



Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н., Н.С.

Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и
мониторинга климатических изменений (ЛВОАМКИ)

Организационные вопросы

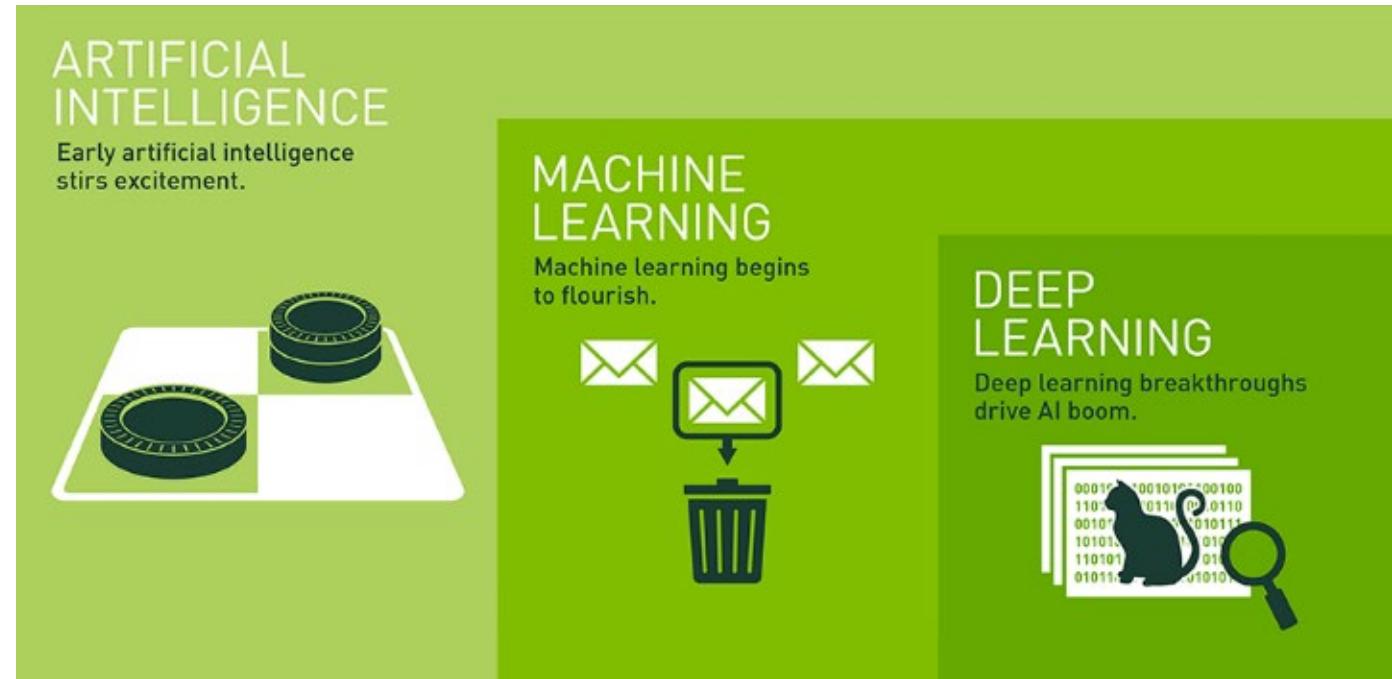
- Материалы курса, 1й год:
 - <https://github.com/MKrinitkiy/ML4ES1-F2021-S2022>
- ДЗ: krinitsky.ma@phystech.edu
- Telegram: @mkrinitkiy (<https://t.me/mkrinitkiy>)
- Расписание: вторник 15:30 – 17:50

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

МАШИННОЕ ОБУЧЕНИЕ

ГЛУБОКОЕ ОБУЧЕНИЕ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, МАШИННОЕ ОБУЧЕНИЕ, ГЛУБОКОЕ ОБУЧЕНИЕ

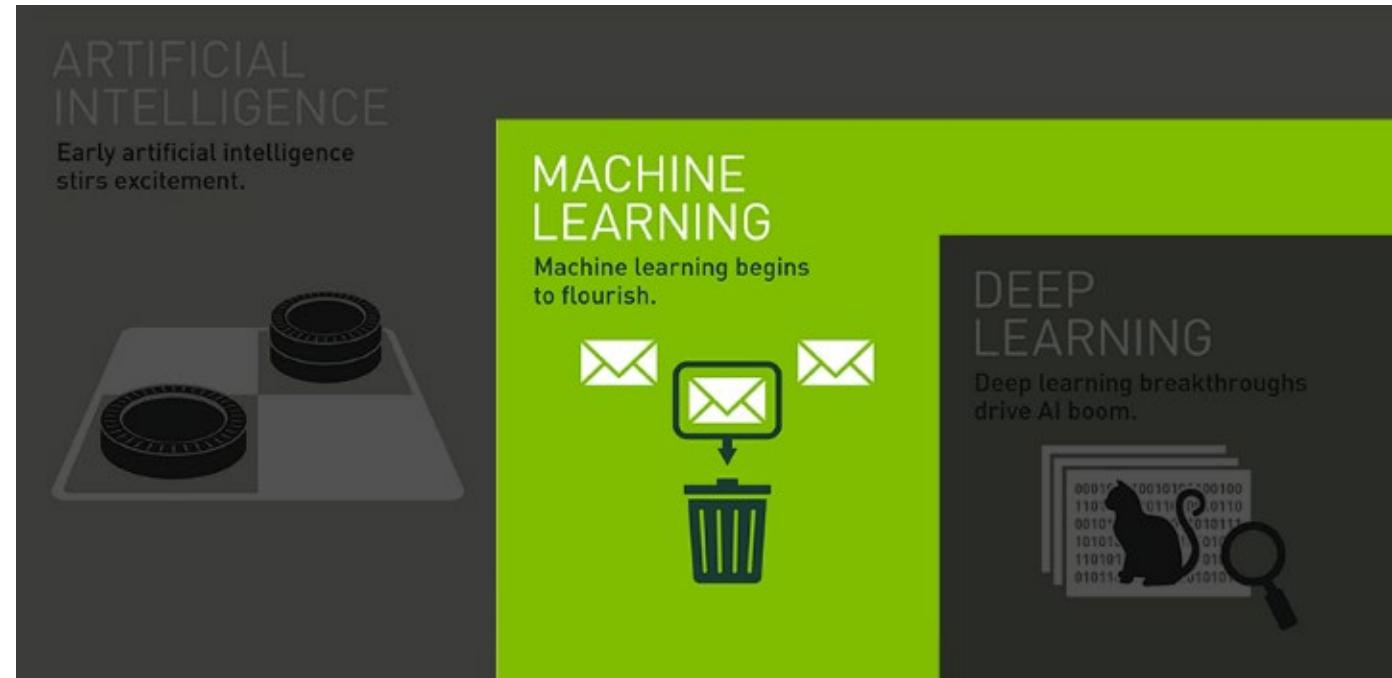


Искусственный интеллект (ИИ) - это наука и инженерная технология создания интеллектуальных машин, и в особенности интеллектуальных компьютерных программ. ИИ связан со сходной задачей использования компьютеров для понимания человеческого интеллекта, но не обязательно ограничивается биологически правдоподобными методами.

Дж. МакКарти, 1956г.

ИИ - научное направление, в рамках которого **ставятся и решаются задачи** аппаратного или программного **моделирования** тех **видов человеческой деятельности**, которые **традиционно считаются интеллектуальными**

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, МАШИННОЕ ОБУЧЕНИЕ, ГЛУБОКОЕ ОБУЧЕНИЕ

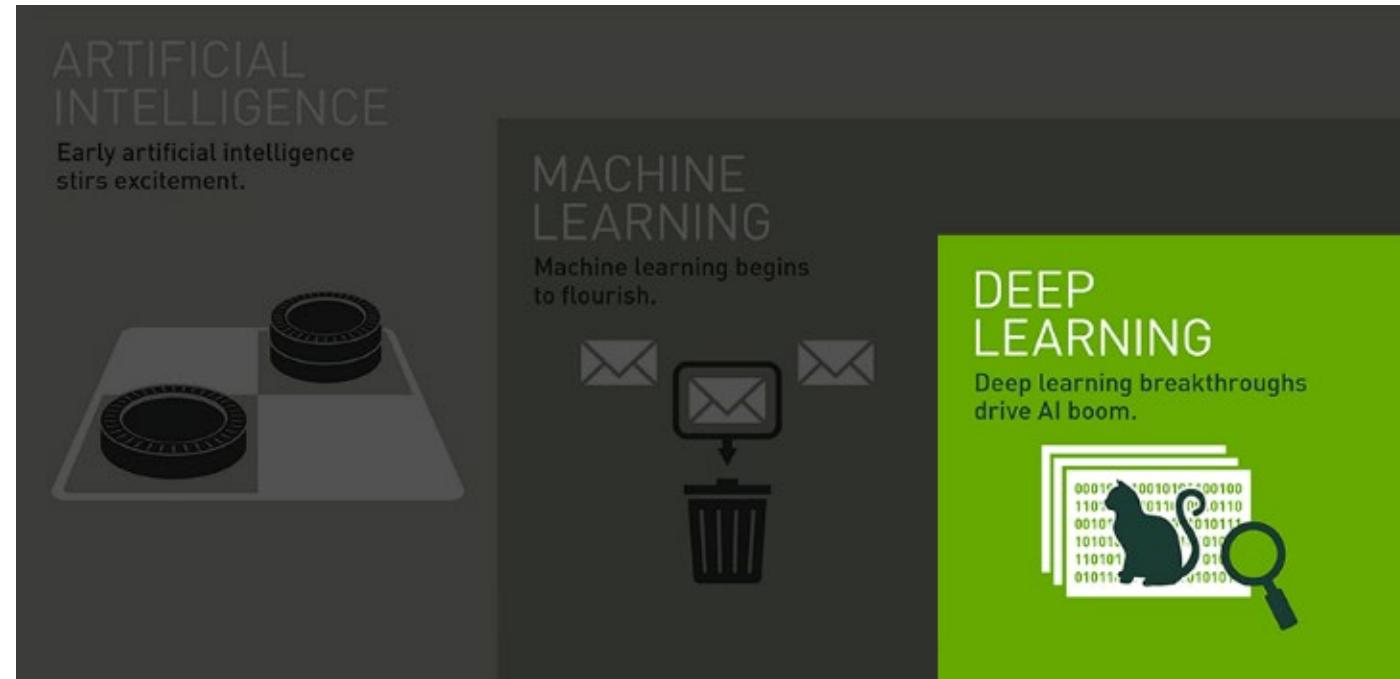


Машинное обучение (МО)

«область компьютерных наук, придающих способность компьютерам обучаться без необходимости явно их программировать»⁽¹⁾

⁽¹⁾A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, vol. 3, no. 3, p.p. 210-229, July 1959. doi: 10.1147/rd.33.0210

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, МАШИННОЕ ОБУЧЕНИЕ, ГЛУБОКОЕ ОБУЧЕНИЕ



Глубинное обучение (Deep Learning, DL)

Совокупность **методов машинного обучения**, основанных на обучении представлениям данных, а не специализированным алгоритмам, предназначенным для решения конкретных задач.



Классификация задач и методов машиинного обучения

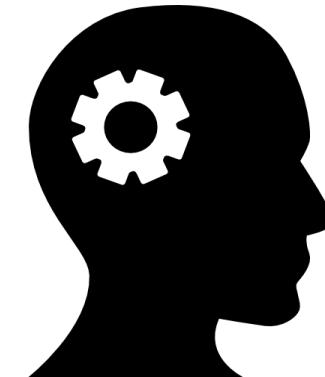
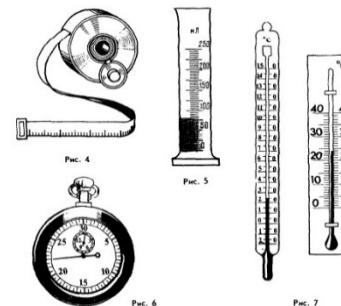
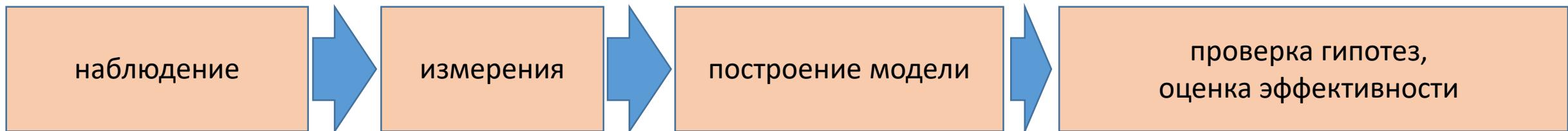
Михаил Криницкий

к.т.н., н.с.

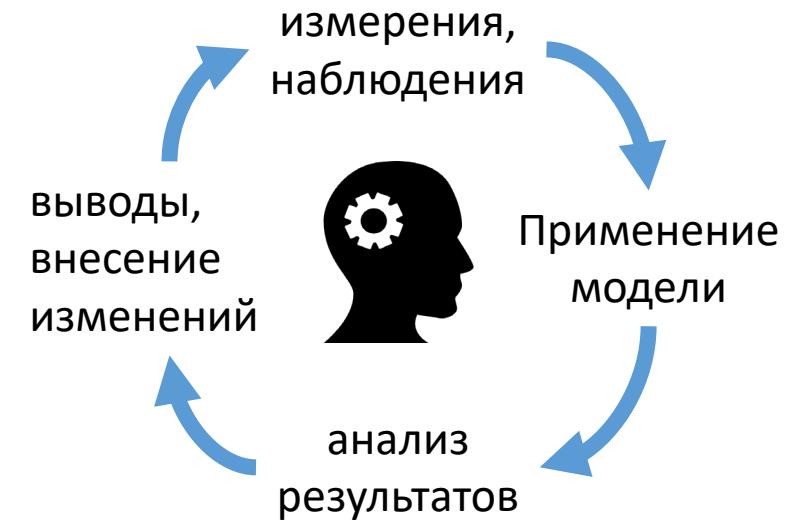
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и
мониторинга климатических изменений (ЛВОАМКИ)

Когда (человеку) непонятно, что происходит все равно строим модель



обобщение ?
введение абстракций ?



ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ НАУК О ЗЕМЛЕ

моделирование «из первых принципов»

- Трактуемость (интерпретируемость)
- В сложных случаях – долгая реализация
- Меняющиеся условия и неучтенные воздействия снижают точность
- Подход основан на понимании описываемого процесса. Модель процесса от начала до конца формулируется человеком.
поэтому нередко модель **недостаточно сложна** для описания сложного процесса с **желаемой точностью**.

моделирование, основанное на данных

- Трактуемость (интерпретируемость) – редко
ММО не всегда подходят для установления причинно-следственных зависимостей
- Скорость реализации
- Модели можно «дообучать» в меняющихся условиях
- Подход основан на данных. Человек может участвовать (помогать) на этапе выбора типа модели общего вида.
модели МО обладают широким спектром сложности, что **позволяет аппроксимировать** очень сложные процессы с **хорошей точностью**.

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: строим модель для решения задачи

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: строим модель для решения задачи

Балансируем:

интерпретируемость

точность («качество»)

способность к обобщению

энерго-, вычислительные затраты

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

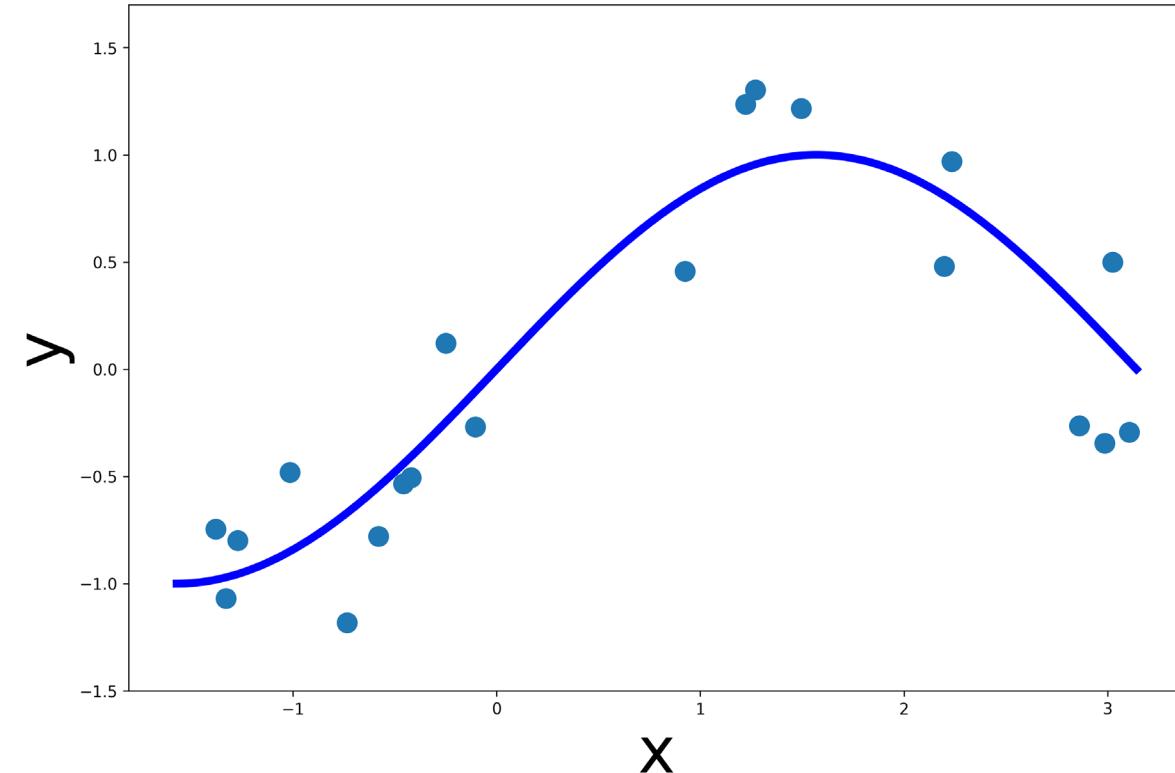
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»
 - восстановление регрессии

ЧТО Я ХОЧУ? – значение y

$$y \in \mathbb{R}^m$$

m – размерность целевой переменной



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

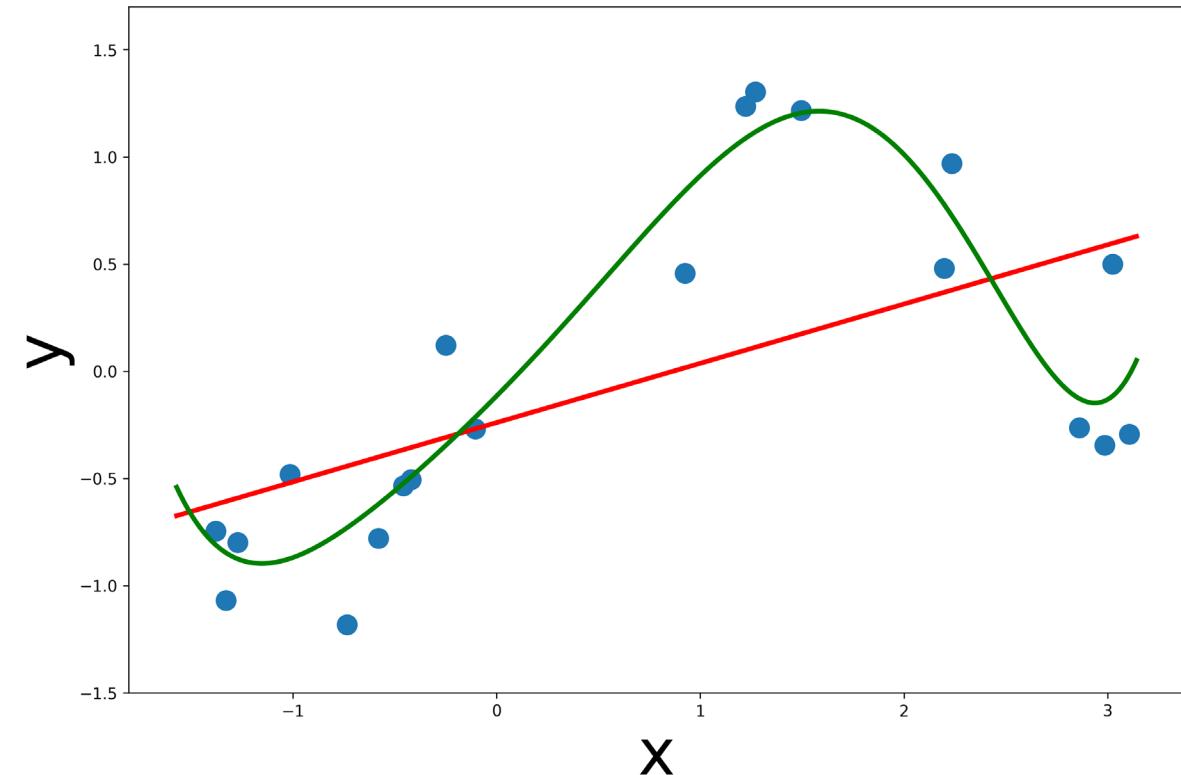
- «Обучение с учителем»
 - восстановление регрессии

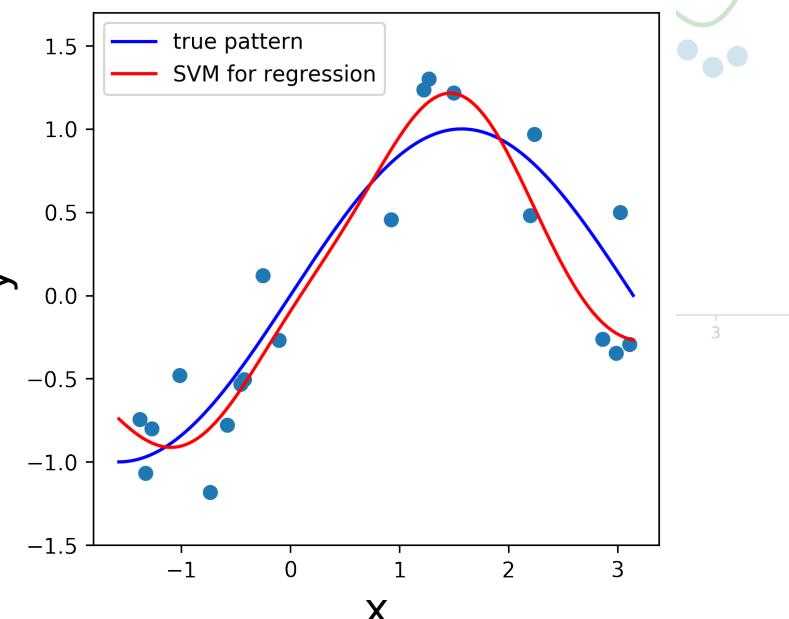
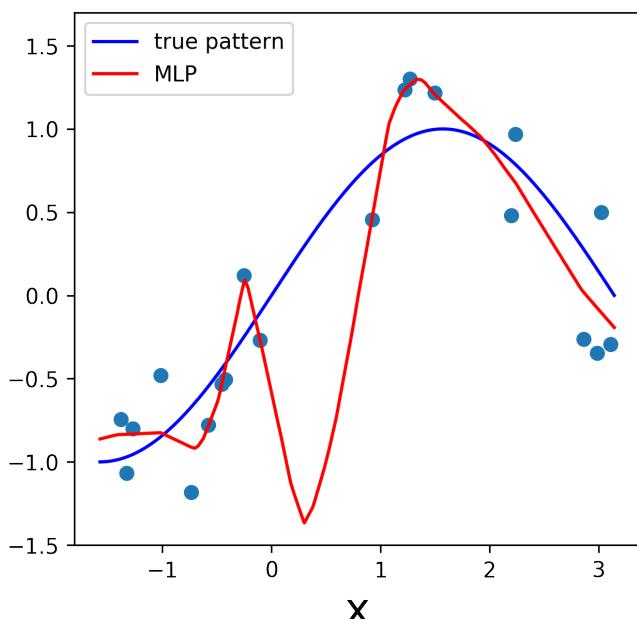
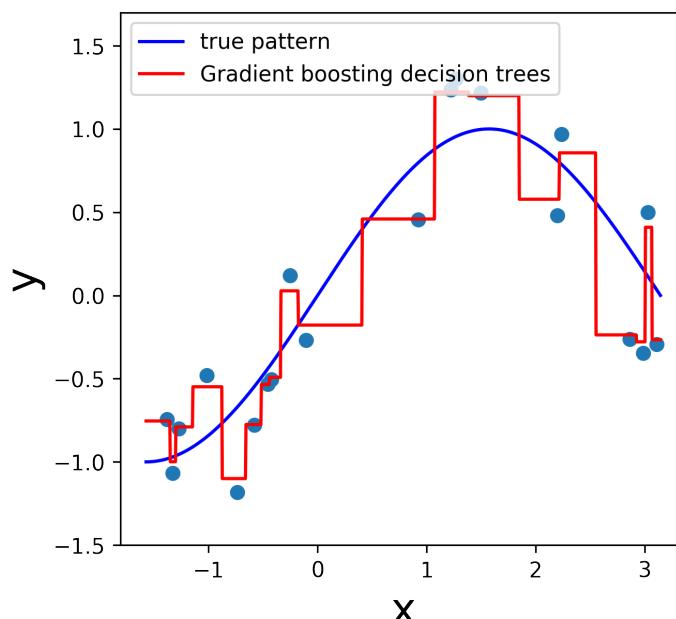
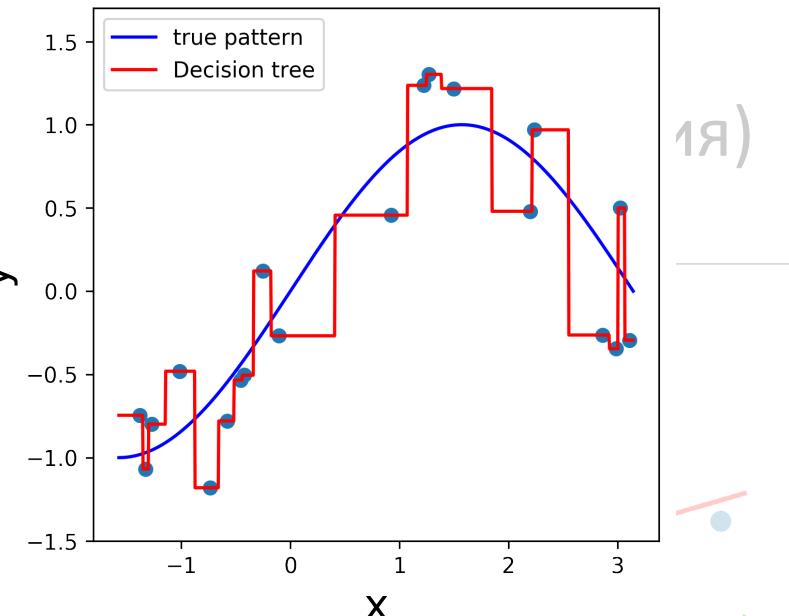
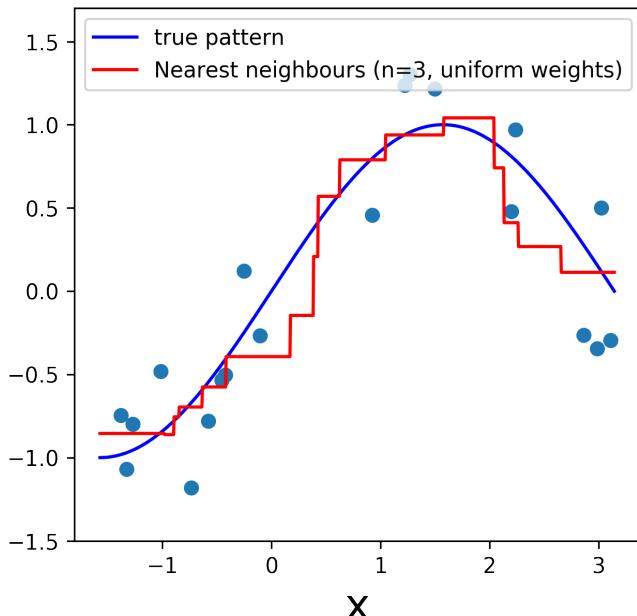
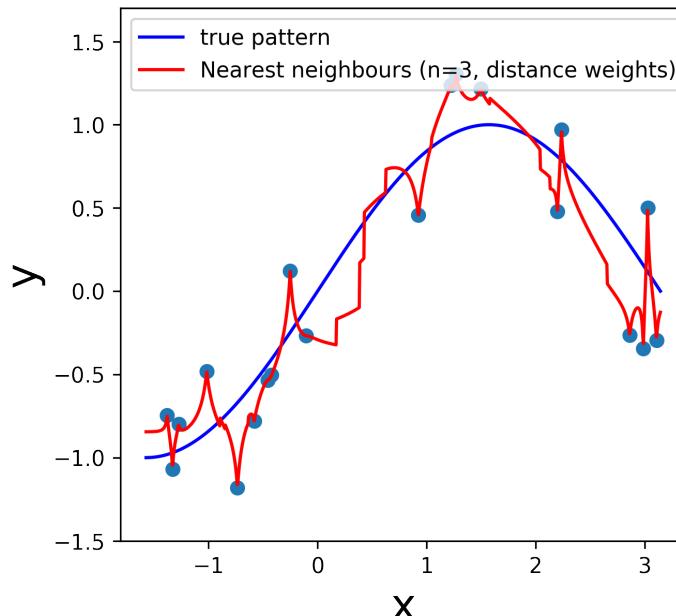
Примеры моделей:

Линейная регрессия

$$\hat{y} = ax + b$$

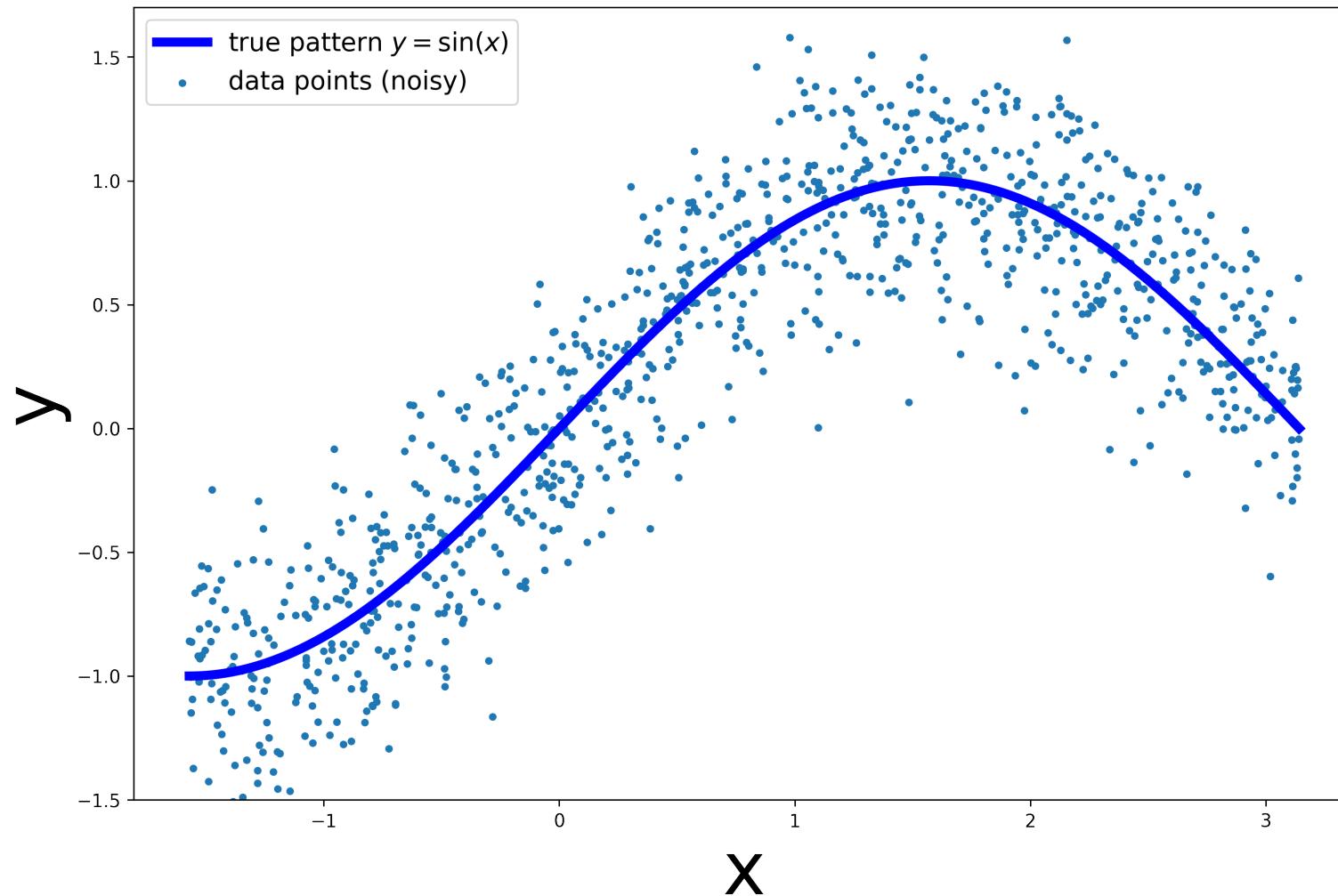
$$\hat{y} = p^{(6)}(x)$$





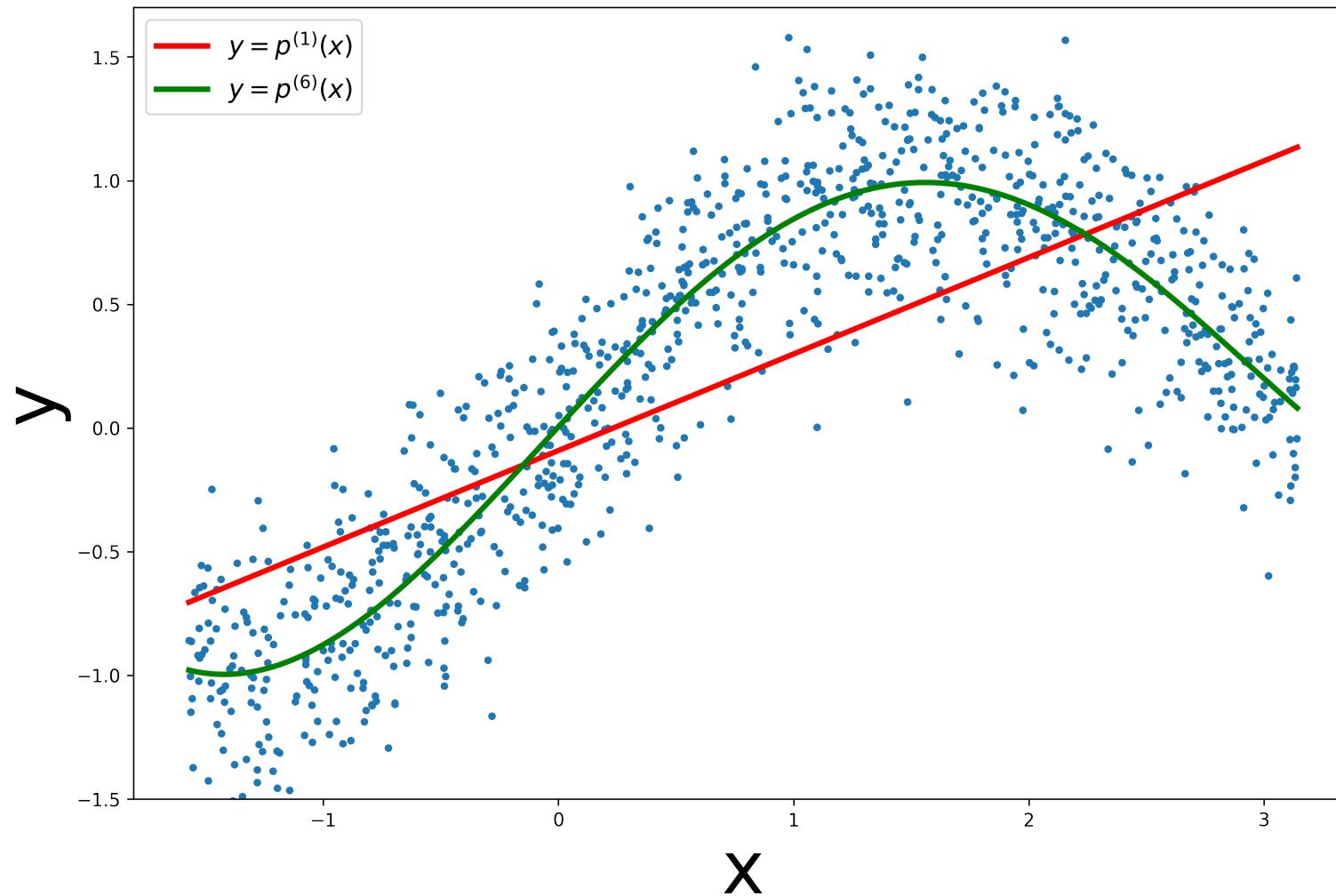
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

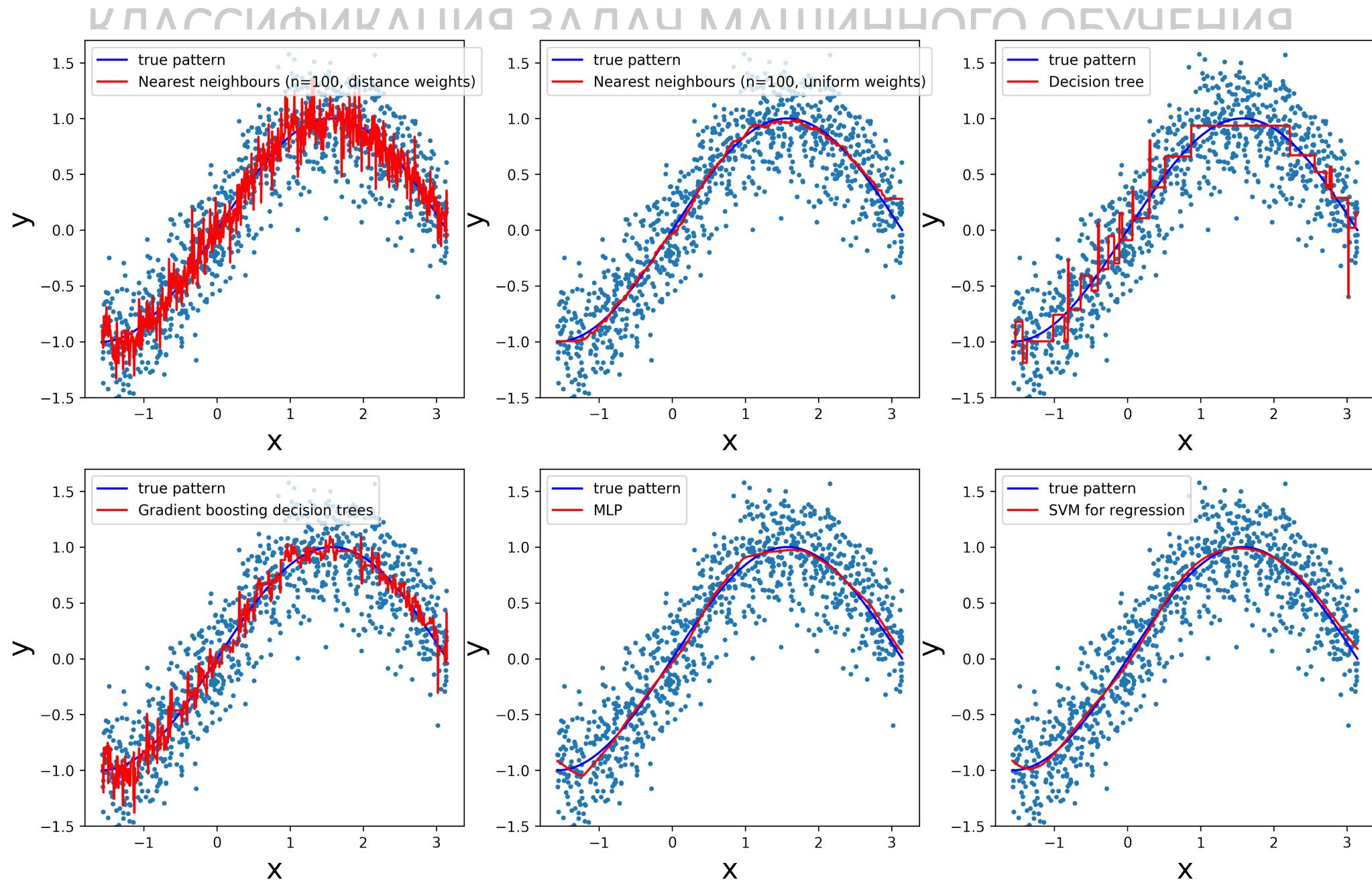
ЦЕЛЬ: сформулировать задачу (в терминах машинного обучения)



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: сформулировать задачу (в терминах машинного обучения)





Что интересного можно сказать
про эти модели
и их результаты?

ремарки

- Количество размеченных данных (зачастую) играет роль
- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных
- Сложная точная модель – не обязательно лучшая для конкретной задачи

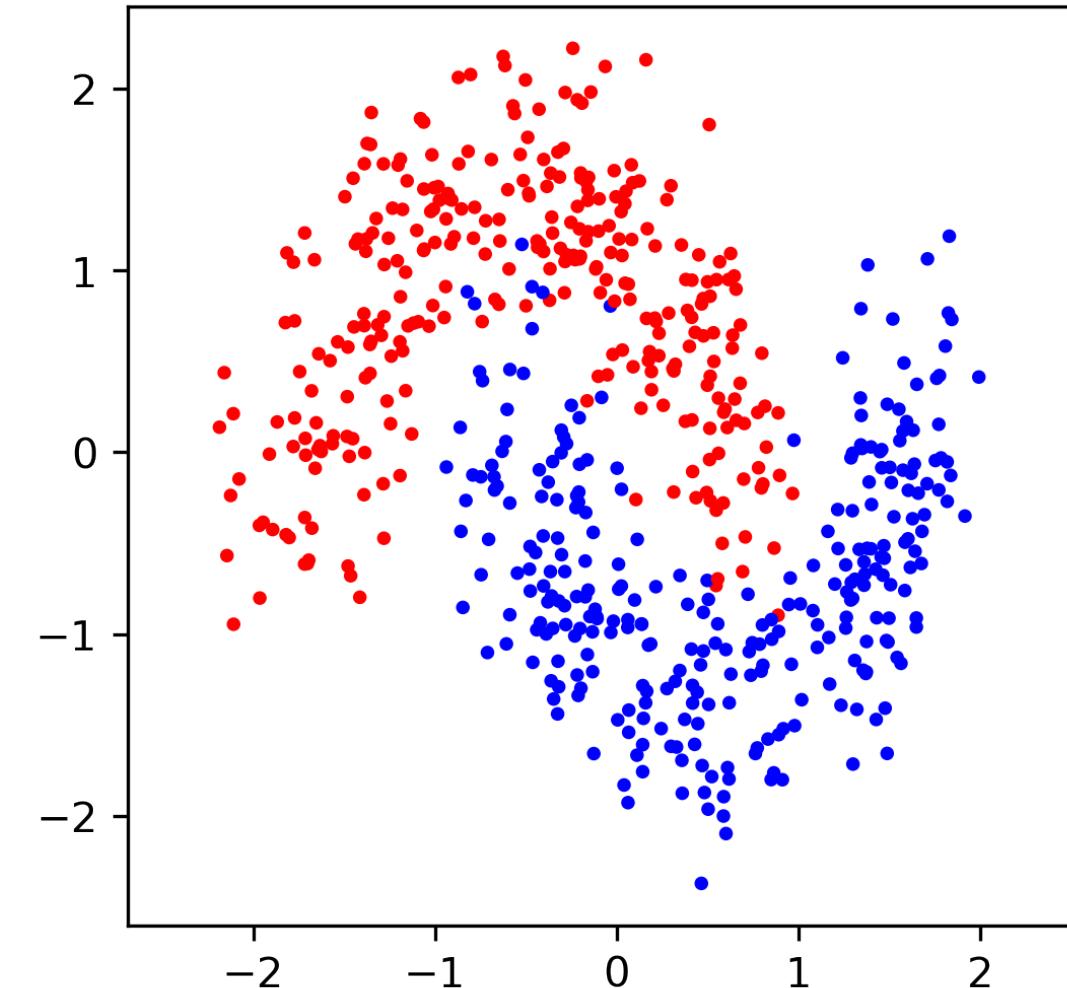
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»

- восстановление регрессии
- классификация

ЧТО Я ХОЧУ? – метку класса
«красный или синий?»
(бинарная классификация)



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

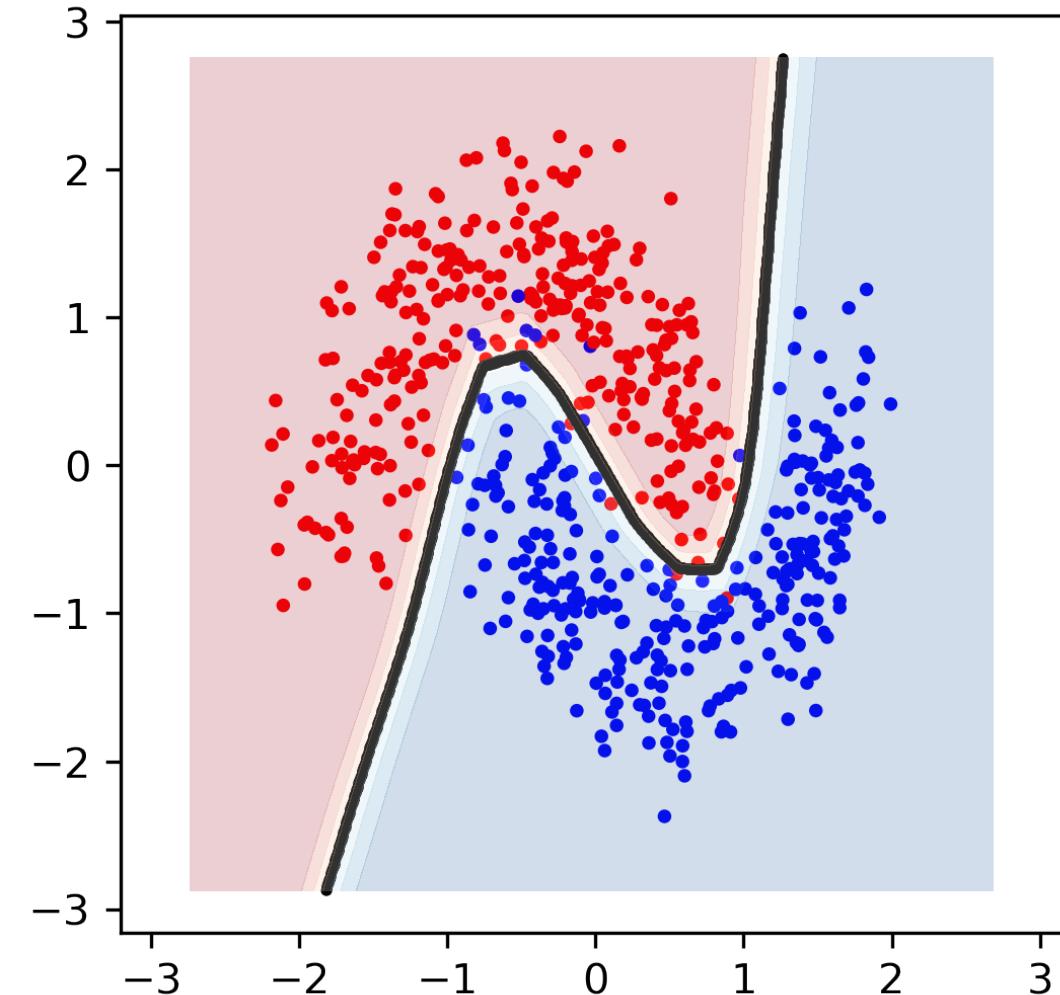
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

- «Обучение с учителем»

- восстановление регрессии
- классификация

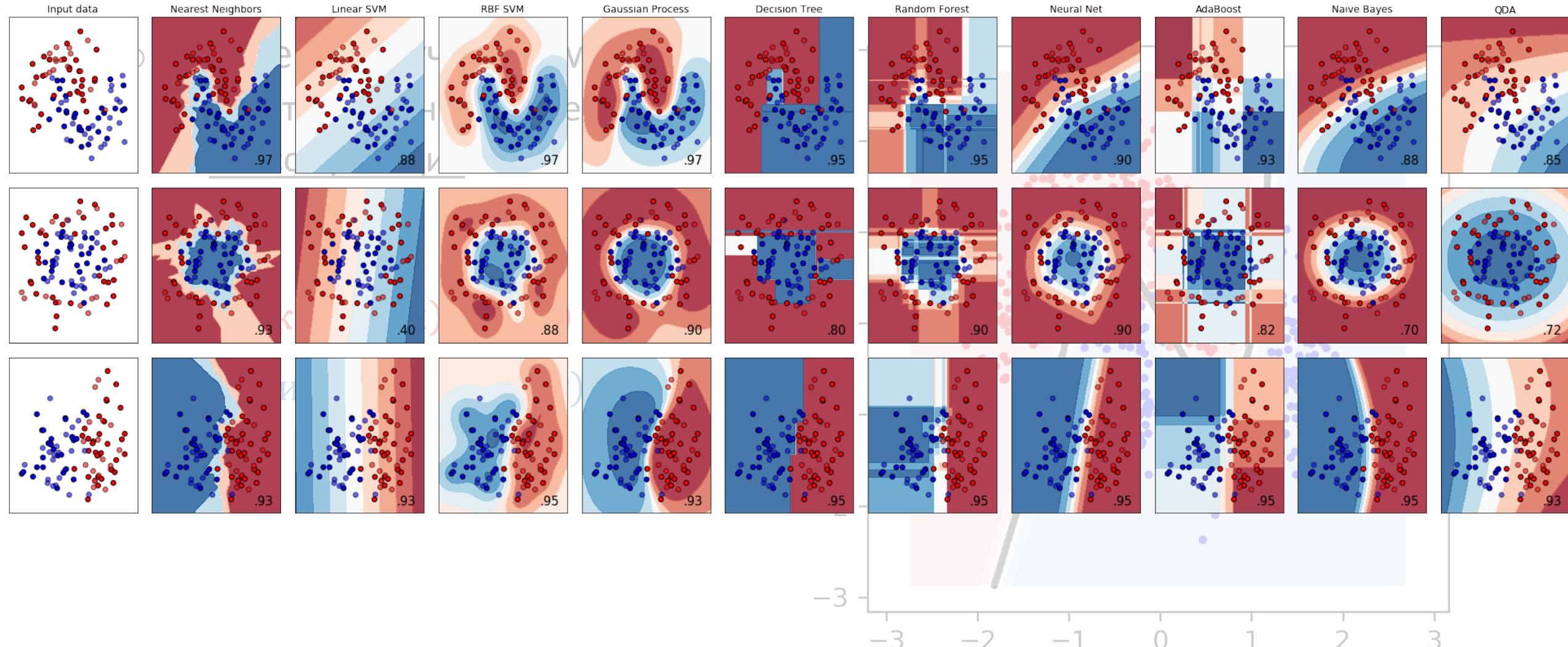
$$\hat{p}(\text{красный}) = f(x)$$

$$\hat{p}(\text{синий}) = 1 - f(x)$$



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Сравнение моделей классификации*



* https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

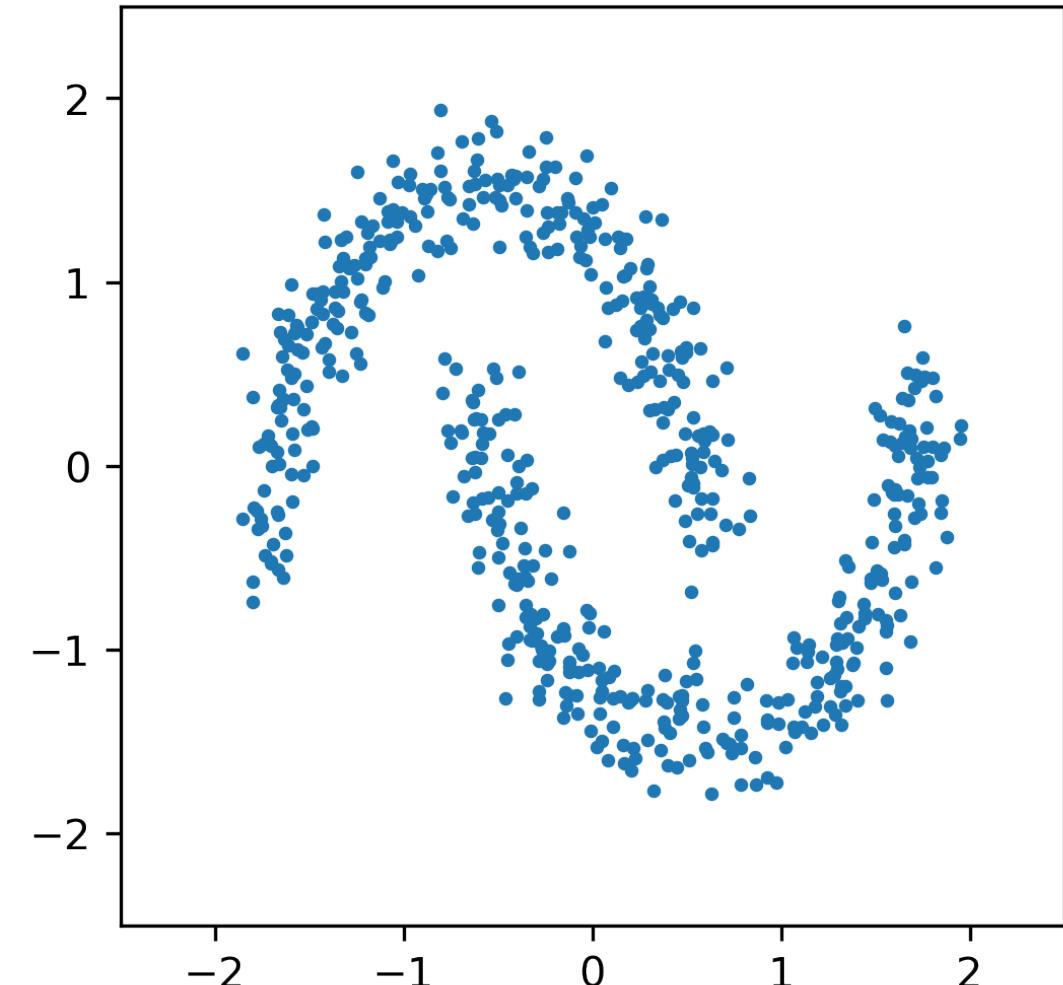
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - поиск структуры в данных

что я хочу?

- метки групп
- знать, есть ли группы?
- сколько групп?



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

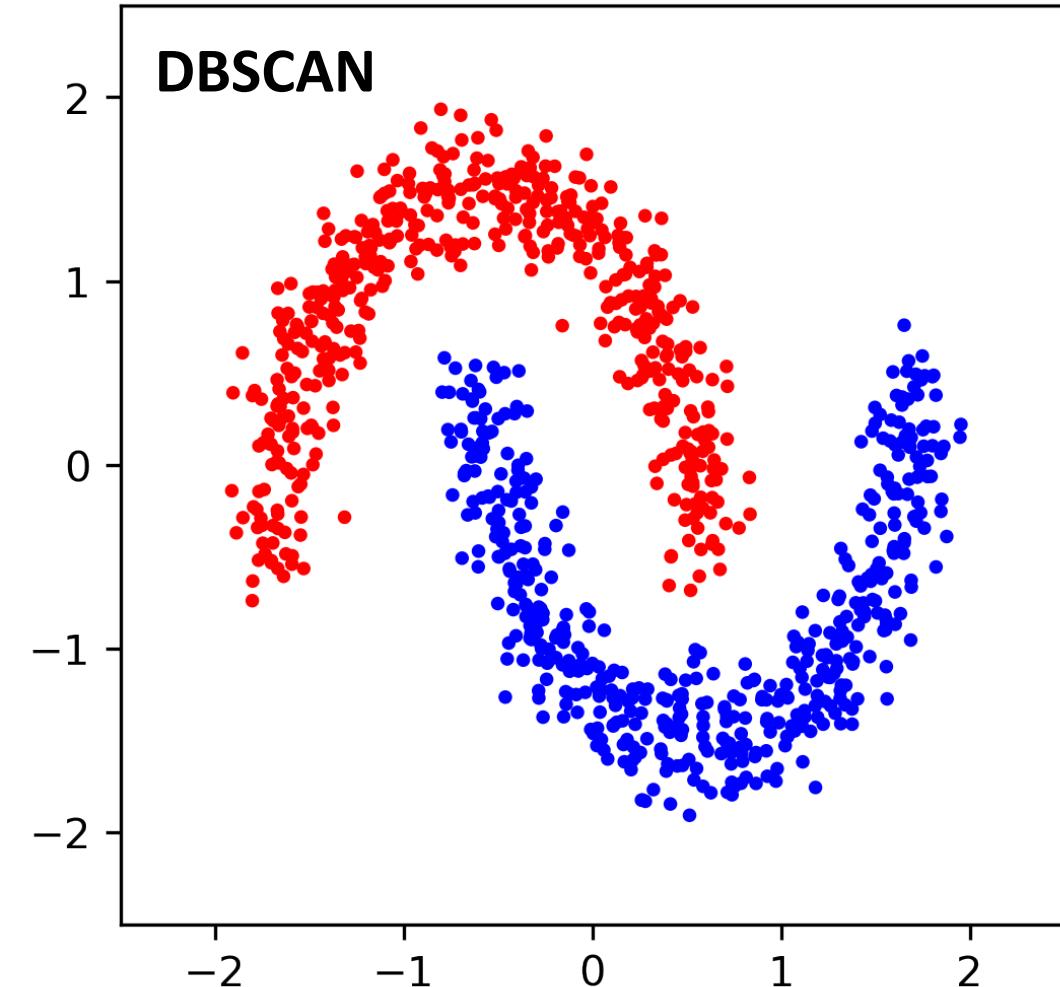
ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - кластеризация

что я хочу?

- метки групп
- знать, есть ли группы?
- сколько групп?



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

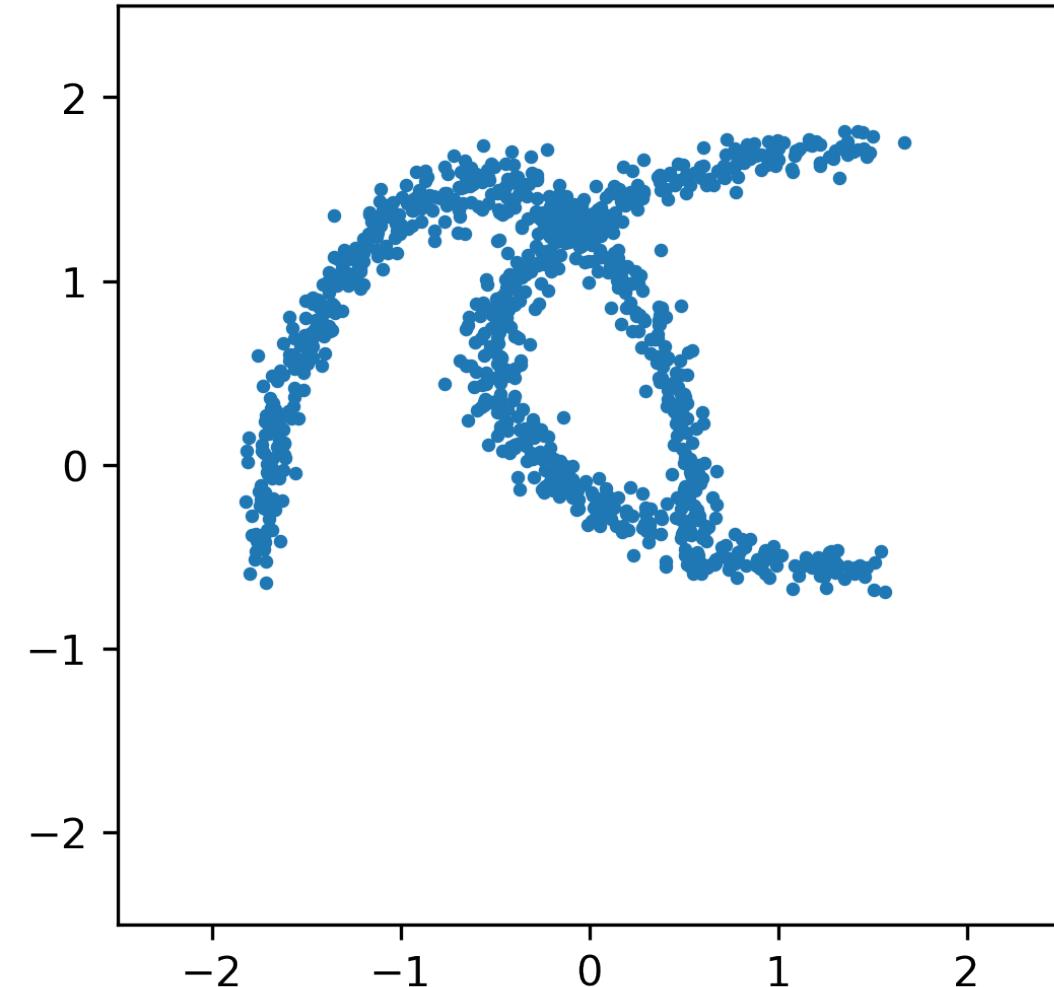
типы задач:

- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - кластеризация

Всегда ли есть решение?

хоть какое-нибудь

ДА



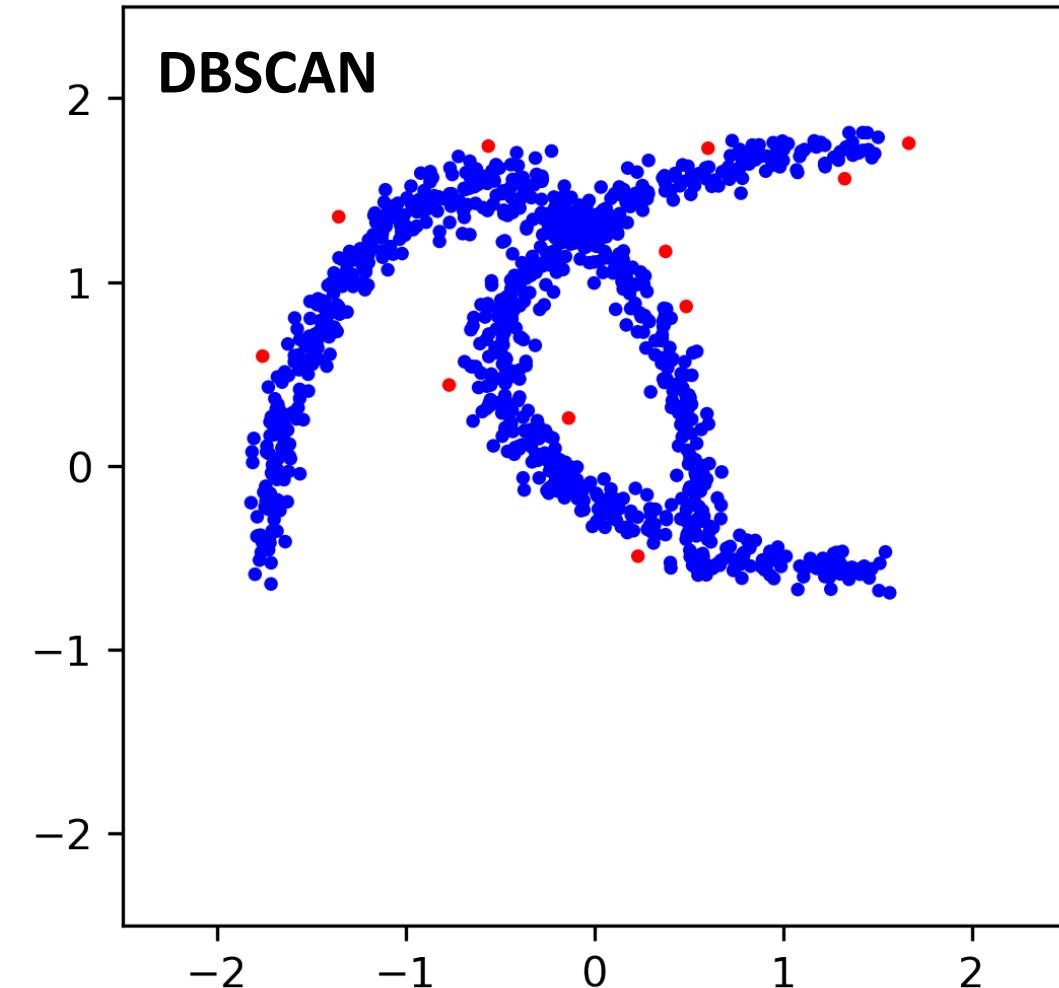
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - кластеризация

Всегда ли есть решение,
которое мне понравится?



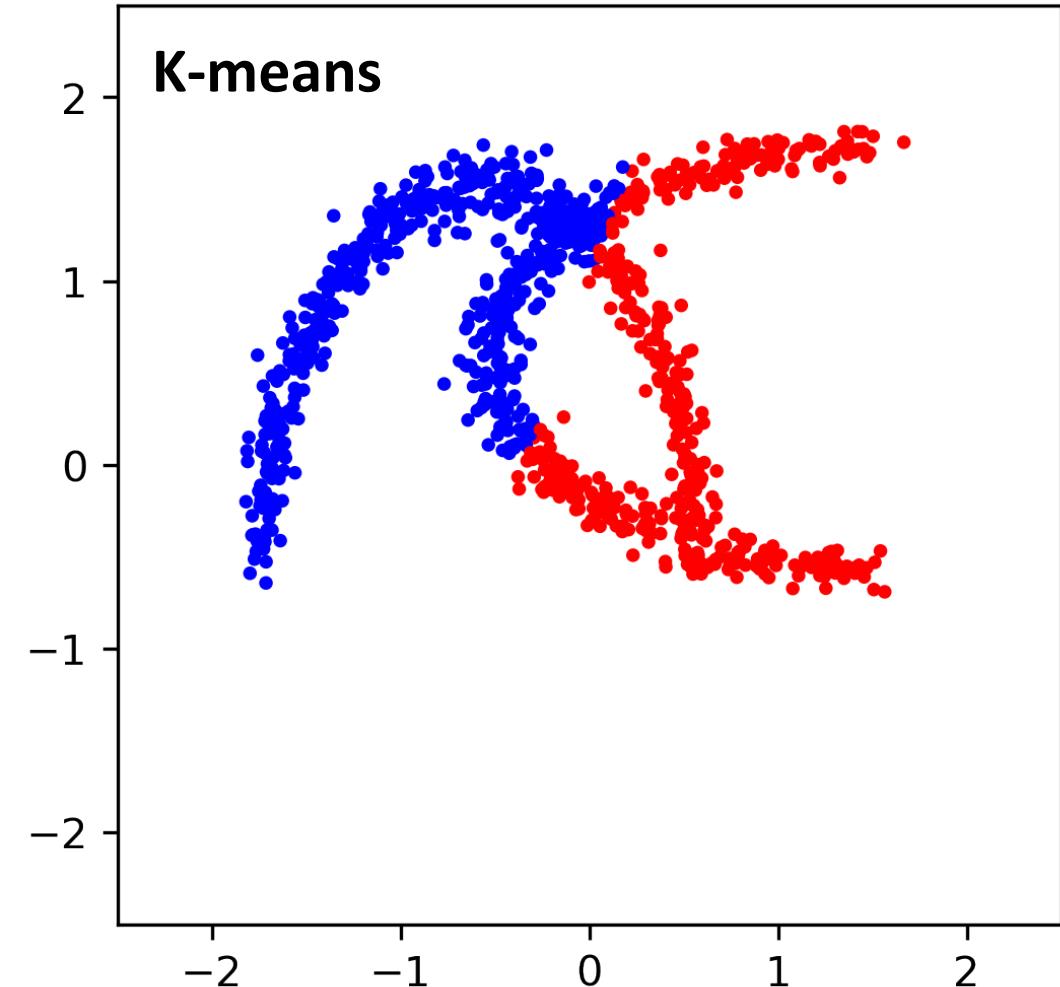
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

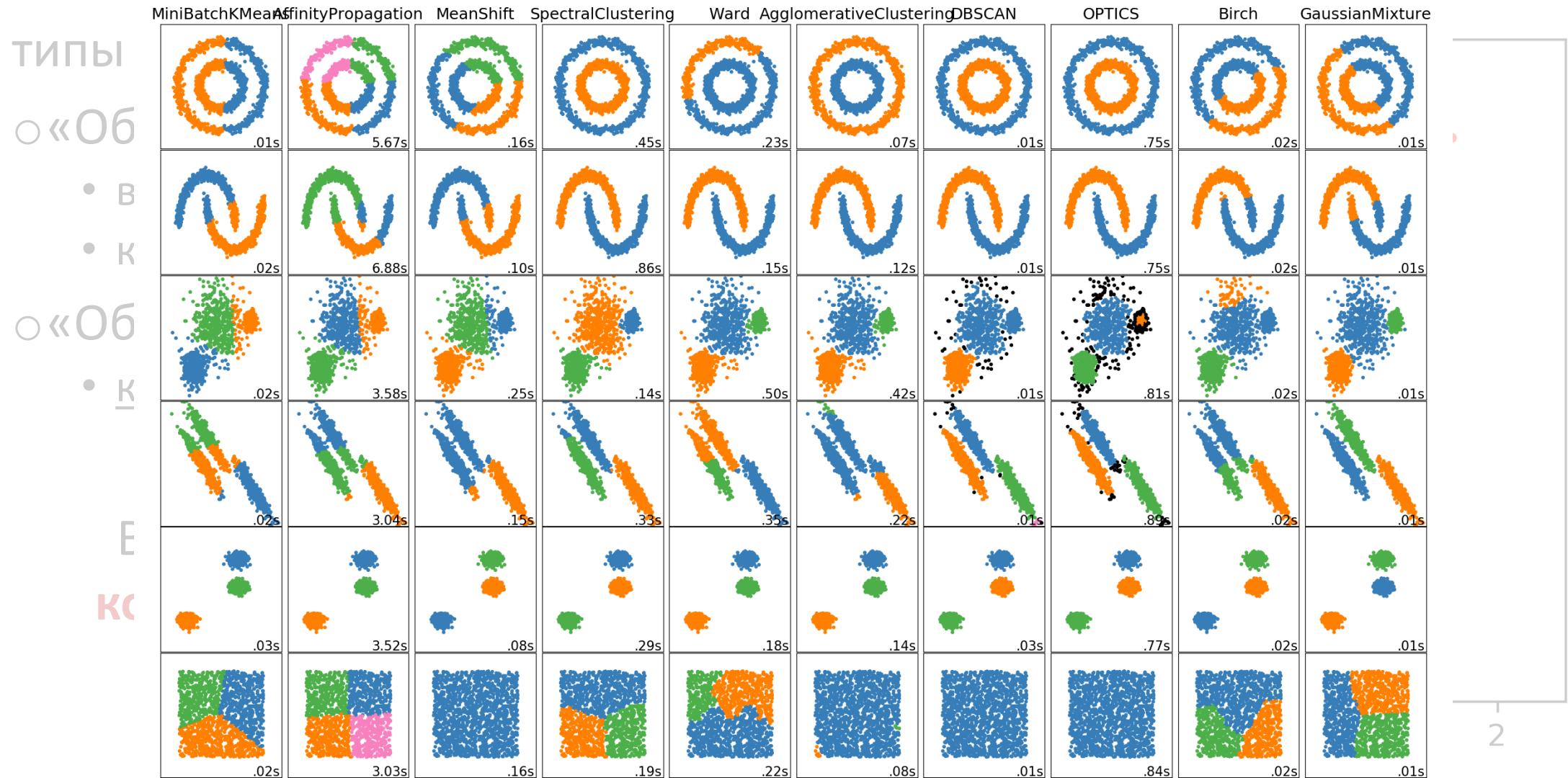
- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - кластеризация

Всегда ли есть решение,
которое мне понравится?



КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Сравнение моделей кластеризации*



Что можно сказать
про эти модели кластеризации
и их результаты?

ремарки

Кластеризация:

- Количество данных часто, но не всегда играет роль
- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных, от наличия структуры в данных
- Разные модели дают разный результат, но нет «более правильного» результата. Есть «более подходящий» для целей конкретного исследования.

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ЦЕЛЬ: **сформулировать задачу** (в терминах машинного обучения)

типы задач:

○ «Обучение с учителем»

- восстановление регрессии
- классификация

○ «Обучение без учителя»

- кластеризация
- понижение размерности

ЧТО Я ХОЧУ?

- новое представление (признаковое описание) данных в пространстве меньшей размерности
- цели:
- Визуализация на плоскости, в 3D
 - Борьба с переобучением (в контексте т.н. «проклятия размерности»)
 - Сжатие данных с минимальными потерями
 - Сокращение вычислительных затрат при обработке данных
 - Извлечение значимых признаков, feature engineering

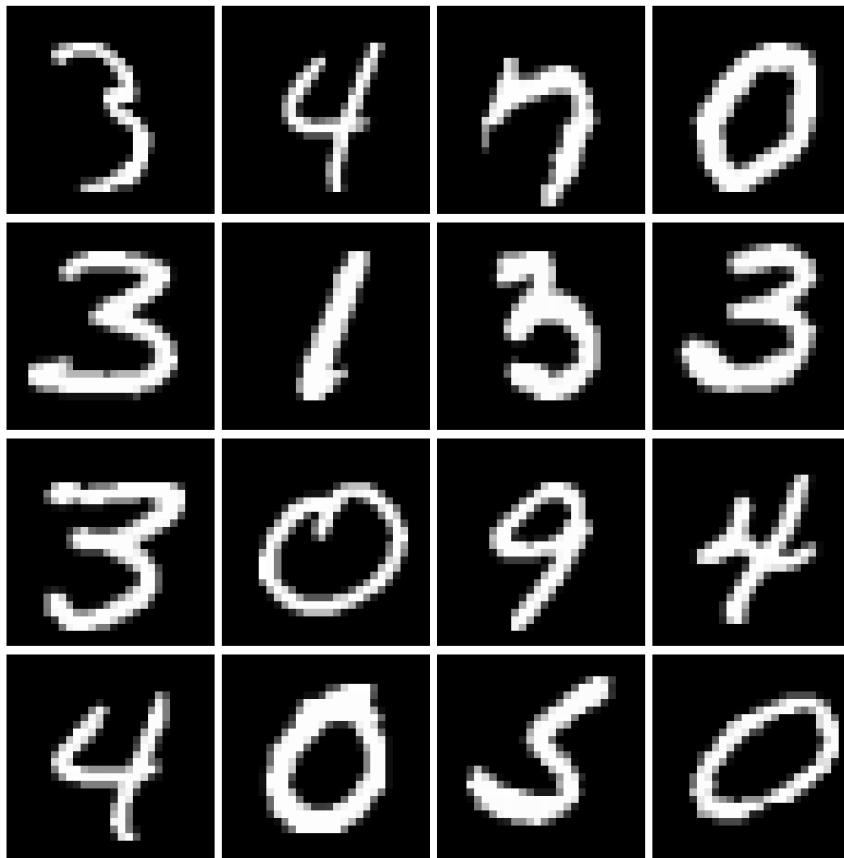
Пожелания:

- Сохранение структуры данных
- Сохранение отношений близости между объектами (событиями)
- Возможность визуализации
- Интерпретируемость новых признаков

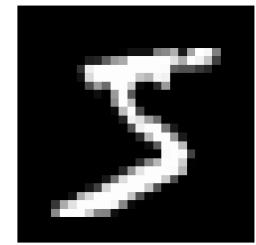
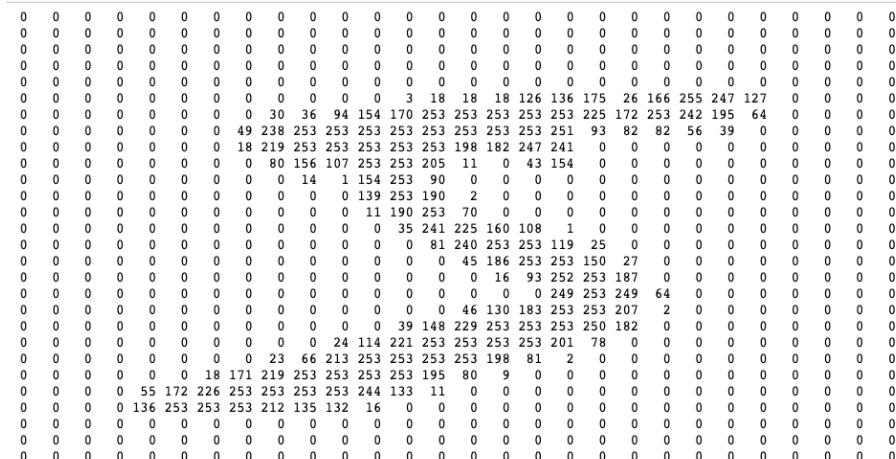
КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Задача понижения размерности: пример

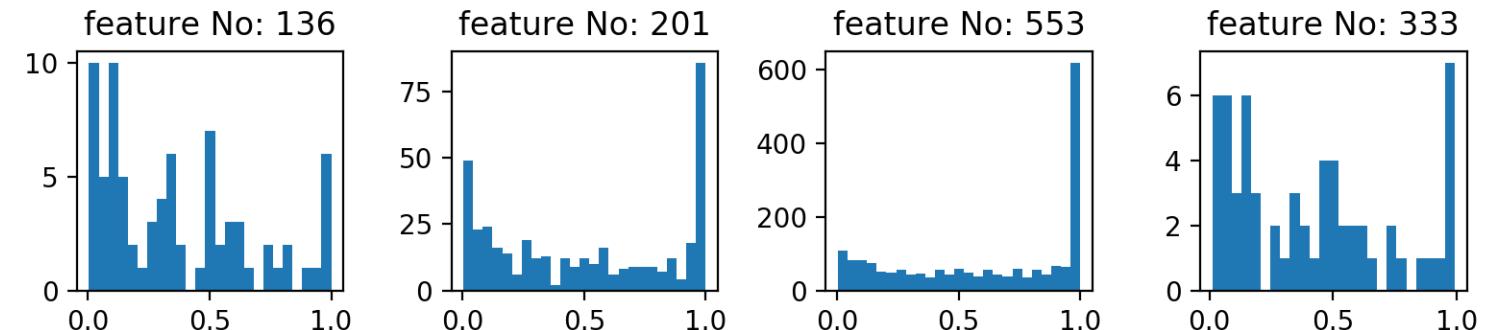
MNIST dataset*



example #0 (label = "5")



data distribution (4 of 784 features)



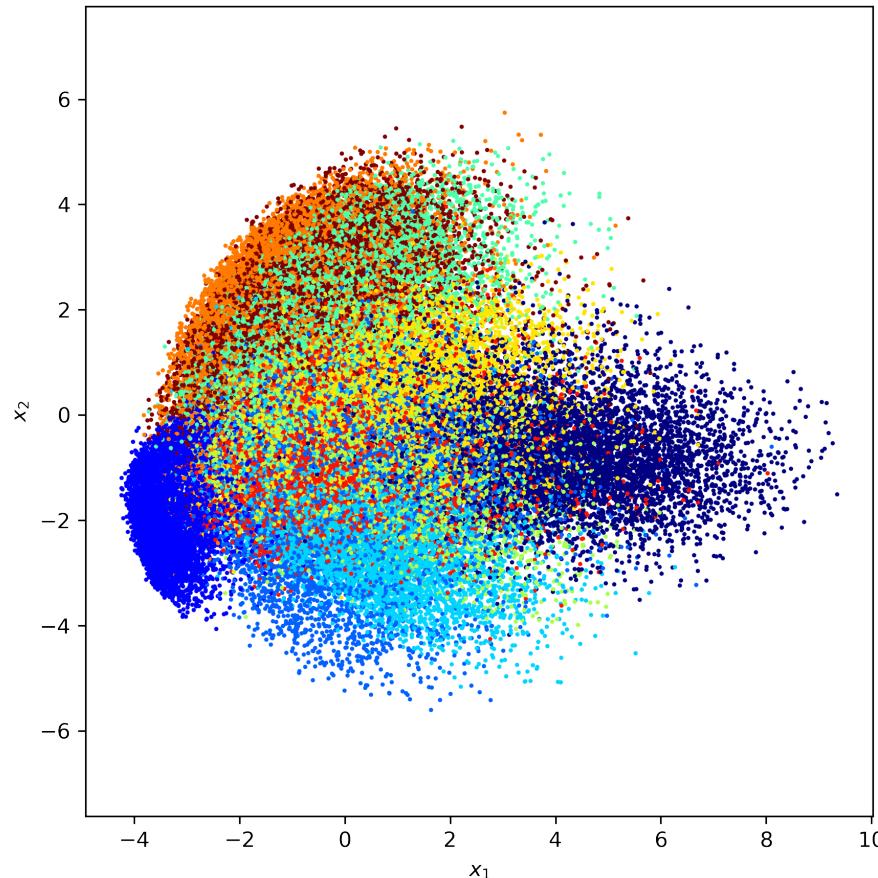
* <http://yann.lecun.com/exdb/mnist/>

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Задача понижения размерности: пример

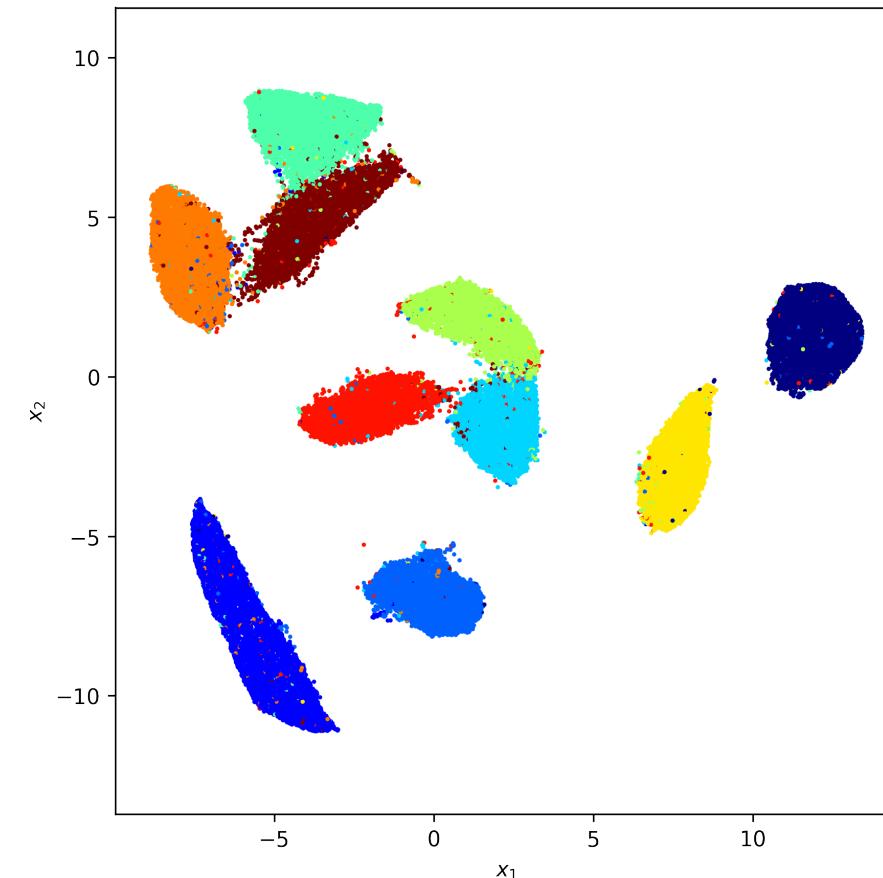
PCA

Principal components analysis
Метод главных компонент



UMAP

Uniform Manifold Approximation and Projection



Что можно сказать
про эти модели кластеризации
и их результаты?

ремарки

Понижение размерности:

- Количество данных всегда играет роль
- Разные модели ведут себя по-разному в зависимости от шума в данных, от количества данных, от наличия выбросов в данных, от наличия структуры в данных
- Разные модели дают разный результат, нет «более правильного» результата. Но есть «более подходящий» для целей конкретного исследования.
- Модели различаются по интерпретируемости (e.g. PCA vs. UMAP)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем»

- восстановление регрессии
- классификация

- «Обучение без учителя»

- кластеризация
- понижение размерности
- восстановление
распределения данных

ЧТО Я ХОЧУ?

- Получить модель, генерирующую примеры, распределение которых совпадает с распределением обучающих данных

Цели:

- дополнение данных
- заполнение пропусков в данных

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем»

- восстановление регрессии
- классификация

- «Обучение без учителя»

- кластеризация
- понижение размерности
- восстановление
распределения данных

Примеры:

- DeepFake - video
- SuperResolution - images
- Text-to-speech - audio (Siri, Алиса, Cortana, Alexa etc.)

КЛАССИФИКАЦИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем» - **Supervised learning**
 - восстановление регрессии
 - классификация
- «Обучение без учителя» - **Unsupervised learning**
 - кластеризация
 - понижение размерности
 - восстановление распределения данных

другие (реже используемые в ES) типы

- «Обучение с частичным привлечением учителя»
 - **Weakly supervised learning**
- «Обучение с подкреплением»
 - **Reinforcement learning**

ДЗ №1

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

обучаем (тренируем) модель на имеющихся данных

