Машинное обучение

Лекция 1
Введение в машинное обучение

Андрей Нарцев
andrei.nartsev@gmail.com
anartsev@hse.ru

НИУ ВШЭ, 2024

Организационное

Команда курса:

- Лекции: Андрей Нарцев @andreynar
- Семинары: Александра Коган и Артем Рыжиков
- Ассистенты: to be continued

Организационное

Устройство курса:

16 лекций и 16 семинаров по «классическому» машинному обучению

Организационное

Что будет в курсе:

Основные алгоритмы машинного обучения

- Метод k-ближайших соседей
- Линейные модели
- Деревья и ансамбли на основе бэггинга и бустинга

Организационное

Что будет в курсе:

Популярные задачи машинного обучения

- Регрессия
- Классификация
- Ранжирование
- Кластеризация
- Снижение размерностей

Организационное

Чего **НЕ** будет в курсе:

- Глубинное обучение (нейронные сети)
- MLOps (деплой ML моделей)
- Внедрение ML моделей (A/B-эксперименты)

Формула оценивания

$$О_{\text{итоговая}} = 0.4 * \overline{О}_{\text{домашние задания}} + 0.1 * \overline{О}_{\text{проверочные работы}} + 0.2 * О_{\text{контрольная работа}} + 0.3 * О_{\text{экзамен}}$$

Из каких компонент состоит оценка:

- Домашние задания: несколько за весь курс (предположительно 4)
- Проверочные работы: несколько несложных теоретических вопросов, пишем в начале некоторых лекций
- Контрольная работа: после первого модуля
- Экзамен: письменный

Критерии автомата

$$О_{\text{итоговая}} = О_{\text{накопленная}} = \frac{10}{7}(0.4 * \overline{О}_{\text{домашние задания}} + 0.1 * \overline{О}_{\text{проверочные работы}} + 0.2 * О_{\text{контрольная работа}})$$

Можно получить автомат **при выполнение следующих условий:**

1. Накопленная оценка должна быть не ниже 5.5
2. Оценка за контрольную должна быть не ниже 5.5
3. Не был установлен факт плагиата ни одной домашней, проверочной и контрольной работ

Организационное

Информация о курсе:

- Канал с объявлениями: https://t.me/+TglJAAxY7cNhNGYy
- Материалы на github: https://github.com/andrewnarts/hse-ml/tree/main/math-faculty-intro-ml/2024

Что нужно для успешного освоения курса?

Для лекций:

- Основы математического анализа и линейной алгебры (векторное и матричное дифференцирование)
- Основы теории вероятностей и математической статистики (когда будем говорить про калибровку моделей)

Для семинаров:

- Уметь писать код на Python
- Понимать основные принципы ООП

О чем еще помнить?

- Все виды работ проверяются на плагиат (помним, что факт плагиата лишает возможности получения автомата, так же работу нельзя будет пересдать)
- Важно коммуницировать с ассистентами по поводу проверки домашних работ и отвечать на вопросы, если они у них возникают
- Можно задавать любые вопросы по курсу лектору, семинаристу, ассистенту
- 9 и 10 – это очень высокие оценки по данному курсу, возможно, для их получения придется потратить значительное время

План лекции

- Небольшой исторический экскурс
- Основные понятия
- Виды задач
- Обучение и оценка качества моделей

Про историю машинного обучения
(очень коротко)

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

Слишком много ручного труда

для создания системы

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

НАЧАЛО 21 ВЕКА — ГЛУБИННОЕ ОБУЧЕНИЕ (DEEP LEARNING)

Решение сложных задач распознавания с точностью, близкой к человеку

ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

Машинное обучение – набор способов воспроизведения связей между событиями и результатом.

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Machine learning – the field of study that gives computers the ability to learn without being explicitly programmed.

Основные понятия

Пример задачи

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?

* см. kaggle.com, TFI Restaurant Revenue Prediction

Обозначения

- x — объект — для чего хотим делать предсказания
  - Конкретное расположение ресторана
- $\mathbb{X}$ — пространство всех возможных объектов
  - Все возможные расположения ресторанов
- y — ответ, целевая переменная, target — что предсказываем
  - Прибыль в течение первого года работы
- $\mathbb{Y}$ — пространство ответов — все возможные значения ответа
  - Все вещественные числа

Обучающая выборка

- Мы ничего не понимаем в экономике
- Зато имеем много объектов с известными ответами
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- ℓ — размер выборки

Признаки

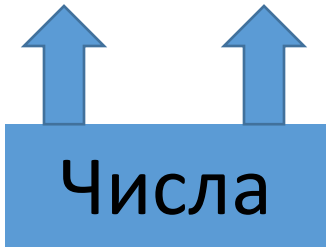
- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание

Признаки

- Объекты — абстрактные сущности
- Компьютеры работают только с числами
- Признаки, факторы, features — числовые характеристики объектов
- d — количество признаков
- $x = (x_1, \dots, x_d)$ — признаковое описание



Признаки

- Про демографию:
  - Средний возраст жителей ближайших кварталов
  - Динамика количества жителей
- Про недвижимость:
  - Средняя стоимость квадратного метра жилья поблизости
  - Количество школ, банков, магазинов, заправок
  - Расстояние до ближайшего конкурента
- Про дороги:
  - Среднее количество машин, проезжающих мимо за день

Алгоритм

- $a(x)$ — алгоритм, модель — функция, предсказывающая ответ для любого объекта
- Отображает $\mathbb{X}$ в $\mathbb{Y}$
- Линейная модель: $a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$
- Например:

$$a(x) = 1.000.000 + 100.000 * (\text{расстояние до конкурента}) \\ -100.000 * (\text{расстояние до метро})$$

Функция потерь

- Не все алгоритмы полезны
- $a(x) = 0$ — не принесет никакой выгоды
- Функция потерь — мера корректности ответа алгоритма
- Предсказали $10000 прибыли, на самом деле $5000 — хорошо или плохо?
- Квадратичное отклонение: $(a(x) - y)^2$

Функционал ошибки

- Функционал ошибки, метрика качества — мера качества работы алгоритма на выборке
- Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2$$

- Чем меньше, тем лучше

Функционал ошибки

- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обучение алгоритма

- Есть обучающая выборка и функционал ошибки
- Семейство алгоритмов $\mathcal{A}$
  - Из чего выбираем алгоритм
  - Пример: все линейные модели
  - $\mathcal{A} = \{w_0 + w_1 x_1 + \cdots + w_d x_d \mid w_0, w_1, \ldots, w_d \in \mathbb{R}\}$
- Обучение: поиск оптимального алгоритма с точки зрения функционала ошибки

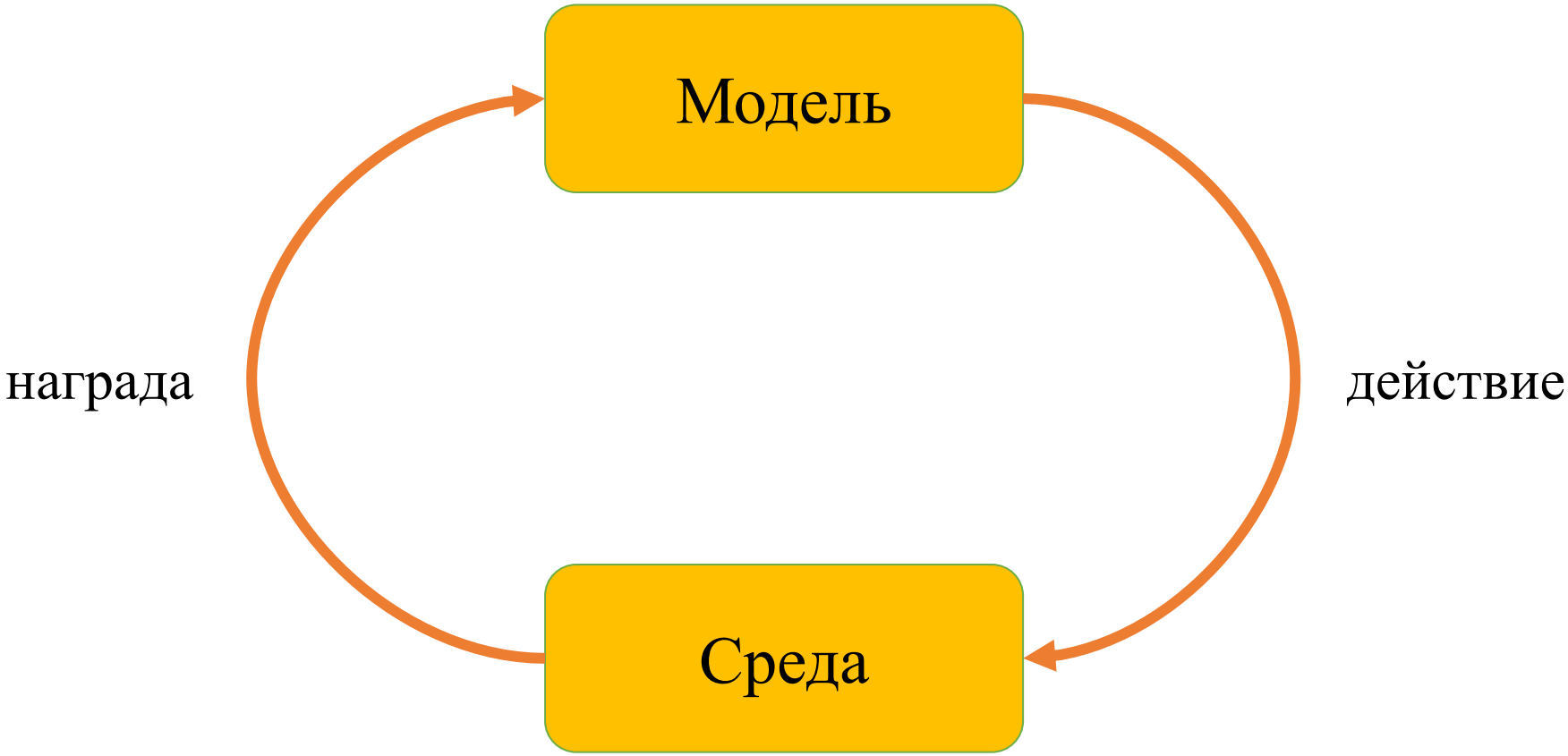
$$a(x) = \arg\min_{a \in \mathcal{A}} Q(a, X)$$

Машинное обучение

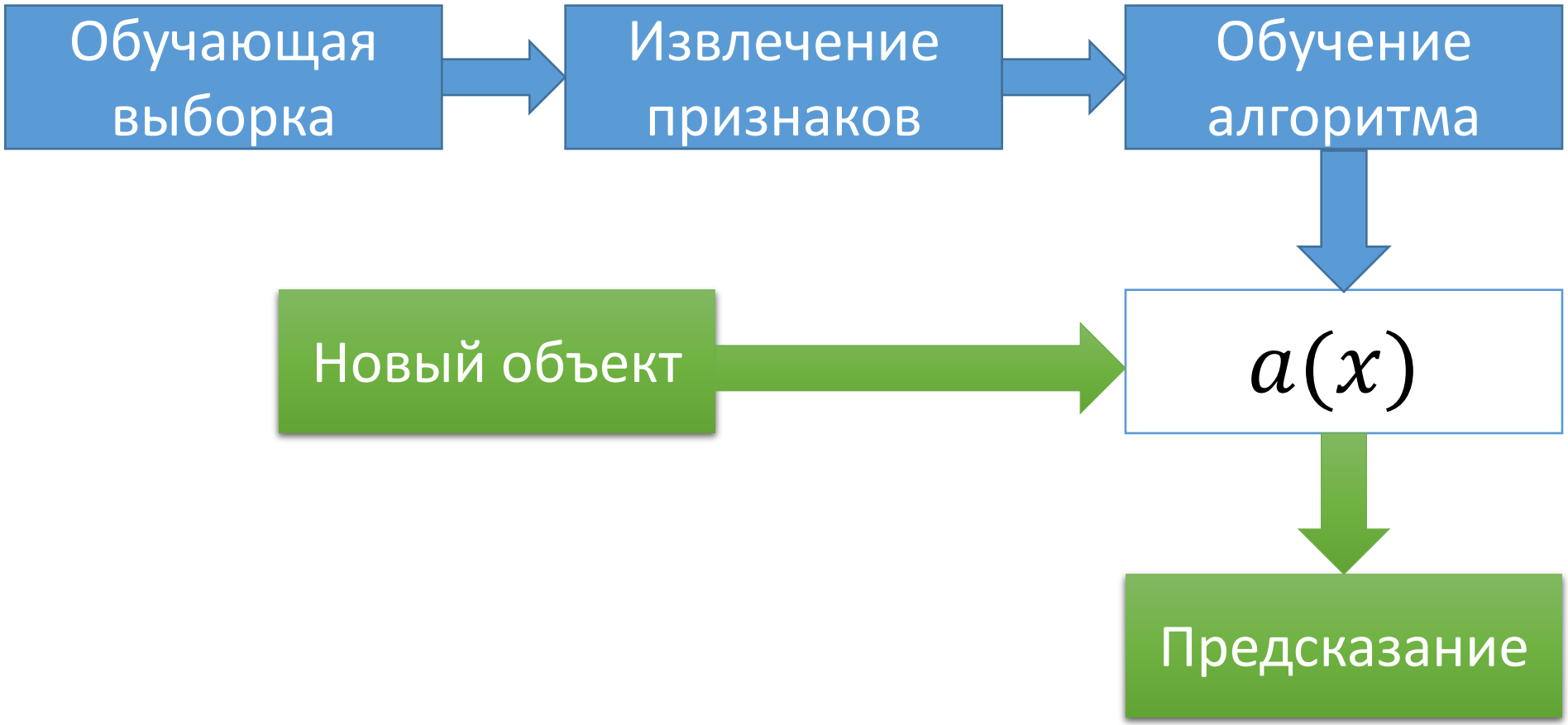
- Не все задачи имеют такую формулировку!

- Обучение без учителя
- Обучение с подкреплением
- И т.д.

Обучение с подкреплением



Машинное обучение



Что нужно знать

1. Как сформулировать задачу?
2. Какие признаки использовать?
3. Откуда взять обучающую выборку?
4. Как подготовить обучающую выборке?
5. Как выбрать метрику качества?
6. Как обучить алгоритм?
7. Как оценить качество алгоритма?
8. Как потом внедрить алгоритм и поддерживать его?

Виды задач

ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0,1\}^M$ - классификация на M классов, которые могут пересекаться

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного скоринга (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

Мультиклассовая классификация

- Определение типа объекта на изображении

Pedestrian

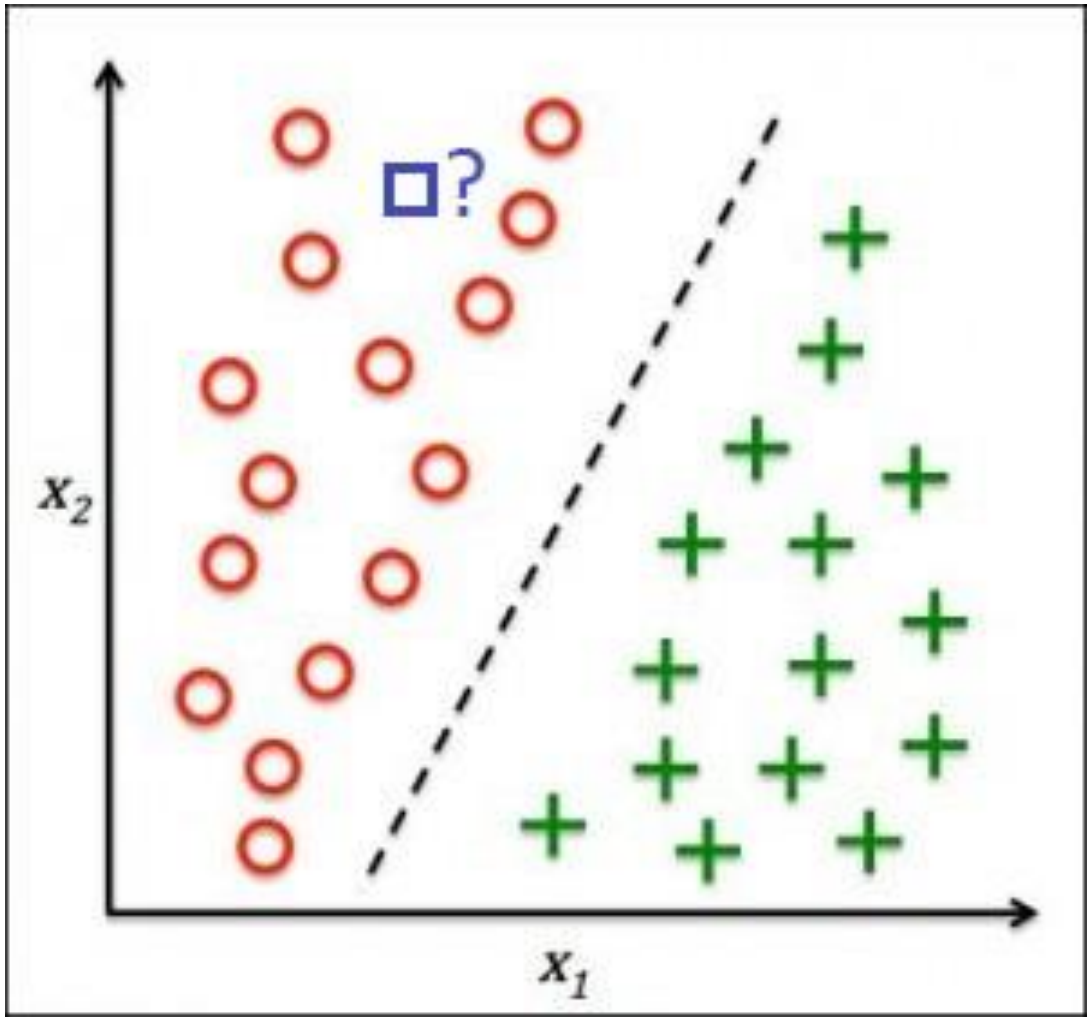
Car

Motorcycle

Truck

- Определение наиболее подходящей профессии для данного кандидата

ЗАДАЧА КЛАССИФИКАЦИИ



ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0,1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

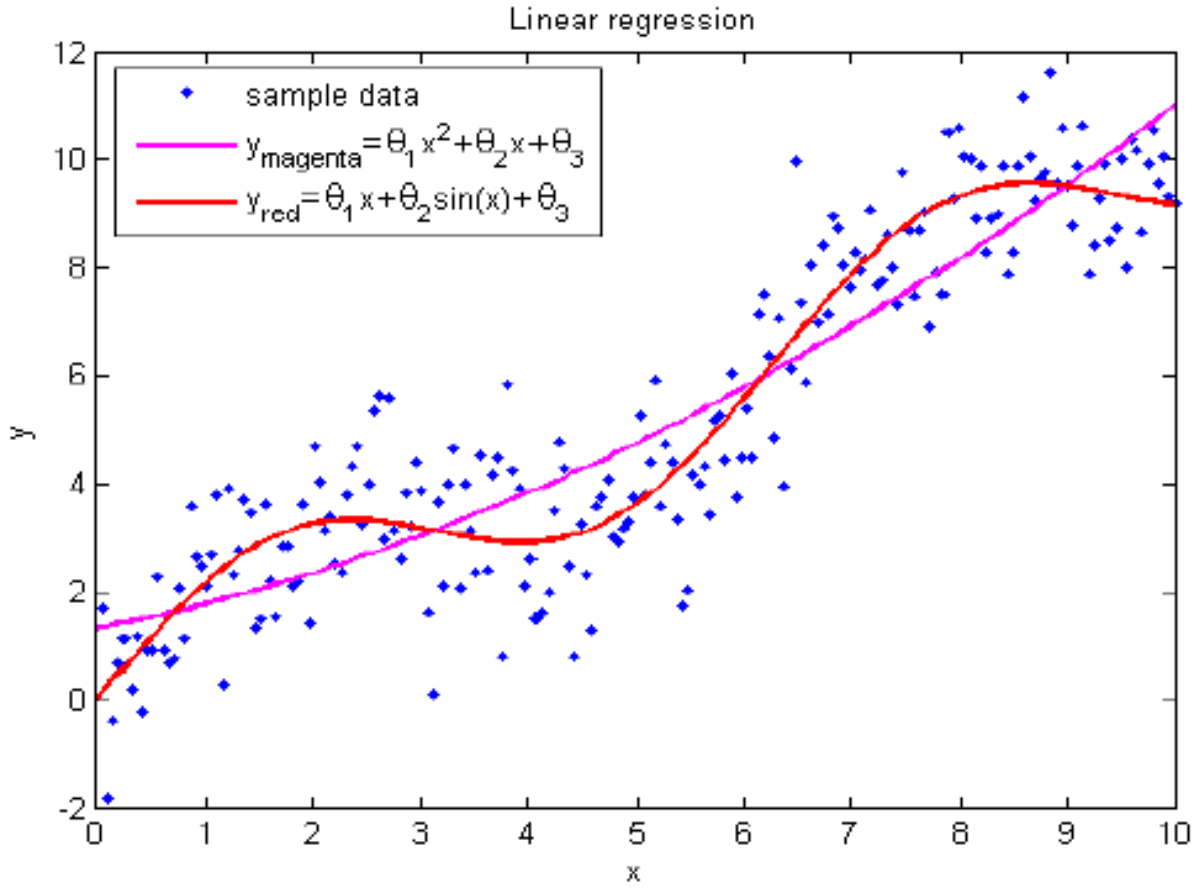
- $Y = R$ или $Y = R^n$

ПРИМЕРЫ ЗАДАЧ РЕГРЕССИИ

- Предсказание стоимости недвижимости (стоимость квартиры в Москве)
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам

ЗАДАЧА РЕГРЕССИИ

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0,1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

- $Y = R$ или $Y = R^n$

Ранжирование

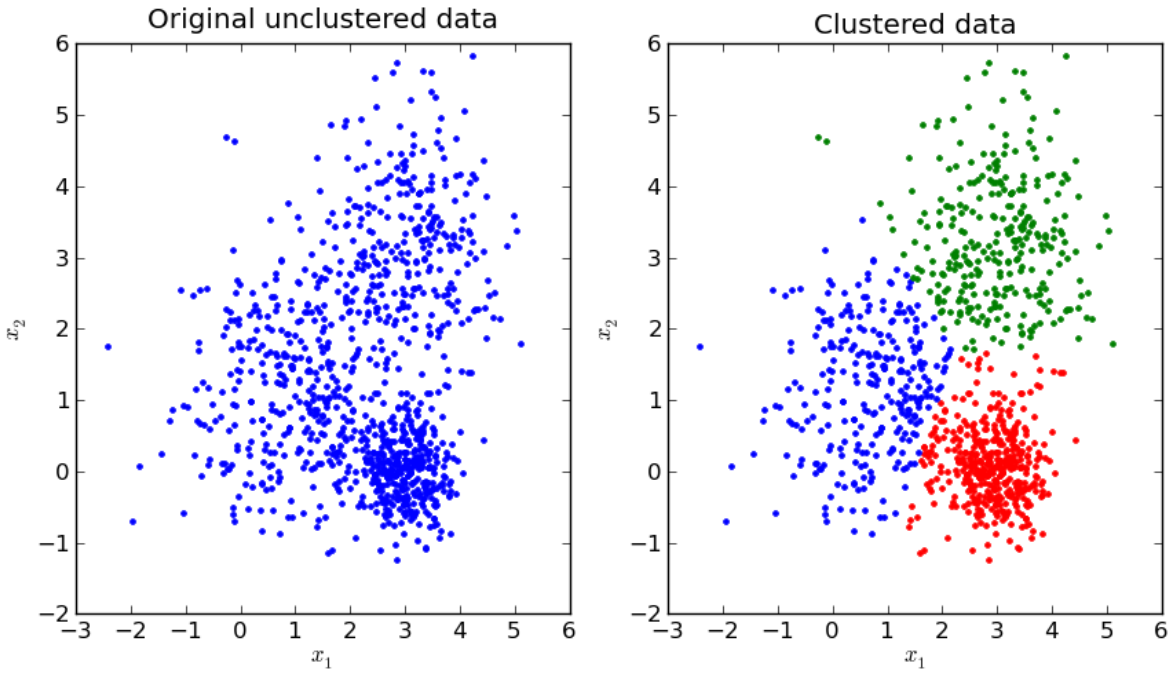
- Y – конечное упорядоченное множество

ПРИМЕРЫ ЗАДАЧ РАНЖИРОВАНИЯ

- Вывести подходящие запросу документы в порядке уменьшения релевантности
- Вывести кандидатов на должность в порядке уменьшения релевантности

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.



ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

- Разбить пользователей на группы, внутри каждой из которых будут похожие пользователи
- Разбить текстовые документы на группы по похожести документов

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

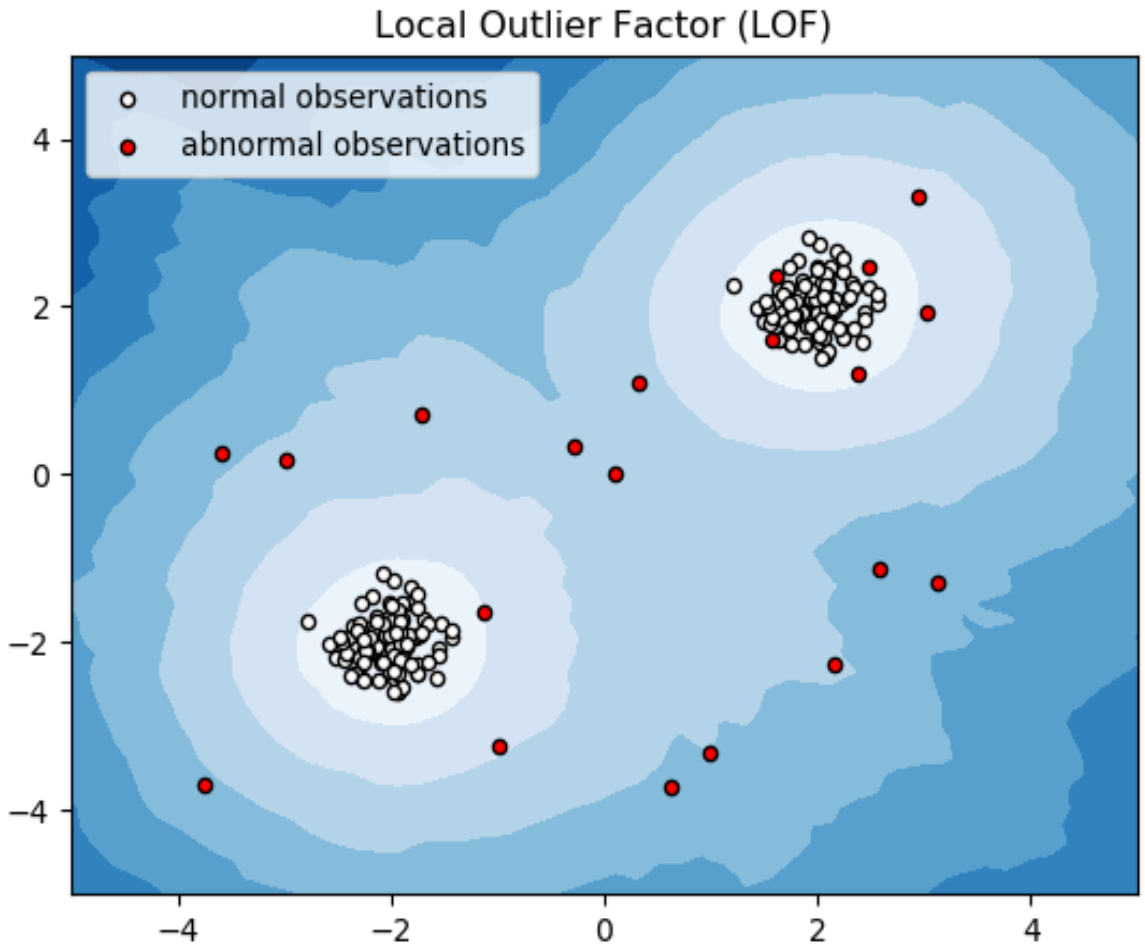
- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности –** задача приближения распределения объектов.

ПРИМЕР ОЦЕНИВАНИЯ ПЛОТНОСТИ

- Поиск аномалий с помощью оценивания плотностей



ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация –** задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности –** задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности –** задача приближения распределения объектов.
- **Визуализация –** задача изображения многомерных объектов в 2х или 3хмерном пространстве с сохранением зависимостей между ними.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.