронную таблицу, текстовый файл или базу данных.

2. Исследование и подготовка данных. Этап подразумевает исправление или очистку загрязнённых данных, удаление ненужной информации и перекодирование сведений в соответствии с задачами, которые ставятся перед моделью.

3. Обучение модели. Алгоритм обучения подбирается с учётом специфики данных анализа и задачи ML.

4. Оценка модели. Поскольку любая модель ML решает задачу обучения предвзято, после получения результата важно оценить, насколько хорошо алгоритм учитывает собственный опыт. Методы оценки точности зависят от типа используемой модели: для одних моделей подойдут тестовые наборы данных, для других требуется подбор специфичных для предполагаемой области использования показателей производительности.

5. Улучшение модели. Для повышения производительности требуется использование улучшенных стратегий, дополнение или дополнительная подготовка данных. В ряде случаев требуется смена модели ML.

Результат, который выдают модели ML, зависит не только от входного набора данных, но и от конфигурации: при разной конфигурации, то есть гиперпараметрах алгоритма, на одном наборе данных будет получаться разный результат [6]. На практике выбор алгоритма и настройка гиперпараметров зачастую выполняется вручную, что требует значительных трудовых и временных затрат. Однако в условиях увеличения вычислительных мощностей компьютеров развитие получают системы автоматизированного ML – AutoML, которые помогают автоматизировать основные процессы обучения модели ML с сохранением предсказательной точности целевой модели [7]:

- подготовка данных – интеграция, преобразование, очистка и сокращение данных;
- извлечение и выбор признаков – выбор подмножества признаков из набора данных, улучшающий обобщение обучения;
- выбор алгоритма, обеспечивающего наиболее точные результаты;
- оптимизация гиперпараметров – настройка параметров алгоритмов для дальнейшего улучшения результатов.

Системы AutoML используют различные методы и приёмы оптимизации гиперпараметров для достижения желаемой точности и производительности модели. Основными методами оптимизации гиперпараметров являются [8, 9]:

1. Полный перебор значений в пространстве гиперпараметров. При таком подходе используются метрика оценки производительности и кросс-валидация, обеспечивающая устойчивость этой оценки. Поскольку гиперпараметры могут иметь значения непрерывного типа, возможно применение эвристических ограничений или дискретизации.

2. Случайный поиск. При таком подходе оценивается заданное число случайным образом сгенерированных наборов значений гиперпараметров. Дискретизация не требуется. Производительность метода можно повышать

посредством априорных знаний, используя определённое распределение.

3. Байесовская оптимизация. Метод основан на обращении к функции чёрного ящика с шумом. Он позволяет адаптировать поиск на основе значимости всех гиперпараметров для конкретных задач: в процессе байесовской оптимизации разрабатывается вероятностная модель зависимости значений оценивающей качество результатов метрики от значений гиперпараметров. При каждой итерации, число которых задаётся вручную, используются наиболее предпочтительные значения гиперпараметров, после чего вероятностная модель обновляется на основе результатов обучения. Байесовская оптимизация позволяет предсказывать производительность до выполнения обучения.

4. Оптимизация на основе градиента. Подходит для методов, пространство значений гиперпараметров которых можно представить в виде градиента.

5. Эволюционная оптимизация. В основе метода лежит генерация случайных популяций наборов значений гиперпараметров, которые итеративно изменяются и отбираются на основе производительности до прекращения её увеличения либо до достижения желаемой величины.

6. Метапризнаки наборов данных. Подход предполагает анализ метапризнаков образцовых наборов данных, вычисление сходства нового набора данных с образцовым и выбор алгоритмов ML, оценённых на наиболее схожих образцовых наборах, данных как наиболее производительные.

7. Метод ансамблей. Метод заключается в обучении нескольких алгоритмов для получения комбинированной модели, которая даёт более качественные результаты. Качество комбинированной модели повышается при задействовании разнообразных моделей.

В число наиболее распространённых систем AutoML входят [10]:

1. Auto-Weka. Алгоритм использует метод байесовской оптимизации со случайными лесами, обрабатывающий дискретные входные данные большой размерности.

2. Auto-Sklearn. Система использует 4 метода предварительной обработки, 14 методов обработки признаков данных, 15 алгоритмов классификации и метод

байесовской оптимизации для настройки гиперпараметров.

3. TPOT. В основе системы лежит генетическое программирование – один из методов эволюционной оптимизации. Алгоритм полностью автоматизирован.

4. Google Vizier. Система предлагает динамический выбор алгоритмов оптимизации, однако по умолчанию использует метод Gaussian Process Bandits, который после запуска может быть динамически заменён более масштабируемыми алгоритмами.

Таким образом, одной из тенденций в области ML являются системы автоматизированного ML, которые заменяют традиционный подход к построению систем ML, требующий знания практически всех существующих алгоритмов, умений их верно конфигурировать и применять. Системы AutoML позволяют подобрать алгоритм оптимизации гиперпараметров выбранной модели, что даёт возможность сократить затраты материальных, финансовых, трудовых и временных ресурсов.

Библиографический список:

1. Лука В.Д. Сравнение алгоритмов Machine Learning в решении задач распознавания изображений // Информатика: проблемы, методы, технологии: материалы XXIII Международной научно-практической конференции им. Э.К. Алгазинова. – Воронеж, 2023. – С. 593-602.

2. Миронов А.М. Машинное обучение. Часть 1: учебник для вузов. – М.: МАКС Пресс, 2018. – 90 с.

3. Платонов А.В. Машинное обучение: учеб. пособие для вузов / А.В. Платонов. – М.: Изд-во Юрайт, 2023. – 85 с.

4. Воронов М.В. Системы искусственного интеллекта: учебник и практикум для вузов / М.В. Воронов, В.И. Пименов, И.А. Небаев. – М.: Изд-во Юрайт, 2023. – 256 с.

5. Ланц Б. Машинное обучение на R: экспертные техники для прогностического анализа. – СПб.: Питер, 2020. – 438 с.

6. Муравьёв С.Б. Автоматическая настройка гиперпараметров алгоритмов

кластеризации с помощью обучения с подкреплением / С.Б. Муравьёв, В.А. Ефимова, В.В. Шаламов, А.А. Фильченков, И.Б. Сметанников // Научно-технический вестник информационных технологий, механики и оптики. – 2019. – Т. 19, № 3. – С. 508-515. – DOI: 10.17586/2226-1494-2019-19-3-508-515.

7. Mustafa A., Rahimi A.M. Automated Machine Learning for Healthcare and Clinical Notes Analysis. *Computers*, 2021, vol. 10, no. 24. DOI: 10.3390/computers10020024.

8. Баймуратов И.Р. Метод комплексной оценки моделей данных для автоматизации машинного обучения без учителя: дис. … канд. техн. наук: 05.13.17 / Ильдар Раисович Баймуратов; Нац. исслед. ун-т ИТМО. – СПб., 2019. – 240 с.

9. Смирнова В.С. Оптимизация гиперпараметров на основе объединения априорных и апостериорных знаний о задаче классификации / В.С. Смирнова, В.В. Шаламов, В.А. Ефимова, А.А. Фильченков // Научно-технический вестник информационных технологий, механики и оптики. – 2020. – Т. 20, № 6. – С. 828-834. – DOI: 10.17586/2226-1494-2020-20-6-828-834.

10. Попова И.А. Системы AutoML как современный инструмент для построения моделей машинного обучения // StudNet. – 2022. – Т. 5, № 1. – С. 815-825.