Машинное обучение и где оно обитает

Задачи, метрики качества, данные, модели и обучение

Автор презентации: Павел Дылевский
Доработал и обновил: Алексей Чернуха

По статье 1.3 «Машинное обучение» из учебника по ML (Яндекс)

Введение

Зачем нужно машинное обучение?

Вопрос на подумать. Алан Тьюринг расшифровывал Энигму перебором вариантов — машина справилась быстрее людей.

Есть задачи, которые человеку даются легко, а программой их почти не записать

Назовите примеры таких задач.

Ответ: легко человеку, трудно запрограммировать

1	Перевести текст с языка на язык
2	Поставить диагноз по симптомам
3	Выбрать, какой из двух документов лучше подходит запросу
4	Сказать, что изображено на картинке
5	Оценить, по какой цене продастся квартира

Что у них общего

1. Решение — функция из объекта в предсказание.
2. Единственно верного ответа нет: переводы бывают разными, врачи ошибаются.
3. Правильных ответов много, а неправильные легко придумать.

Вопрос.

Объясните своими словами, как человеку, далёкому от IT.

Определение машинного обучения

Компьютер сам находит закономерности в данных. Вместо того чтобы человек вручную прописывал правила, алгоритм выучивает зависимость между объектами и ответами по примерам и применяет её к новым объектам.

Данные

Примеры: объекты и правильные ответы к ним.

Модель

Функция $f(x) \to y$: по объекту x выдаёт предсказание y.

Обучение

Подбор параметров модели так, чтобы ответы на данных были как можно лучше.

Из чего состоит задача



Объекты и таргеты

Объект — картинка, история болезни, пользователь сервиса. Таргет — ответ: подпись, диагноз, уход пользователя.

Модель

Функция $f(x) \to y$. Принимает признаки объекта x и выдаёт предсказание y.

Датасет

Обучающая выборка: объекты вместе с известными ответами. По ней учится модель.

Постановка задачи

Три парадигмы обучения

С учителем

Для каждого объекта известен правильный ответ. Учимся предсказывать ответ для новых объектов.

Пример: диагноз по симптомам.

Без учителя

Есть только данные, ответов нет. Ищем структуру: например, группы похожих объектов.

Пример: группировка новостей по сюжетам.

С подкреплением

Агент действует в среде и получает награду за удачные действия.

Пример: робот учится запрыгивать на стол.

Что предсказывает модель?

Вопрос на подумать.

1. Время поездки на каршеринге
2. Кликнет ли пользователь по рекламе
3. Тема научной статьи: математика, биология, психология...
4. Теги ресторана: «кофе», «веранда», «живая музыка»
5. Порядок документов в поисковой выдаче

Чем различаются ответы? Какие группы задач вы видите?

Обучение с учителем: виды задач

Задача	Что предсказываем
Регрессия	Таргет — число или вектор. Время поездки на каршеринге, спрос на товар.
Бинарная классификация	Да или нет. Кликнет ли по рекламе, вернёт ли кредит.
Многоклассовая	Один класс из нескольких. Тема научной статьи.
Многометочная	Несколько классов сразу, они могут пересекаться. Теги ресторана.
Ранжирование	Упорядочить объекты. Поисковая выдача по релевантности.
Сложные ответы	Сегментация изображения, перевод текста, генерация новых объектов.

Определите тип задачи

Вопрос на подумать.

1. Предсказать курс евро к доллару на следующий день
2. Стилизовать текст: «Пиппина и Мерри похитили!» → «Граждане Тук и Брендибак были похищены неустановленными лицами»
3. Определить, котик или собачка на картинке
4. Научить робокота запрыгивать на стол из любой позы
5. Найти наборы товаров, которые покупают вместе

Ответы: тип задачи

Задача	Тип
1. Курс евро	Регрессия: предсказываем число
2. Стилизация текста	Генерация новых объектов на основе существующих
3. Котик или собачка	Бинарная классификация: два класса
4. Робокот	Обучение с подкреплением: перебирать траектории и узнавать ответ слишком долго и дорого
5. Наборы товаров	Обучение без учителя: ответов нет

Критерии качества

Как понять, что модель хороша

Бизнес	Сколько денег и довольных клиентов получает компания. Зависит от всего сразу, измерить трудно.
Онлайн	Что видно на работающем сервисе: например, сколько времени люди сидят в приложении.
Асессоры	Живые люди смотрят на ответы модели и ставят оценки. Нужно, когда качество — дело вкуса.
Офлайн	Считаем по старым данным до запуска: насколько ответы модели отличаются от правильных.

Офлайн и онлайн метрики

	Офлайн	Онлайн
Когда	До запуска модели	После деплоя, на работающей системе
На чём	Исторические данные, тестовая выборка, оценки асессоров	Реальное поведение пользователей и магазинов
Примеры	accuracy, MAE, MSE, precision, recall	Медианная длина сессии, число бананов на полках в конце дня
Плюсы	Быстро, дёшево, без риска для пользователей	Ближе к бизнес-метрикам
Минусы	Хороший результат не гарантирует рост бизнес-метрик	Нужны деплой, A/B-тест и риск для пользователей

Путь модели от офлайна до бизнеса

Как всё связано.

Офлайн

Метрики на тесте. Быстро отсеиваем слабые модели.

Деплой

Выкатка в продакшен. Нужна возможность отката.

Онлайн и A/B

Сравниваем с прошлой версией на случайных группах.

Бизнес

Доход и лояльность — итоговая цель.

Офлайн-метрика — прокси онлайн-метрики, онлайн — прокси бизнеса. Хорошее число на тесте не гарантирует эффект в продакшене.

Офлайн или онлайн?

Голосуем.

1. Доля верно поставленных диагнозов на отложенной выборке
2. Среднее число бананов на полках в конце дня
3. Средний модуль разницы между предсказанной и истинной ценой квартиры
4. Медианная длина сессии в онлайн-игре
5. Оценка качества перевода асессорами

Ответы: вид метрики

Метрика	Вид
1. Доля верных диагнозов	Офлайн: считаем по историческим данным
2. Бананы на полках	Онлайн: нужна работающая система
3. Ошибка в цене квартиры	Офлайн: сравниваем с известными ценами
4. Длина сессии	Онлайн: считается по реальным пользователям
5. Оценка асессоров	Офлайн: оценивают до выпуска в продакшен

Какую метрику выбрать?

Вопрос на подумать. Модель отвечает «да» или «нет». Ошибиться можно двумя способами: сказать «да» зря (ложная тревога) или сказать «нет» зря (пропуск).

Обычная поликлиника

Нужно, чтобы диагнозы в целом чаще были верными.

Проходная завода

Камера решает, сотрудник ли перед ней. Пустить чужого — беда, а лишний раз не пустить своего — мелочь.

Опасная болезнь

Пропущенный больной заразит других. Ложная тревога куда дешевле.

Какую метрику выбрать в каждом случае?

Ответ: разные нужды — разные метрики

Accuracy обычная поликлиника	**верные ответы ÷ все ответы** Какая доля ответов оказалась правильной. Подходит, когда обе ошибки одинаково неприятны.
Precision проходная завода	**верно найденные ÷ всё, что модель назвала «да»** Можно ли верить модели, когда она говорит «да». Важна, когда ложное «да» дорого стоит: нельзя пустить чужого.
Recall опасная болезнь	**верно найденные ÷ все, кого нужно было найти** Сколько из настоящих «да» модель нашла. Важна, когда нельзя пропустить: больной заразит других.

Другие метрики

Вопрос на подумать. Какую метрику выбрать здесь?

Видеозвонок

Приложение «надевает» на участника виртуальный костюм рыцаря. Костюм должен ровно лечь на фигуру.

Безопасность

Охрана важного объекта: нужно заметить любого, кто вошёл в кадр.

Пандемия

Контроль, не превышает ли число людей в помещении порог.

Ответ: зависит от цели

Точность положения

Среднеквадратичное отклонение координат опорных точек тела от истинных: костюм должен сесть ровно.

Полнота (recall)

На какой доле кадров с людьми модель их заметила. Точность рамки не так важна.

Число людей

Среднеквадратичное отклонение числа найденных людей от истинного.

Не всё сводится к метрикам

Скорость

Предсказания в режиме реального времени. Требование и к модели, и к серверам.

Компактность

Модель должна помещаться на телефон или другое устройство.

Интерпретируемость

Можно объяснить решение: дадут ли кредит, согласуют ли лечение.

Данные

Данные: сколько и откуда

Чем сложнее задача, тем больше данных нужно. Они должны быть и многочисленными, и качественными.

Открытые датасеты

Их много, но чаще для простых задач, например классификации картинок.

Краудсорсинг

Люди размечают данные. Волонтёры без оплаты — citizen science.

Покупка

Иногда датасет можно купить. Для некоторых задач данных нет нигде.

Неразмеченные данные

Self-supervised learning и обучение представлений: используем данные без ответов.

Типы признаков

Признаки — свойства объекта, по которым модель его описывает. Удобнее всего таблица: строки — объекты, столбцы — признаки.

Численный

Рост, доход. Иногда отдельно выделяют вещественные и целочисленные.

Бинарный

Два значения: да или нет, 0 или 1. Можно считать и числом, и категорией.

Категориальный

Значение из набора вариантов: профессия, день недели.

Ординальный

Категории с порядком: класс опасности вещества от 1 до 4, год обучения.

Какого типа признак?

Вопрос на подумать. Численный · Категориальный · Бинарный · Ординальный

1. Рост человека
2. Профессия
3. Оплатил ли клиент заказ
4. Класс энергопотребления (от A до G)
5. День недели
6. Год обучения студента

Ответ: типы признаков

Признак	Тип
Рост человека	**Численный**
Профессия	**Категориальный**
Оплатил ли заказ	**Бинарный**
Класс энергопотребления, A–G	**Ординальный**
День недели	**Категориальный**
Год обучения	**Ординальный**

Что не так с этими данными?

Найдите проблемы.

Клиент	Возраст	Доход, руб.	Вернул кредит
Анна	34	90 000	да
Борис	N/A	70 000	нет
Виктор	140	85 000	да
Галина	29	2 400 000	да
Дмитрий	41	65 000	да

Часть проблем в таблице не видна вовсе. Какие это?

Ответ: четыре типа проблем в данных

Пропуски

Возраст Бориса неизвестен. Удалить, заполнить или оставить как признак: пропуск сам может нести информацию.

Выбросы

140 лет и доход 2,4 млн — ошибка сбора или реальная аномалия. Важные выбросы обрабатывают отдельно.

Ошибки разметки

Разметчики ошибаются: бывают недобросовестные, а задача может быть сложной.

Data drift

Данные меняются со временем: формат или распределение. Нужен мониторинг.

Модель и обучение

Модель и алгоритм обучения

Модель — способ описать мир

«Земля плоская» — тоже модель. В масштабе города она работает отлично, а на маршруте Париж — Лос-Анджелес ломается, и её приходится заменить.

Предсказательная модель

$$f(x;\theta) \to y$$

x — признаки объекта, y — таргет, θ — параметры, которые подбираются по данным.

Простейшая модель: одна константа

Всем квартирам предсказываем одну и ту же цену c, то есть $f(x) = c$. Цены пяти квартир, млн руб.: **5, 6, 7, 8, 24**.

MAE

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|f(x_i) - y_i|$$

Средний модуль ошибки: насколько в среднем промахиваемся.

Лучшая c — медиана: 7

MSE

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\big(f(x_i) - y_i\big)^2$$

Средний квадрат ошибки: большие промахи штрафуются сильнее.

Лучшая c — среднее: 10

Медиана — число посередине, если выстроить цены по порядку. Одна дорогая квартира в 24 млн сдвинула среднее вверх, а медиану почти не тронула. Здесь N — число квартир, i — номер квартиры.

Что нужно, чтобы обучить модель

Модель

Формула с настраиваемыми числами θ. Меняем θ — меняются предсказания.

Функция потерь

Число, которое показывает, насколько модель ошибается. Чем меньше, тем лучше.

Алгоритм обучения

Подбирает θ так, чтобы потери стали меньше. Например, градиентный спуск: понемногу меняем θ туда, где ошибка убывает.

Градиентный спуск в одну строку

$$\theta \leftarrow \theta - \eta \, \nabla L(\theta)$$

Здесь η — размер шага, ∇L — направление, в котором ошибка растёт быстрее всего. Идём в противоположную сторону.

Параметры и гиперпараметры

Параметры

Числа внутри модели. Их подбирает алгоритм обучения по данным.

- коэффициенты k и b в прямой $y = kx + b$
- веса нейросети
- параметры θ из прошлого слайда

Гиперпараметры

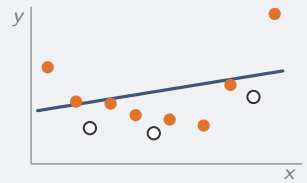
Настройки самого обучения. Их задаёт человек до начала обучения.

- шаг η в градиентном спуске
- степень многочлена
- число слоёв нейросети
- сколько раз пройти по данным

Недообучение и переобучение

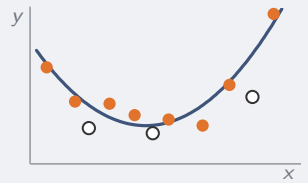
● обучающие точки ○ тестовые точки (новые данные)

Слишком простая



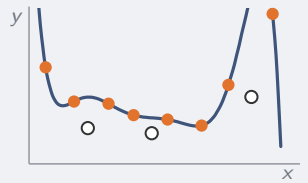
Прямая не ловит изгиб данных.

Нормальная



Улавливает общий тренд, не гонясь за каждой точкой.

Переобученная



Проходит почти через каждую обучающую точку, но между ними и на новых данных ошибается.

Ошибка на train и на test



train (тёмная линия) падает с ростом сложности, иногда до нуля.

test (оранжевая линия) сначала падает, потом снова растёт.

Оптимум — модель не слишком простая и не слишком сложная.

Метрики считаем на тесте и никогда не смешиваем его с обучающей выборкой. По горизонтали растёт сложность модели (степень многочлена от 0 до 5), по вертикали — MSE.

После обучения

Деплой

Выкатка в продакшен: эффективный код, параллельность, дружба с фреймворками.

Мониторинг

Следим за качеством, дообучаем и переобучаем на новых данных.

A/B-тест

Сравнение с предыдущей версией на случайных группах. Если хуже — откат.

Drift

Data drift: меняются данные. Concept drift: меняется связь признаков и таргета.