

ВВЕДЕНИЕ

Искусственный интеллект (ИИ) и машинное обучение (МО) сегодня стали ключевыми направлениями развития информационных технологий и прочно вошли в практику обеспечения информационной безопасности (ИБ). Современные интеллектуальные системы способны анализировать огромные объемы данных, выявлять сложные, неочевидные закономерности и принимать решения в условиях неопределенности. Эта способность делает их незаменимым инструментом в условиях цифровой трансформации, когда традиционные, основанные на сигнатурах и статических правилах, подходы к защите демонстрируют растущую неэффективность перед лицом адаптивных, целевых и скрытых атак.

Основу большинства современных систем ИИ составляют нейронные сети (НС) — математические модели, способные обучаться на примерах. От простых перцептронов, решающих линейно разделимые задачи, развитие привело к созданию глубоких архитектур. Исторически прогресс был связан с появлением многослойных сетей и алгоритмов обратного распространения ошибки, что позволило решать принципиально более сложные, нелинейные задачи.

Современные подходы, основанные на глубоком обучении, используют большое количество слоев для автоматического выявления иерархических признаков данных. Ключевыми архитектурами стали сверточные сети (CNN) для обработки изображений и видео, рекуррентные сети (RNN, LSTM, GRU) для анализа последовательностей и временных рядов, а также трансформеры, учитывающие сложные взаимосвязи внутри данных с помощью механизма внимания. Эти технологии формируют мощный и гибкий инструментарий для решения широкого круга практических задач.

В области кибербезопасности применение ИИ стало повсеместным, затрагивая все уровни защиты. Методы машинного

обучения лежат в основе высокоточных систем биометрической аутентификации личности. Современные системы в этой области используют комбинированные подходы: извлечение признаков с помощью CNN, построение эмбедингов для сравнения объектов и обучение на больших датасетах с учетом возможных вариаций и помех, что минимизирует риск подделки. Алгоритмы машинного обучения являются ядром интеллектуальных систем мониторинга (SIEM/SOAR) для обнаружения аномалий в сетевом трафике, анализа вредоносного ПО и корреляции инцидентов.

Еще одним важным направлением является прогнозирование и предотвращение угроз, связанных с социальной инженерией и человеческим фактором. Системы ИИ могут анализировать поведенческие паттерны пользователей, выявляя аномалии, которые могут указывать на внешнее давление, стресс или попытку несанкционированных действий. Современные генеративные модели, такие как GAN, создают новые угрозы в виде сверхубедительного фишингового контента и дипфейк-атак, но в то же время служат инструментами для тестирования устойчивости защитных систем.

Важно подчеркнуть, что стремительная интеграция ИИ в критические инфраструктуры порождает и новые классы уязвимостей. Системы, основанные на машинном обучении, сами становятся мишенью для изощренных атак: отравления обучающих данных (data poisoning), генерации состязательных примеров (adversarial examples) или кражи модели (model stealing).

Для успешного применения ИИ в ИБ важно использовать надежные методы обучения и регуляризации моделей. Функции потерь и алгоритмы оптимизации позволяют сети корректно обрабатывать зашумленные и неполные данные, минимизировать ошибки и учитывать статистическую значимость признаков. Таким образом, современная кибербезопасность приобретает двойственный характер, требуя от специалиста не только навыков защиты с помощью ИИ, но и глубокого понимания принципов работы этих технологий для защиты самих интеллектуальных систем.

В связи с этим подготовка кадров, обладающих комплексными знаниями на стыке ИИ и ИБ, является одной из наиболее актуальных задач. Настоящий учебник призван системно и всесторонне ответить на этот вызов. Его структура отражает логику построения компетенций: от фундаментальных основ архитектуры и обучения нейронных сетей — к их конкретному применению

в биометрической аутентификации, противодействии социальной инженерии, построении систем обнаружения угроз и, наконец, к вопросам безопасности самих алгоритмов ИИ.

Теоретический материал каждой главы подкреплён практическими заданиями, которые содержат ключевые фрагменты кода. Для успешного выполнения заданий необходимо владение основами программирования на языке Python. Для их выполнения используется облачная платформа Google Colab (Colaboratory), которая предоставляет интерактивную среду на основе Jupyter Notebooks, что позволяет абстрагироваться от трудоёмкой настройки локального программного окружения и сосредоточиться на освоении методов ИИ и их приложений в области безопасности. Платформа предлагает бесплатный доступ к вычислительным ресурсам, включая графические (GPU) и тензорные (TPU) процессоры, а также включает предустановленные библиотеки, такие как TensorFlow, PyTorch, Keras, Scikit-learn и OpenCV, что создаёт готовую среду для экспериментов с нейронными сетями, обработки данных и реализации алгоритмов.

В рамках практикума на базе Google Colab рассматриваются ключевые задачи безопасности: построение моделей для обнаружения аномалий в сетевом трафике, создание систем биометрической аутентификации, анализ поведения программного обеспечения на предмет вредоносности, генерация синтетических данных для тестирования защитных механизмов, а также эксперименты с состязательными атаками и методами защиты моделей машинного обучения. Платформа также облегчает визуализацию процессов обучения, внутренних представлений данных и динамики метрик, что способствует глубокому пониманию работы алгоритмов.

Учебник адресован студентам и аспирантам, обучающимся по направлениям информационной безопасности и искусственного интеллекта, а также практикующим специалистам, стремящимся систематизировать и углубить свои знания в этой быстро развивающейся и критически важной сфере.

Авторы выражают искреннюю признательность рецензентам за внимательное прочтение рукописи, ценные замечания и конструктивные предложения, которые способствовали значительному улучшению содержания и структуры учебника.

Особую благодарность за глубокий анализ, профессиональные рекомендации и поддержку в подготовке издания авторы выражают канд. техн. наук Юрию Николаевичу Чернышову.

Р а з д е л I

МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Идея искусственного интеллекта появилась в 1950-х годах, когда группа энтузиастов из только что зарождающейся области информатики задались вопросом, можно ли заставить компьютеры «думать». Коротко эту область можно определить так: автоматизация интеллектуальных задач, обычно выполняемых людьми. Соответственно, *искусственный интеллект* (ИИ) как научная область включает не только методы машинного и глубокого обучения, но и подходы, не связанные с обучением на данных, — символические вычисления, экспертные системы, эвристический поиск.

Другими словами, *искусственный интеллект* — это научная дисциплина и область компьютерной науки, занимающаяся созданием программных и аппаратных систем, способных к решению интеллектуальных задач, традиционно считающихся прерогативой человеческого мышления. К таким задачам относятся: логический вывод, принятие решений в условиях неопределенности, планирование, обучение, восприятие (компьютерное зрение, обработка естественного языка) и целенаправленное поведение.

Ключевой признак систем ИИ — наличие у них *интеллектуального поведения*, т. е. способности адаптироваться к изменяющимся условиям и решать новые задачи, т. е. способности адаптироваться к изменяющимся условиям и решать задачи, не сводимые к заранее заданному алгоритму.

Долгое время многие эксперты полагали, что ИИ уровня человека можно создать, если имеется достаточный набор явных правил для манипулирования знаниями. Этот подход, известный как символический ИИ, и является доминирующей парадигмой ИИ с 1950-х до конца 1980-х годов. Пик его популярности пришелся на бум экспертных систем в 1980-х годах.

Символический ИИ прекрасно справлялся с решением четко определенных логических задач, таких как игра в шахматы,

но, как оказалось, невозможно задать строгие правила для решения более сложных, нечетких задач, таких как классификация изображений, распознавание речи и перевод на другие языки. На смену символическому ИИ пришли новые подходы — *машинное обучение* и *глубокое обучение*. Эти направления являются основополагающими в современных технологических решениях. На рис. 1.1 представлена диаграмма Венна, иллюстрирующая, что глубокое обучение является частным случаем репрезентативного обучения, которое, в свою очередь, относится к классу методов машинного обучения и используется во многих подходах искусственного интеллекта.



Рис. 1.1. Диаграмма Венна для ИИ

В классическом программировании, в парадигме символического ИИ, программист вводит правила и данные для обработки, в соответствии с этими правилами получает ответ. В машинном обучении программист вводит данные и ответы, соответствующие этим данным, а на выходе получает правила. Эти правила затем можно применить к новым данным для получения оригинальных ответов. Парадигма программирования представлена на рис. 1.2.

Машинное обучение — это подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться на основе эмпирических данных. В отличие от классического программирования, где поведение системы определяется явно заданными правилами, в машинном обучении модель

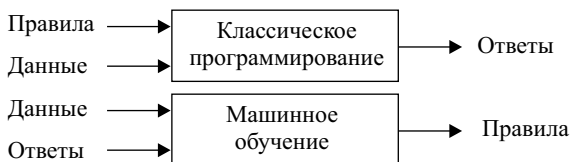


Рис. 1.2. Парадигма программирования

выявляет статистические закономерности и скрытые паттерны в предоставленном наборе данных (тренировочной выборке) и на их основе строит прогностическую модель или принимает решения.

В машинном обучении система обучается, а не программируется явно. Ей передаются многочисленные примеры, имеющие отношение к решаемой задаче, а она находит в этих примерах статистическую структуру, которая позволяет системе выработать правила для автоматического решения задачи.

Машинное обучение тесно связано с математической статистикой, но имеет несколько важных отличий. В отличие от классической статистики, машинное обучение чаще ориентировано на работу с большими и сложными наборами данных, где традиционные статистические методы (например, байесовские подходы в их классической формулировке) не всегда применимы напрямую. Как следствие, машинное и особенно глубокое обучение в значительной мере опираются на эмпирические и инженерные подходы, а их теоретическое обоснование зачастую развивается вслед за практическими достижениями.

Для машинного обучения нужны три составляющие: контрольные входные данные, примеры ожидаемых результатов и способ оценки качества работы алгоритма. Модель машинного обучения трансформирует исходные данные в значимые результаты, обучаясь на известных примерах входных данных и результатов. Формально, задача машинного обучения сводится к минимизации эмпирического риска или максимизации целевой функции на множестве входных данных. Для решения более сложных и специфичных задач применяется глубокое машинное обучение.

1 Классические нейронные сети

Нейронные сети — это одно из ключевых направлений в области искусственного интеллекта, вдохновленное принципами работы биологической нервной системы. Основная идея заключается в том, чтобы воспроизвести такие ключевые способности мозга, как обучение на опыте и адаптация, что позволяет создавать вычислительные системы, приближенно имитирующие когнитивные функции.

Искусственная нейронная сеть (ИНС) — это математическая модель и вычислительная архитектура, используемая в машинном обучении (в частном случае — в глубоком машинном обучении). Она представляет собой систему из большого числа взаимосвязанных простых процессоров (искусственных нейронов), организованных в слои. Каждый нейрон преобразует входной вектор сигналов в один выходной сигнал с помощью функции активации, зависящей от взвешенной суммы входов.

С математической точки зрения ИНС реализует сложную функцию множества переменных. Конкретный вид этой функции определяется архитектурой сети (например, количеством и типом слоев, связями между нейронами) и значениями параметров (весов синаптических связей). Процесс обучения сети по сути является оптимизацией этих весов на основе предоставленных тренировочных данных.

Исторически в 60–80-е годы XX века фокус исследований ИИ сместился в сторону экспертных систем. Несмотря на их эффективность в узких предметных областях, создание более универсальных и гибких интеллектуальных систем требовало иного подхода. Это стало катализатором для роста интереса к биологически инспирированным моделям, в частности, к архитектурам, повторяющим структуру нейронных сетей человеческого мозга.

Таким образом, ИНС созданы по биологическому прототипу. Они состоят из элементарных единиц — искусственных нейронов, — функционал которых аналогичен базовым функциям

биологического нейрона. Эти элементы организуются в слоистые структуры, отдаленно напоминающие анатомию мозга. Несмотря на это упрощенное сходство, ИНС демонстрируют ряд впечатляющих свойств, характерных для живого интеллекта: способность к обучению, обобщению информации и абстрагированию ключевых признаков. ИНС индуцированы биологией, так как они состоят из элементов, функциональные возможности которых аналогичны большинству элементарных функций биологического нейрона. Эти элементы затем организуются по способу, который соответствует анатомии мозга. Даже при таком поверхностном сходстве искусственные нейронные сети демонстрируют удивительное число свойств, присущих мозгу. Например, они обучаются на основе опыта, обобщают предыдущие прецеденты на новые случаи и извлекают существенные свойства из поступающей информации, содержащей излишние данные.

Рассмотрим некоторые свойства нейронных сетей.

Обучение. ИНС могут адаптировать свое поведение в ответ на изменения внешней среды. Именно это свойство вызывает основной интерес к ним. Получая входные данные (иногда вместе с ожидаемым выходом), сеть самостоятельно настраивает внутренние параметры для достижения целевой реакции. Существует множество алгоритмов обучения, каждый со своими преимуществами и ограничениями, при этом фундаментальные вопросы о пределах обучаемости сетей и оптимальных методах обучения остаются предметом исследований по сей день.

Обобщение. После обучения реакция сети становится устойчивой к незначительным искажениям во входных данных. Эта врожденная способность «видеть суть сквозь шум» критически важна для задач распознавания образов в реальных условиях. Она снимает требование абсолютной точности входных данных, характерное для классических алгоритмов, и позволяет создавать системы, работающие в условиях неидеальной информации.

Абстрагирование. Некоторые типы ИНС способны выявлять инвариантные, основополагающие черты входных сигналов. К примеру, обучившись на множестве искаженных изображений буквы «А», сеть может сгенерировать ее эталонный, идеальный образ. Фактически, она учится воспроизводить то, с чем никогда не сталкивалась в чистом виде, — качество, высоко ценимое в человеческом интеллекте.

Несмотря на большое разнообразие вариантов нейронных сетей, все они имеют общие черты. Так, все они, как и мозг чело-

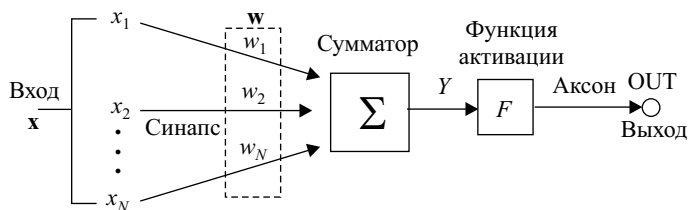


Рис. 1.1. Модель нейрона

века, состоят из большого числа однотипных элементов — нейронов, которые имитируют нейроны головного мозга, связанные между собой.

Нейрон — это элементарная процессорная единица в сети. Его упрощенная модель, лежащая в основе большинства ИНС, включает три ключевых компонента (рис. 1.1):

- *синапсы* — каждый синапс характеризуется своим весом (силой связи). Он выполняет роль канала связи между нейронами, умножая входной сигнал на соответствующий весовой коэффициент;
- *сумматор* — этот компонент, аналогичный телу биологической нервной клетки, агрегирует все входящие сигналы (от других нейронов или внешних источников), определяя общий уровень возбуждения нейрона;
- *функция активации* — нелинейная функция, которая преобразует суммарный входной сигнал в окончательный выходной сигнал нейрона. Именно этот сигнал передается далее на синапсы следующих нейронов.

Архитектура нейронных сетей варьируется в зависимости от количества слоев, типов связей между нейронами, методов обучения и вида используемых функций активации.

1.1. Персептрон

Многослойным персептроном называют нейронную сеть прямого распространения. Входной сигнал в такой сети распространяется в прямом направлении, от слоя к слою. Многослойный персептрон в общем представлении состоит из следующих элементов:

- множества входных узлов, которые образуют входной слой;
- одного или нескольких скрытых слоев вычислительных нейронов;
- одного выходного слоя нейронов.

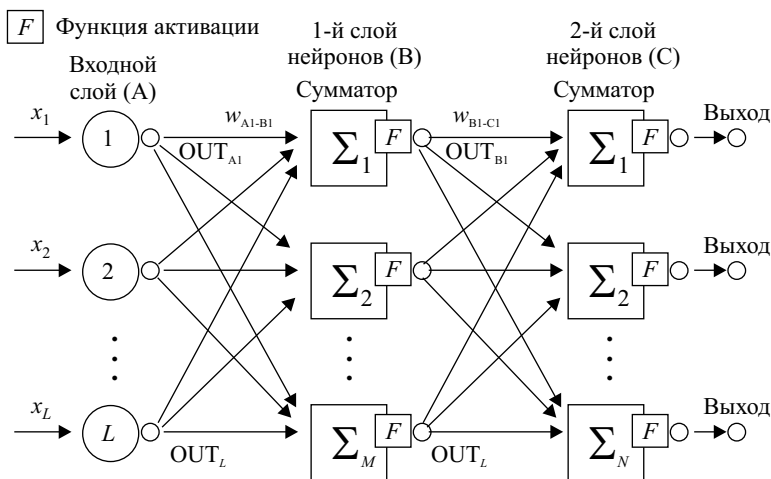


Рис. 1.2. Многослойный персептрон

Многослойный персептрон представляет собой обобщение однослойного персептрона Розенблатта. На рис. 1.2 представлен пример модели нейронной сети типа многослойный персептрон.

Многослойные персептроны успешно применяются для решения разнообразных сложных задач и имеют три отличительных признака.

Каждый нейрон сети имеет нелинейную функцию активации. Важно подчеркнуть, что такая нелинейная функция должна быть гладкой (т. е. всюду дифференцируемой), в отличие от жесткой пороговой функции, используемой в персептроне Розенблатта. Наиболее часто используемой формой функции, удовлетворяющей этому требованию, является сигмоидальная. Примером сигмоидальной функции может служить логистическая функция, задаваемая выражением

$$F = \frac{1}{1 + e^{-\alpha Y}},$$

где α — параметр наклона сигмоидальной функции; Y — результат работы нейрона. Изменяя параметр α , можно построить функции с различной крутизной. График этой функции представлен на рис. 1.3.

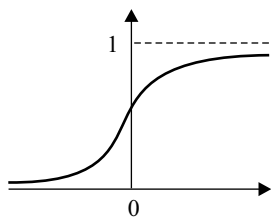


Рис. 1.3. Сигмоидальная функция

Наличие нелинейности играет очень важную роль, так как в противном случае отображение «вход-выход» сети можно свести к обычному однослойному персептрону.

Многослойный персептрон содержит один или несколько слоев скрытых нейронов, не являющихся частью входа или выхода сети. Эти нейроны позволяют сети обучаться решению сложных задач, последовательно извлекая наиболее важные признаки из входного образа.

Многослойный персептрон обладает высокой степенью связности, реализуемой посредством синаптических соединений. Изменение уровня связности сети требует изменения множества синаптических соединений или их весовых коэффициентов. Комбинация всех этих свойств наряду со способностью к обучению на собственном опыте обеспечивает вычислительную мощь многослойного персептрона.

Самым важным свойством нейронных сетей является их способность обучаться на основе данных окружающей среды и со временем повышать качество выполняемых ими задач. Процесс обучения представляет собой итеративную процедуру адаптации параметров модели — синаптических весов и порогов (смещений). В результате повторяющихся обновлений веса постепенно принимают такие значения, которые обеспечивают оптимальное поведение сети на заданной задаче.

Обучение нейронной сети — это процесс настройки ее свободных параметров на основе данных, отражающих свойства среды, в которой она применяется. Конкретный тип обучения определяется тем, каким образом происходит обновление этих параметров и какая информация используется для их подстройки.

Существуют два концептуальных подхода к обучению нейронных сетей: обучение с учителем и обучение без учителя.

Обучение нейронной сети с учителем предполагает, что для каждого входного вектора из обучающего множества существует требуемое значение выходного вектора, называемого целевым. Эти векторы образуют обучающую пару. Веса сети изменяют до тех пор, пока для каждого входного вектора не будет получен приемлемый уровень отклонения выходного вектора от целевого.

Обучение нейронной сети без учителя является намного более правдоподобной моделью обучения с точки зрения биологических корней искусственных нейронных сетей. Обучающее множество состоит лишь из входных векторов. Алгоритм обучения нейронной сети подстраивает веса сети так, чтобы получались согласованные выходные векторы, т. е. чтобы предъявление достаточно близких входных векторов давало одинаковые выходы.