ЛЕКЦИЯ 5

ОСНОВНЫЕ МЕТОДЫ И ПОДХОДЫ МАШИННОГО ОБУЧЕНИЯ

Основные понятия • определения • схемы • таблицы • графики • примеры

Преподаватель: проф.Ахмедиярова А.Т.

Структура и основные разделы

01 Понятие машинного обучения

02 Основные подходы машинного обучения

03 Обучение с учителем

04 Линейная регрессия

05 Метод k-ближайших соседей (kNN)

06 Деревья решений

07 Логистическая регрессия

08 Обучение без учителя

09 Обучение с подкреплением

10 Обучающая, валидационная и тестовая выборки

11 Переобучение и недообучение

12 Оценка качества моделей

Понятие машинного обучения

Машинное обучение (Machine Learning, ML) — направление искусственного интеллекта, в котором модель формируется на основе данных и используется для прогнозирования, классификации или выявления структуры новых данных. В отличие от программы, где все правила поведения явно задаются разработчиком, в машинном обучении значительная часть закономерностей извлекается из обучающей выборки.

Ключевая идея: Машинное обучение (Machine Learning, ML) — направление искусственного интеллекта, в котором модель формируется на основе данн

Понятие машинного обучения — продолжение

Важно различать алгоритм и модель. Алгоритм задаёт процедуру обучения, а модель является результатом применения этой процедуры к конкретным данным. Например, алгоритм дерева решений определяет, как выбирать признаки и строить разбиения, а конкретное обученное дерево с определёнными узлами и порогами — это модель. Пример.

Понятие машинного обучения — продолжение

В системе обнаружения подозрительных входов объектом является отдельная попытка входа. Признаками могут быть число неудачных попыток, время суток, страна подключения, тип устройства и длительность сессии. Целевой переменной может быть класс normal или suspicious. Рисунок 1 – Типовой цикл построения модели машинного обучения Таблица 1. Основные компоненты задачи машинного обучения

Схема: Понятие машинного обучения

Машинное обучение (Machine Learning, ML) — напра

Важно различать алгоритм и модель. Алгоритм зада

Пример. В системе обнаружения подозрительных вхо

Рисунок 1 – Типовой цикл построения модели машин

Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Основные подходы машинного обучения

В учебном курсе выделяются три базовых подхода: обучение с учителем, обучение без учителя и обучение с подкреплением. Они различаются тем, какая информация доступна алгоритму в процессе обучения и какой результат требуется получить. Рисунок 2 – Основные подходы машинного обучения Таблица 2 – Основные подходы

Ключевая идея: В учебном курсе выделяются три базовых подхода: обучение с учителем, обучение без учителя и обучение с подкреплением. Они раз

Сравнительная таблица: Основные подходы машинного обучения

Основные подходы машинного обучения	В учебном курсе выделяются три базовых подхода: обучение с учителем, обучение без учителя и обучение с подкрепление
Обучение с учителем	Рисунок 2 – Основные подходы машинного обучения
Линейная регрессия	Таблица 2 – Основные подходы

Обучение с учителем

При обучении с учителем каждый обучающий пример содержит входные признаки X и известный правильный ответ y. Алгоритм подбирает зависимость между X и y, а затем применяется к новым объектам. Две основные задачи этого подхода — классификация и регрессия. 1. Классификация Классификация используется, когда целевая переменная принимает дискретные значения.

Ключевая идея: При обучении с учителем каждый обучающий пример содержит входные признаки X и известный правильный ответ y. Алгоритм подбирае

Обучение с учителем — продолжение

При двух классах задача называется бинарной, при большем числе классов — многоклассовой. Пример. Если модель оценивает вероятность атаки как 0,92, система может отнести событие к классу attack. Однако сам порог принятия решения зависит от прикладной задачи и цены ошибок. 2. Регрессия Регрессия применяется, когда необходимо прогнозировать числовое значение.

Обучение с учителем — продолжение

В загруженном учебном пособии линейная регрессия рассматривается как базовая параметрическая модель, а подбор весов формулируется как задача минимизации функции потерь. Для линейной регрессии в качестве функции потерь приводится среднеквадратическая ошибка MSE. Иллюстративный пример. Требуется прогнозировать время обработки запроса сервера.

Обучение с учителем — продолжение

Признаки: число активных пользователей, объём передаваемых данных и загрузка процессора. Целевая переменная — время ответа в миллисекундах.

Линейная регрессия

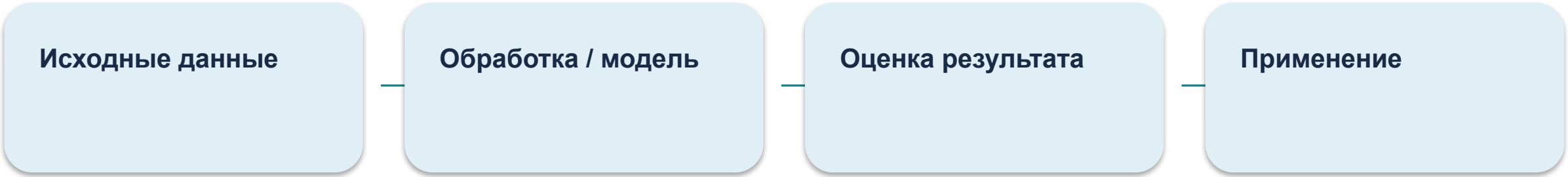
Линейная регрессия описывает зависимость числовой целевой переменной от одного или нескольких признаков линейной функцией. В простейшем случае модель можно записать как $y = b_0 + b_1x$, где b_0 — свободный член, b_1 — вес признака. Для нескольких признаков используется вектор весов.

Ключевая идея: Линейная регрессия описывает зависимость числовой целевой переменной от одного или нескольких признаков линейной функцией. В

Линейная регрессия — продолжение

Обучение заключается в выборе таких коэффициентов, при которых расхождение между фактическими и прогнозируемыми значениями минимально. В учебном пособии подчёркивается чувствительность линейной регрессии к мультиколлинеарности — сильной зависимости между объясняющими признаками.

Схема: Линейная регрессия



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Метод k-ближайших соседей (kNN)

Метод k-ближайших соседей используется и для классификации, и для регрессии. Для нового объекта находятся k наиболее близких объектов обучающей выборки. В классификации решение обычно принимается голосованием соседей, а в регрессии — по среднему или медианному значению их целевой переменной. Пример. Новое сетевое соединение описывается признаками duration, packets и bytes.

Ключевая идея: Метод k-ближайших соседей используется и для классификации, и для регрессии. Для нового объекта находятся k наиболее близких

Метод k-ближайших соседей (kNN) — продолжение

Если среди пяти ближайших по этим признакам соединений четыре относятся к классу attack, алгоритм может классифицировать новый объект как attack. Выбор k существенно влияет на результат. Слишком малое k повышает чувствительность к шуму, а слишком большое может сглаживать границы между классами. В пособии рекомендуется определять подходящее значение с помощью кросс-валидации.

Метод k-ближайших соседей (kNN) — продолжение

Практически важно масштабировать признаки: если bytes измеряется тысячами, а duration — единицами, немасштабированный признак bytes может непропорционально влиять на расстояние.

Деревья решений

Дерево решений представляет правила в иерархической структуре. Внутренние узлы содержат проверки признаков, ветви соответствуют результатам проверок, а листья содержат итоговое решение. В пособии построение дерева описывается как последовательное разбиение исходного множества примеров сверху вниз. Для выбора признака разбиения используются критерии, оценивающие качество разделения.

Ключевая идея: Дерево решений представляет правила в иерархической структуре. Внутренние узлы содержат проверки признаков, ветви соответству

Деревья решений — продолжение

В описании алгоритма C4.5 рассматриваются энтропия и нормированный прирост информации. Чтобы уменьшить излишнюю сложность дерева, применяется отсечение ветвей (pruning). Рисунок 3 – Иллюстративный пример структуры дерева решений На рисунке показан учебный пример, а не готовое правило реальной системы безопасности.

Деревья решений — продолжение

Реальные пороги должны определяться на обучающих данных и проверяться на независимой выборке.

Сравнительная таблица: Деревья решений

Деревья решений	Дерево решений представляет правила в иерархической структуре. Внутренние узлы содержат проверки признаков, ветви с
Логистическая регрессия	Для выбора признака разбиения используются критерии, оценивающие качество разделения. В описании алгоритма C4.5 рас
Обучение без учителя	Рисунок 3 – Иллюстративный пример структуры дерева решений
Обучение с подкреплением	На рисунке показан учебный пример, а не готовое правило реальной системы безопасности. Реальные пороги должны опред

Логистическая регрессия

Логистическая регрессия применяется, когда исследуемая переменная является качественной и может быть представлена индикатором 0 или 1. В учебном пособии показано, что для такой задачи линейный отклик неудобен, поскольку прогноз вероятности должен оставаться в интервале от 0 до 1. Поэтому используется сигмоидальная логистическая функция.

Ключевая идея: Логистическая регрессия применяется, когда исследуемая переменная является качественной и может быть представлена индикатором

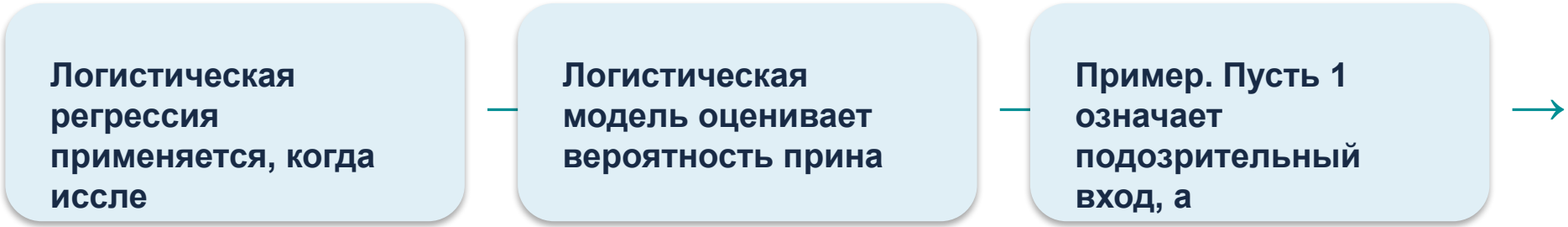
Логистическая регрессия — продолжение

Логистическая модель оценивает вероятность принадлежности к классу. Связь между линейным предиктором и вероятностью выражается через logit-преобразование — логарифм отношения шансов. Пример. Пусть 1 означает подозрительный вход, а 0 — нормальный. Модель по признакам failed_logins, unusual_time и new_device оценивает вероятность класса 1.

Логистическая регрессия — продолжение

Значение 0,86 означает высокую оценённую вероятность подозрительного события, но окончательное действие системы зависит от выбранного порога и политики безопасности.

Схема: Логистическая регрессия



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Обучение без учителя

При обучении без учителя целевые метки отсутствуют. Алгоритм анализирует структуру признакового пространства и пытается обнаружить группы, сходство, скрытые закономерности или компактное представление данных. 1. Кластеризация Кластеризация объединяет похожие объекты в группы.

Ключевая идея: При обучении без учителя целевые метки отсутствуют. Алгоритм анализирует структуру признакового пространства и пытается обнар

Обучение без учителя — продолжение

В информационной безопасности это может использоваться для исследования профилей сетевой активности: обычные рабочие станции, серверы, устройства с необычным поведением могут формировать различные группы. Кластер сам по себе не означает «атака» или «норма» — его смысл должен интерпретироваться после анализа. 2.

Обучение без учителя — продолжение

Снижение размерности При большом количестве признаков данные становятся сложнее анализировать и визуализировать. Методы снижения размерности формируют более компактное представление, сохраняя существенную структуру данных. Это может быть полезно перед визуальным анализом, кластеризацией или построением последующих моделей.

Обучение с подкреплением

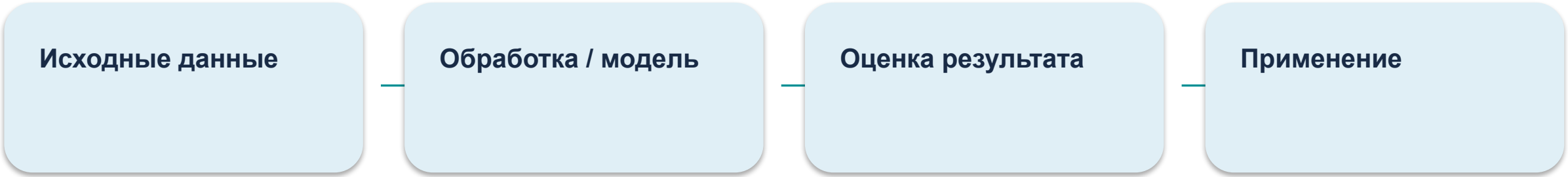
В обучении с подкреплением агент взаимодействует со средой: наблюдает состояние, выбирает действие и получает награду. Цель обучения — сформировать стратегию поведения, увеличивающую ожидаемую суммарную награду. Этот подход принципиально отличается от обычной классификации: здесь важно не только отдельное предсказание, но и последовательность действий и их отложенные последствия.

Ключевая идея: В обучении с подкреплением агент взаимодействует со средой: наблюдает состояние, выбирает действие и получает награду. Цель о

Обучающая, валидационная и тестовая выборки

Качество модели нельзя оценивать только на тех данных, на которых она обучалась. Для объективной проверки данные разделяют. Обучающая часть используется для подбора параметров модели; валидационная — для выбора гиперпараметров и сравнения вариантов; тестовая — для итоговой независимой оценки.

Схема: Обучающая, валидационная и тестовая выборки



Последовательность отражает основные этапы рассматриваемого процесса; конкретная реализация зависит от задачи.

Сравнительная таблица: Обучающая, валидационная и тестовая выборки

Обучающая, валидационная и тестовая выборки

Качество модели нельзя оценивать только на тех данных, на которых она обучалась. Для объективной проверки данные ра

Переобучение и недообучение

Переобучение возникает, когда модель слишком хорошо приспосабливается к обучающим данным, включая их случайные особенности, но хуже работает на новых объектах. Недообучение возникает, когда модель слишком проста или обучена недостаточно и не описывает даже основные закономерности.

Оценка качества моделей

Метрика должна соответствовать задаче. Для классификации одной accuracy может быть недостаточно, особенно если классы несбалансированы. В задачах безопасности редкий класс атаки может быть значительно важнее многочисленного нормального класса. Пример. В наборе 10 000 событий только 100 атак. Модель, которая всегда отвечает normal, получит accuracy 99%, но не обнаружит ни одной атаки.

Ключевая идея: Метрика должна соответствовать задаче. Для классификации одной accuracy может быть недостаточно, особенно если классы несбала

Оценка качества моделей — продолжение

Поэтому оценка должна учитывать структуру данных и последствия ошибок.

Как выбирать подход

Выбор начинается не с названия алгоритма, а с постановки задачи. Если известны правильные ответы и требуется категория — это классификация. Если известны ответы и требуется число — регрессия. Если меток нет и требуется обнаружить группы — кластеризация. Если задача состоит в выборе последовательности действий на основе награды — обучение с подкреплением.

Ключевая идея: Выбор начинается не с названия алгоритма, а с постановки задачи. Если известны правильные ответы и требуется категория — это

Как выбирать подход — продолжение

После определения типа задачи сравнивают несколько подходящих моделей. При этом учитывают качество, интерпретируемость, объём данных, масштаб признаков, вычислительную стоимость и требования к времени получения прогноза. Развёрнутый пример: обнаружение подозрительного сетевого события Предположим, имеется набор исторических сетевых событий.

Как выбирать подход — продолжение

Для каждого события известны duration, packets, bytes, failed_logins и метка normal/attack. Это задача обучения с учителем и бинарной классификации. Проверить качество и состав данных, исключить некорректные записи и определить признаки. Разделить данные на обучающую и тестовую части; при настройке модели использовать отдельную валидацию или кросс-валидацию.

Как выбирать подход — продолжение

Выбрать базовую модель, например логистическую регрессию, а затем сравнить её с kNN и деревом решений. Для kNN масштабировать числовые признаки и подобрать k. Для дерева решений подобрать ограничения сложности и проверить риск переобучения. Сравнить модели не только по accuracy, но и по precision, recall и F1-score.

Как выбирать подход — продолжение

Проанализировать ошибки: какие атаки пропускаются и какие нормальные события ошибочно вызывают тревогу. Такой пример показывает, что выбор алгоритма — только одна часть ML-задачи. Корректность данных, схема проверки и подходящая метрика не менее важны. 17. Практический пример на Python Ниже приведён компактный учебный пример классификации.

Сравнительная таблица: Как выбирать подход

Как выбирать подход

Выбор начинается не с названия алгоритма, а с постановки задачи. Если известны правильные ответы и требуется катего

График: иллюстрация зависимости



Иллюстративные значения (не результаты эксперимента).

Итоги и вопросы для обсуждения

1. Объясните понятие «Понятие машинного обучения» и приведите пример.
2. Объясните понятие «Основные подходы машинного обучения» и приведите пример.
3. Объясните понятие «Обучение с учителем» и приведите пример.
4. Объясните понятие «Линейная регрессия» и приведите пример.
5. Объясните понятие «Метод k-ближайших соседей (kNN)» и приведите пример.
6. Объясните понятие «Деревья решений» и приведите пример.
7. Объясните понятие «Логистическая регрессия» и приведите пример.