– Полуконтролируемое: Полуконтролируемое обучение можно определить как гибрид вышеупомянутых контролируемых и неконтролируемых методов, поскольку оно работает как с маркированными, так и с немаркированными данными [41, 105]. Таким образом, оно находится между обучением «без контроля» и обучением «с контролем». В реальном мире маркированные данные могут быть редкими в нескольких контекстах, а немаркированные данные многочисленны, где полуконтролируемое обучение полезно [75]. Конечной целью модели полуконтролируемого обучения является обеспечение лучшего результата прогнозирования, чем тот, который был получен с использованием только размеченных данных модели.

– Неконтролируемое: Неконтролируемое обучение анализирует немаркированные наборы данных без необходимости вмешательства человека, т. е. процесс, управляемый данными [41]. Это широко используется для извлечения генеративных признаков, выявления значимых тенденций и структур, группирования результатов и исследовательских целей. Наиболее распространённые задачи неконтролируемого обучения — кластеризация, оценка плотности, изучение признаков, снижение размерности, поиск правил ассоциации, обнаружение аномалий и т. д.

– Контролируемое: Подход, основанный на задачах, при котором алгоритмы обучаются на размеченных данных для выполнения задач классификации или регрессии. Классификация текста — пример контролируемого обучения.

– Подкрепление: Обучение с подкреплением — тип алгоритма машинного обучения, в котором агент учится через взаимодействие со средой, получая вознаграждения или штрафы [52]. Применяется в робототехнике, автономном вождении, логистике.

Рис. 2 Различные типы методов машинного обучения.

Таблица 1. Различные типы методов машинного обучения с примерами.

Классификация, кластеризация — Без присмотра Под наблюдением — Классификация, регрессия Укрепление — Модели основаны на вознаграждении или наказании Полуконтролируемый — Модели строятся с использованием комбинированных данных (маркированных + немаркированных)

Machine Translated by Google

Классификация рассматривается как контролируемый метод обучения в машинном обучении, относящийся к проблеме предиктивного моделирования, где метка класса предсказывается для заданного примера [41]. Наивный Байес (NB): Наивный Байесовский алгоритм основан на теореме Байеса с предположением о независимости признаков. Он хорошо работает как для бинарных, так и для многоклассовых задач, таких как классификация документов и фильтрация спама, требуя при этом небольшого объема обучающих данных [94]. Гауссовский, многочленный, бернуллиевский и категориальный варианты — распространенные формы NB.

Линейный дискриминантный анализ (LDA): Линейный классификатор, проецирующий данные в пространство меньшей размерности, минимизируя внутриклассовую дисперсию. Предполагает гауссовское распределение признаков и одинаковую ковариацию для всех классов.

Логистическая регрессия (LR): Вероятностная модель, использующая сигмоидальную функцию для оценки вероятностей принадлежности к классам. Применяется для задач классификации, легко интерпретируется, но требует регуляризации для избежания переобучения.

Дерево решений (DT): Непараметрический метод, создающий древовидную структуру решений на основе критерия прироста информации (энтропия) или индекса Джини. Прост в интерпретации, но склонен к переобучению без ограничения глубины.

Случайный лес (RF): Ансамблевый метод, объединяющий множество деревьев решений для повышения точности и устойчивости. Снижает риск переобучения за счет усреднения результатов отдельных деревьев.

Машина опорных векторов (SVM): Создает гиперплоскость максимального отступа между классами, эффективна в многомерных пространствах, но чувствительна к шуму и требует настройки ядра.

K-ближайших соседей (KNN): Ленивый алгоритм классификации, присваивающий метку на основе большинства голосов k ближайших соседей. Простой, но вычислительно затратный на больших данных.

Алгоритмы кластеризации: – K-Means: алгоритм разбиения на k кластеров путём минимизации внутрикластерной дисперсии. – Иерархическая кластеризация: строит дендрограмму, позволяющую анализировать группы данных на разных уровнях. – DBSCAN: выявляет кластеры любой формы и автоматически отделяет шум.

Методы уменьшения размерности: – PCA (Principal Component Analysis): линейное снижение размерности, сохраняющее максимум дисперсии. – t-SNE: нелинейный метод для визуализации высокоразмерных данных.

Методы ассоциативных правил: – Apriori: поиск частых наборов элементов и правил ассоциаций. – FP-Growth: эффективный алгоритм ассоциаций без многократного сканирования данных.

Методы обнаружения аномалий: – Isolation Forest: изолирует аномальные точки через случайные разбиения. – Local Outlier Factor (LOF): оценивает аномальность объектов относительно соседей.

Методы глубокого обучения: – Нейронные сети (MLP): многослойные перцептроны для классификации и регрессии. – CNN (Convolutional Neural Networks): для обработки изображений. – RNN/LSTM/GRU: для последовательных данных (текст, временные ряды). – Autoencoder: уменьшение размерности и обнаружение аномалий. – GAN (Generative Adversarial Networks): генерация реалистичных данных.