УДК 004.021

АЛГОРИТМЫ АНАЛИЗА ДАННЫХ В МАШИННОМ ОБУЧЕНИИ

Н. А. Глинский[1)], Б. А. Бадак[2)]

[1)]Белорусский национальный технический университет,
пр. Независимости, 65, г. Минск, Беларусь
[2)]Научный руководитель: Белорусский национальный технический университет,
пр. Независимости, 65, г. Минск, Беларусь, badak.bazhena@bk.ru

В работе представлен обзор современных методов машинного обучения и анализа данных, основанный на принципах индуктивного рассуждения; рассматриваются основные алгоритмы машинного обучения, такие как классификация, регрессия и кластеризация, а также их приложения в различных областях; описаны основные принципы работы каждого метода (k-средние, дерево решений, метод опорных векторов, алгоритмы линейной регрессии и кластерного анализа), их преимущества и недостатки, а также практические аспекты их применения.

***Ключевые слова*:** машинное обучение; индуктивное рассуждение(вывод); анализ данных; регрессия; кластеризация; отбор признаков; визуализация данных; оценка моделей.

В последние десятилетия машинное обучение стало неотъемлемой частью многих областей науки, технологий и бизнеса. **Машинное обучение** – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач [1]. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме. То есть суть машинного обучения заключается в том, чтобы запрограммировать компьютер так, чтобы он смог обучаться на доступных ему данных. Успешный обучаемый должен уметь совершать переход от отдельных примеров к более широкому обобщению. Это называется также индуктивным рассуждением, или индуктивным выводом. Однако индуктивное рассуждение может приводить к ложным заключениям.

Средства машинного обучения – это программы, которые изменяют своё поведение в зависимости от входных данных, они естественным образом приспосабливаются к изменениям окружающей среды, с которой они взаимодействуют. Примеры успешного применения машинного обучения в таких областях включают программы распознавания рукописного текста (где одна программа способна адаптироваться к почерку разных пользователей) и программы распознавания речи (где каждый человек имеет уникальный голос).

Одним из ключевых компонентов машинного обучения являются *алгоритмы анализа данных*, которые позволяют извлекать ценные знания из больших объемов информации.

Алгоритм – совокупность точно заданных правил решения некоторого класса задач или набор инструкций, описывающих порядок действий исполнителя для решения определённой задачи [4]. Эти алгоритмы обеспечивают обработку данных, выявление закономерностей и прогнозирование результатов на основе имеющихся данных. В работе рассматриваются основные алгоритмы анализа данных (линейная регрессия, кластерный анализ, деревья решений, метод опорных векторов, нейронные сети) в машинном обучении, их принципы работы и практическое применение.

Кластеризация – это процесс группировки элементов или объектов на основе их сходства по определенным характеристикам. В результате каждая группа, или кластер, содержит объекты, которые наиболее похожи между собой. Представим переезд: нужно разложить по коробкам вещи по категориям (кластерам) – например одежда, посуда, декор, канцелярия, книги. Так удобнее перевозить и раскладывать предметы в новом жилье. Процесс сбора вещей по

коробкам и будет кластеризацией. Примеры: маркетинговые исследования (сегментация аудитории на основе их предпочтений, поведения или демографических характеристик), медицинская диагностика (выявление сходств между пациентами на основе результатов медицинских тестов или симптомов), финансовая аналитика (выявление сегментов клиентов с различным уровнем риска или доходности инвестиций, а также для обнаружения аномального поведения или мошенничества), техническое обслуживание и ремонт (анализ данных о состоянии оборудования или машин с целью выявления сходных характеристик и определения оптимальных стратегий обслуживания или предотвращения отказов).

Линейная регрессия является наиболее простым и широко используемым методом анализа данных. Это модель машинного обучения, основанная на предположении, что зависимость в наблюдаемых данных можно описать простой прямой [2]. Оказывается, такой моделью можно объяснить большое количество явлений. Помимо этого, линейная регрессия хорошо изучена, имеет ряд полезных теоретических свойств, и ввиду своей простоты легко поддаётся интерпретации. На рис. 1 приведён пример линейной регрессии, где x – товарооборота X (тыс. руб.) в 20 магазинах за квартал, y – данные средней выработки на одного рабочего.

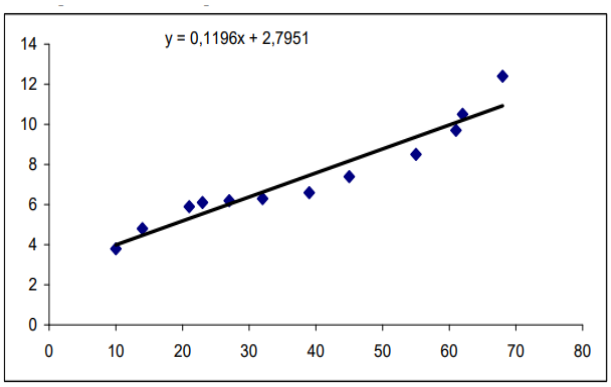

Рис. 1. Линейная-регрессия для товарооборота

Кластерный анализ используется для разделения набора данных на группы или кластеры на основе сходства между наблюдениями. Этот метод позволяет выявить внутренние структуры данных и выделить группы, обладающие схожими характеристиками, т.е. основная идея кластерного анализа заключается в том, чтобы разбить объекты на группы или кластеры таким образом, чтобы внутри группы эти наблюдения были более похожи друг на друга, чем на объекты другого кластера [3]. Как же разбить данные на кластеры? Чтобы разбить данные на кластеры достаточно измерить расстояние между точками и на основе этого измерения принять решение к какому кластеру отнести то или иное наблюдение.

Метод k-средних является одним из наиболее популярных и простых алгоритмов в задачах кластеризации в машинном обучении. Он используется для разделения набора данных на группы или кластеры на основе их признаков.

Дерево решений представляют собой графическую модель, используемую для принятия решений на основе последовательности условий. Она широко применяется в задачах классификации и регрессии. Дерево решений часто используют в банковском секторе и в тех сферах, где применяют скрипты для общения с клиентами и нужно управлять процессами принятия решений. Пример такой сферы – финансовые услуги, где банки и страховые компании проверяют информацию о клиенте в строгой последовательности, чтобы оценить риски перед заключением договора. Дерево принятия решений состоит из «узлов» и «листьев». Вверху дерева – начальный корневой узел, в который попадает вся выборка. Далее происходит проверка на выполнение условия или наличие признака. В результате такой проверки группа данных разбивается на подгруппы: подгруппа данных, которые прошли проверку, и подгруппа данных, которые не соответствуют заданному условию. Далее подгруппы данных попадают в следующий узел с новой проверкой. И так до конечного узла дерева задач, который отвечает заданной цели анализа данных или завершает процесс принятия решения (рис. 2).

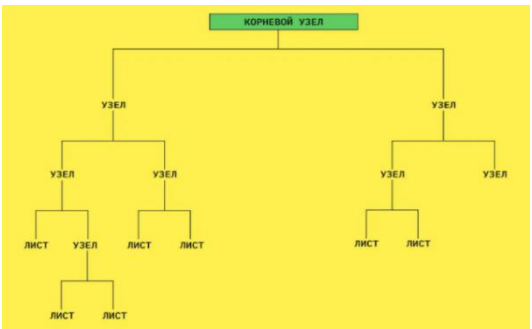

Рис. 2. Пример дерева решений

Метод опорных векторов (SVM) используется для задач классификации и регрессии. Он строит гиперплоскость в пространстве признаков, максимально разделяющую классы или аппроксимирующую зависимости в данных. Метод опорных векторов использует обучающие данные для нахождения гиперплоскости, которая разделяет два класса данных в многомерном пространстве. Метод принимает входные данные и возвращает выходные результаты на основе обученной модели. Ограничения метода зависят от вычислительной мощности и размерности входной выборки. Его задача – найти гиперплоскость, которая наилучшим образом разделяет обучающую выборку. На рис. 3 синими крестиками отмечены точки первого класса, а красными кружками точки второго класса.

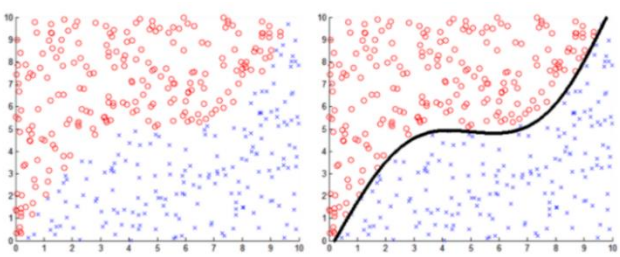

Рис. 3. Пример гиперплоскости

Нейронные сети моделируют работу человеческого мозга и используются для решения различных задач в машинном обучении, включая классификацию, регрессию, кластеризацию и обработку естественного языка.

Алгоритмы анализа данных играют ключевую роль в развитии машинного обучения и его применении в различных областях [5]. Они позволяют извлекать информацию из данных, делать предсказания и принимать решения на основе имеющихся знаний. Понимание принципов работы основных алгоритмов анализа данных в машинном обучении является важным шагом для развития компетенций в этой области и применения их на практике.

Библиографические ссылки

1. Машинное обучение [Электронный ресурс].
URL: https://ru.wikipedia.org/wiki/Машинное_обучение (дата обращения: 30.03.2024).

2. Что такое линейная регрессия [Электронный ресурс].
URL: https://sysblok.ru/glossary/chto-takoe-linejnaja-regressija (дата обращения: 30.03.2024).

3. Кластерный анализ [Электронный ресурс].
URL: https://www.dmitrymakarov.ru/intro/clustering-16/ (дата обращения: 30.03.2024).

4. Алгоритм [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Алгоритм (дата обращения: 30.03.2024).

5. Для чего начинающим аналитикам нужны деревья решений [Электронный ресурс]. URL: https://practicum.yandex.ru/blog/chto-takoe-derevo-reshenii-kak-ego-postroit (дата обращения: 30.03.2024).