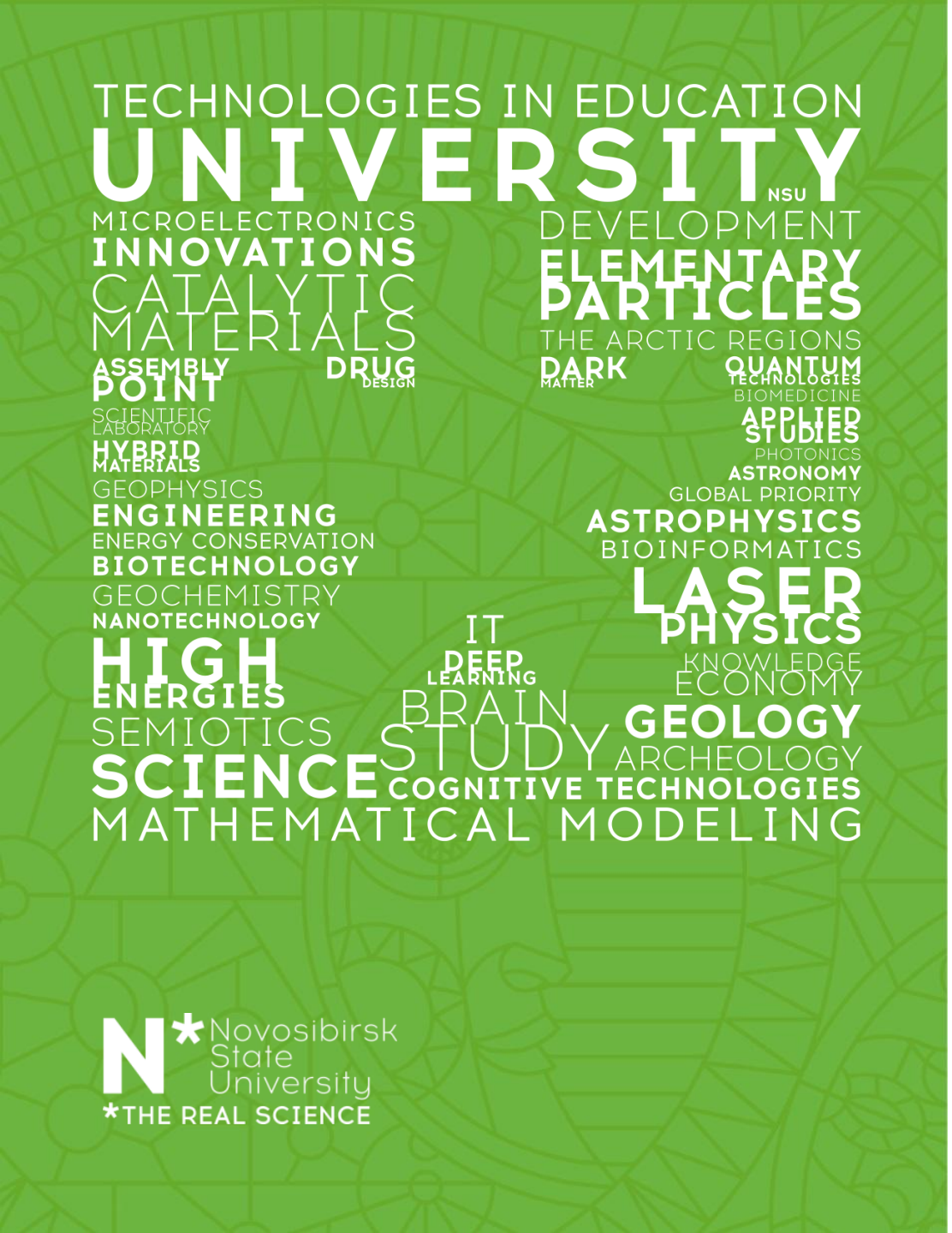


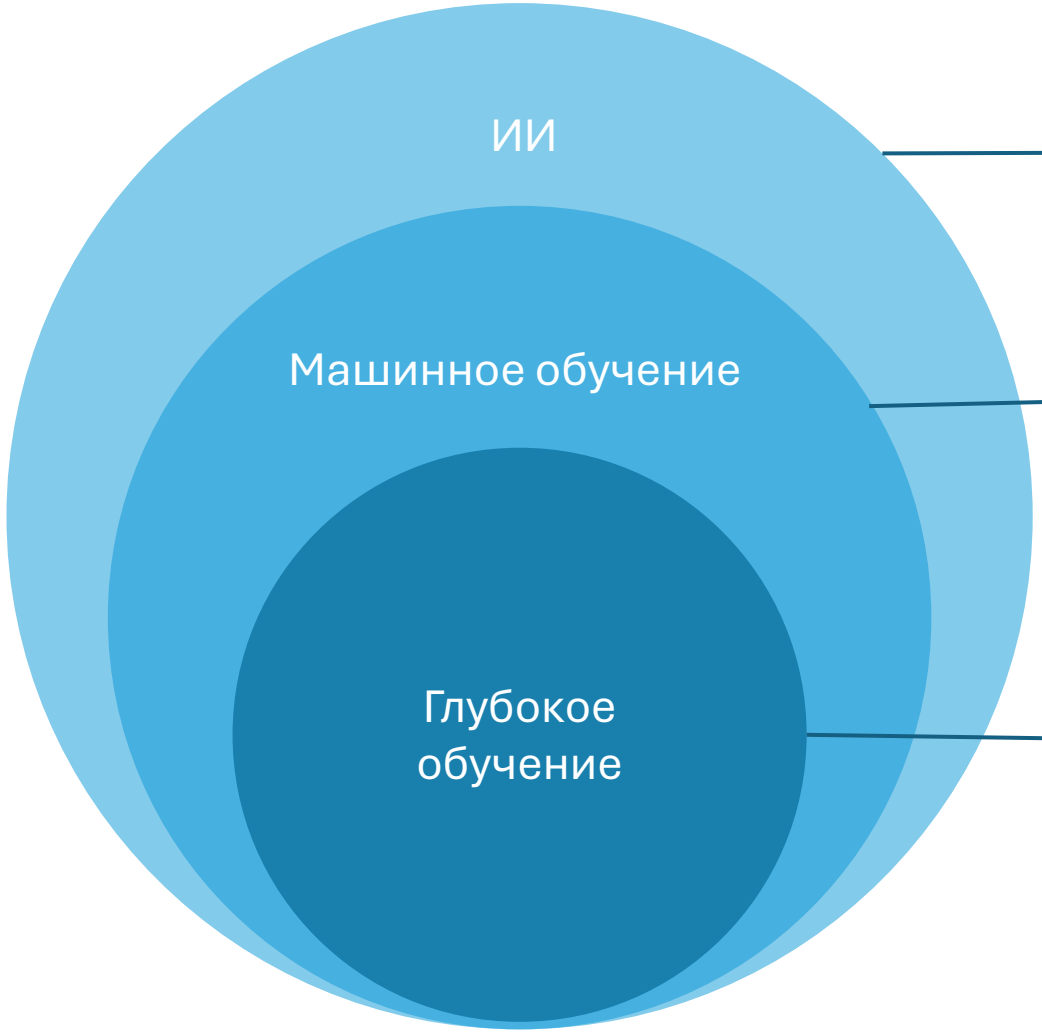
Машинное обучение
Семинар 1

Глушенко Андрей Валерьевич
Институт Интеллектуальной Робототехники

Обзор курса

- Github: https://github.com/andreymarlin/ml-iir
- Email: a.glushenko@g.nsu.ru
- Telegram: @gsswgg
- Форма аттестации – диф. зачёт по следующим критериям:
  - ❑Посещаемость
  - ❑Выполнение лабораторных работ
  - ❑Выполнение кейса в команде (3 чел.)

Что такое машинное обучение?

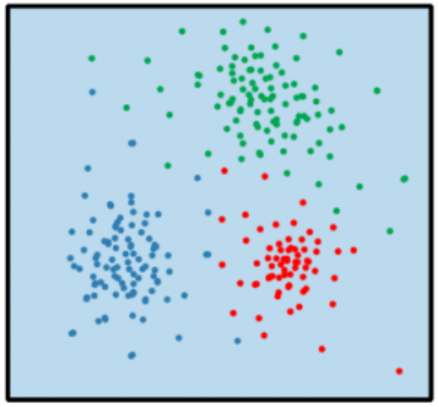
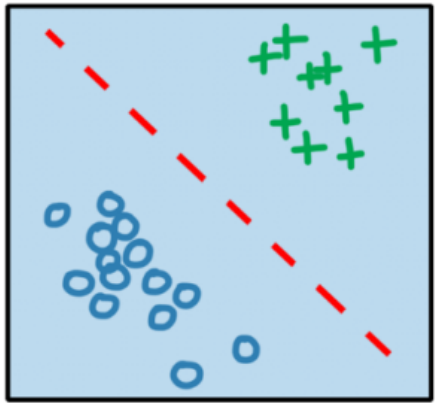
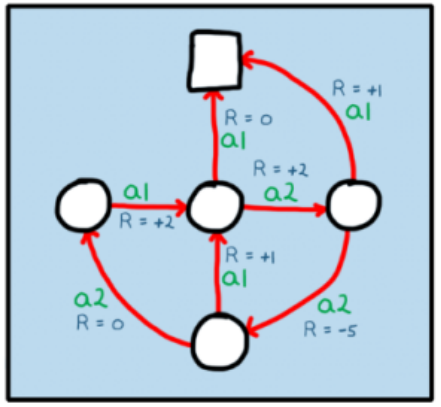


Широкая область создания систем, имитирующих человеческий интеллект для автономного выполнения задач.

Область искусственного интеллекта, позволяющая компьютерам извлекать закономерности из данных и совершенствоваться со временем без явного программирования каждой задачи.

Специализированная область машинного обучения, использующая многослойные нейронные сети для моделирования сложных паттернов.

Типы машинного обучения

Обучение без учителя (Unsupervised learning)	Обучение с учителем (Supervised learning)	Обучение с подкреплением (Reinforcement learning)
		
Неразмеченные данные Нет обратной связи Найти скрытую структуру	Размеченные данные Предсказание следующего значения	Принятие оптимальных решений Взаимодействие со средой Вознаграждения (или наказания) за результат

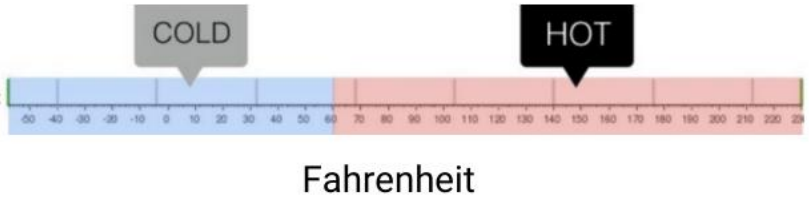
Source: https://www.theaiops.com/machine-learning-compare-supervised-learning-vs-unsupervised-learning-vs-reinforcement-learning/

Задачи обучения с учителем

Регрессия (Regression)

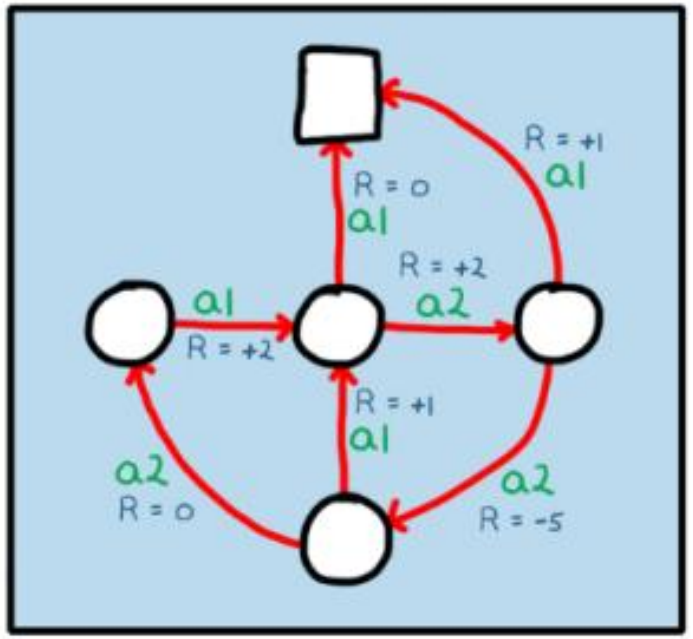
Классификация (Classification)

Бинарная/
Многоклассовая



Source: https://www.enjoyalgorithms.com/blogs/classification-and-regression-in-machine-learning

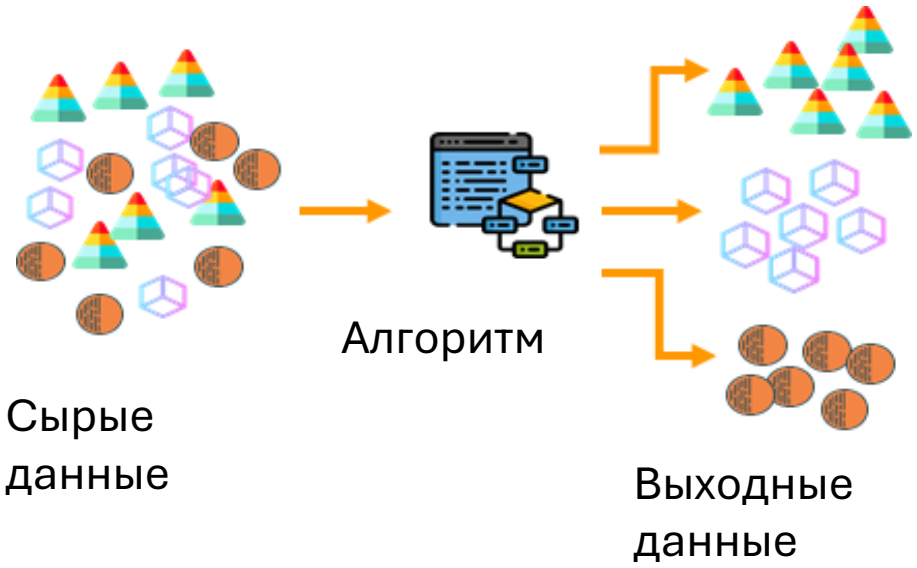
Обучение с подкреплением (Reinforcement learning)



Принятие оптимальных решений
Взаимодействие со средой
Вознаграждения (или наказания) за результат

Source: https://www.theaiops.com/machine-learning-compare-supervised-learning-vs-unsupervised-learning-vs-reinforcement-learning/

Задачи обучения без учителя



Source: https://medium.com/analytics-vidhya/beginners-guide-to-unsupervised-learning-76a575c4e942

- Кластеризация
- Понижение размерности

Любой крупный бизнес сталкивается с потоком обращений клиентов. Ручная сортировка («это в бухгалтерию», «это к сисадминам») занимает время и стоит дорого. Автоматизация этого процесса — классическая задача в индустрии

Разработать веб-сервис, который позволяет пользователю ввести параметры недвижимости (площадь, район, этаж и т.д.) и получить прогноз стоимости, основанный на обученной модели машинного обучения.

В любой компании документооборот занимает огромное количество времени. Автоматическое извлечение таких полей, как «Номер договора», «Дата», «Сумма» и «Стороны», позволяет мгновенно каталогизировать архивы и искать нужные документы.

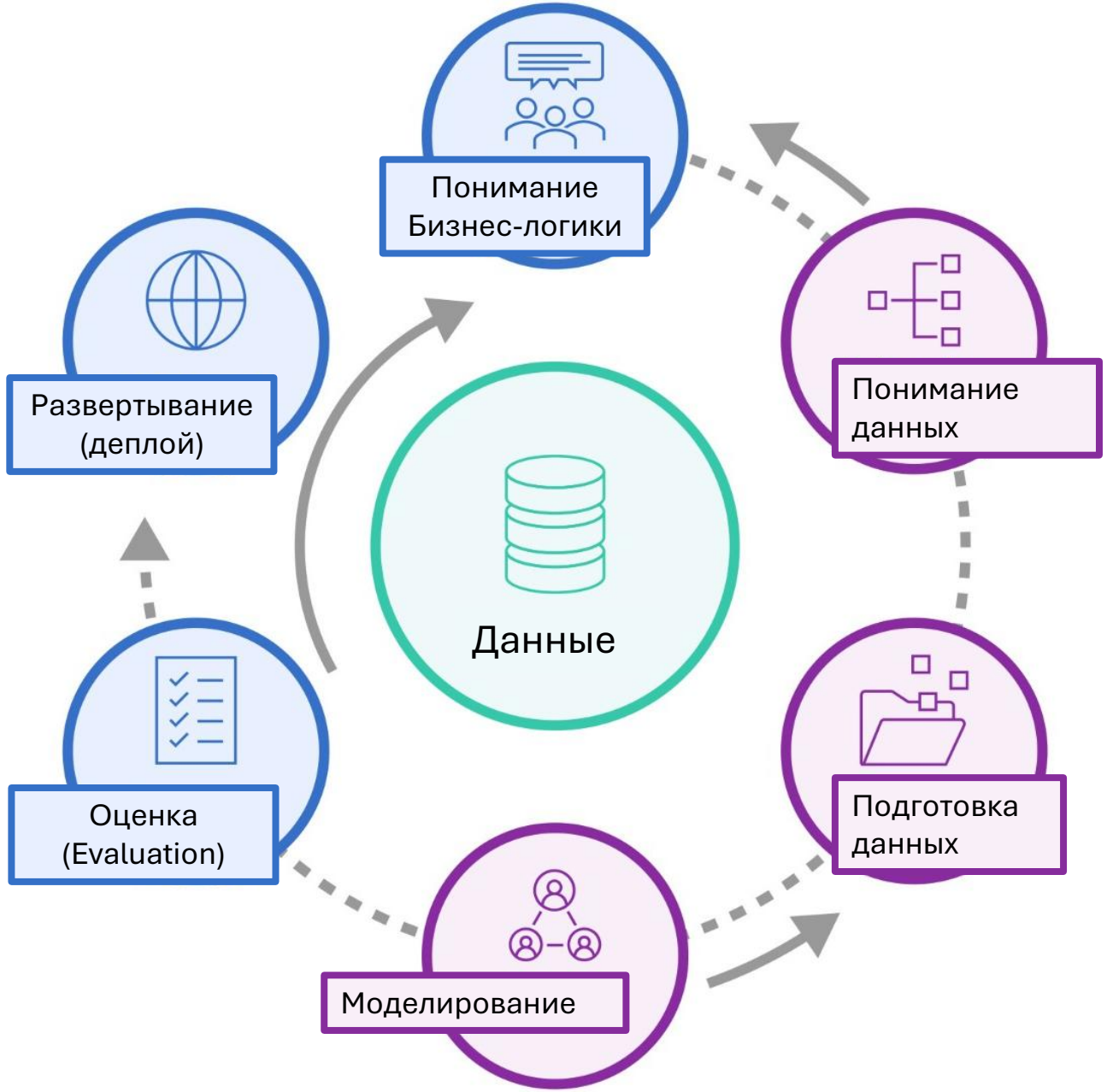
Определите тип задачи (Регрессия, Классификация или Кластеризация) для следующих сценариев:

А) Банк хочет предсказать вероятность (0 или 1), вернет ли клиент кредит.
Б) Риелтор хочет оценить стоимость квартиры в рублях на основе её площади и района.
В) Маркетолог хочет разбить базу клиентов на сегменты по покупательскому поведению, не зная заранее, какие это сегменты.

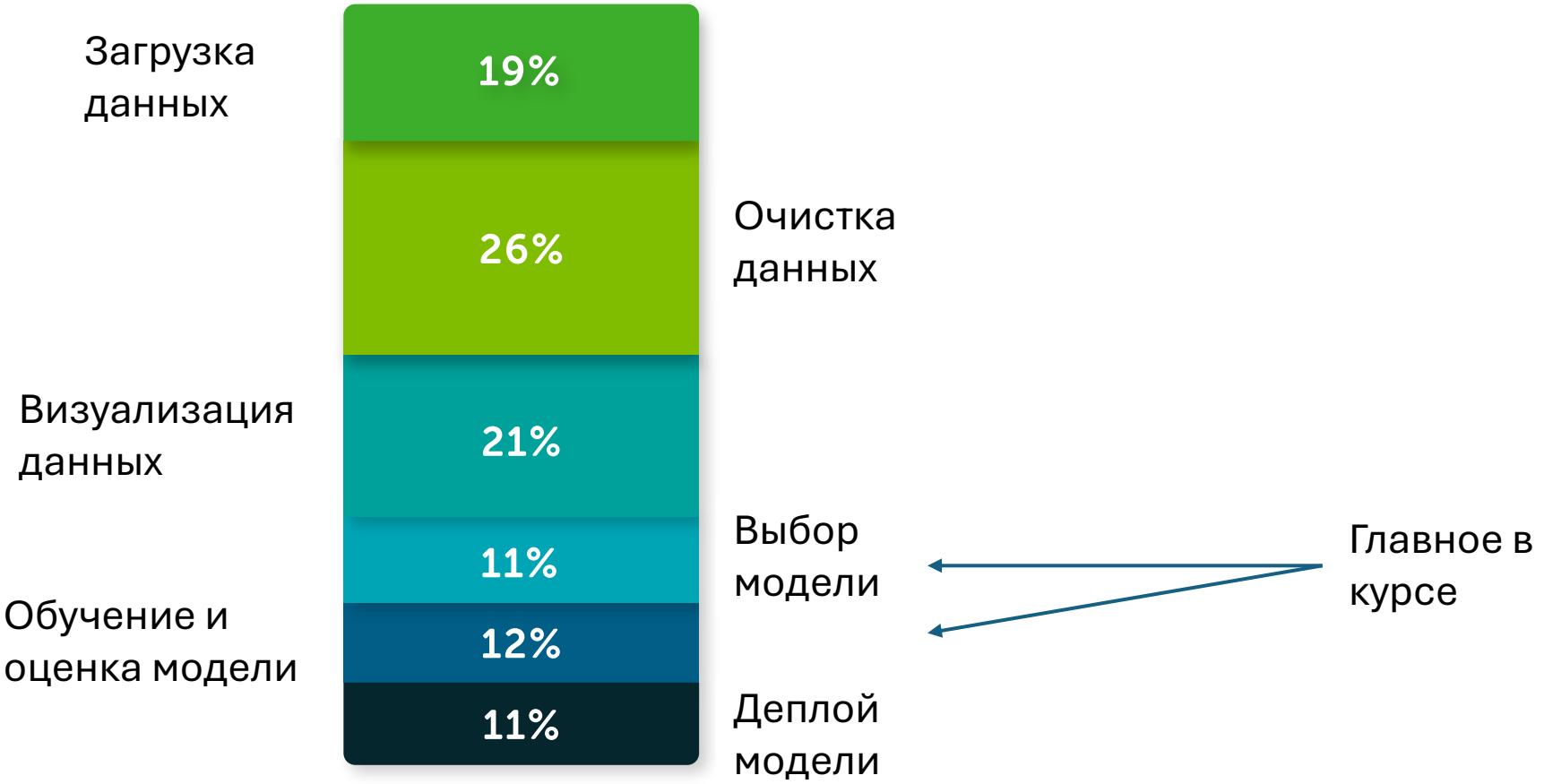
Почему для предсказания погоды (температуры) классическая логистическая регрессия (классификатор) может не подойти, а линейная регрессия подойдет?

Цикл жизни дата-майнинга

Cross-Industry Standard Process for Data Mining (CRISP-DM)



Затраты для выполнения задач ML



Source: https://www.anaconda.com/resources/whitepapers/state-of-data-science-2020

Компоненты пайплайна машинного обучения

- **Входные данные**
- **Признаки**
- **Модель**
- **Предсказание модели**
- **Метрики оценки**

Практические аспекты

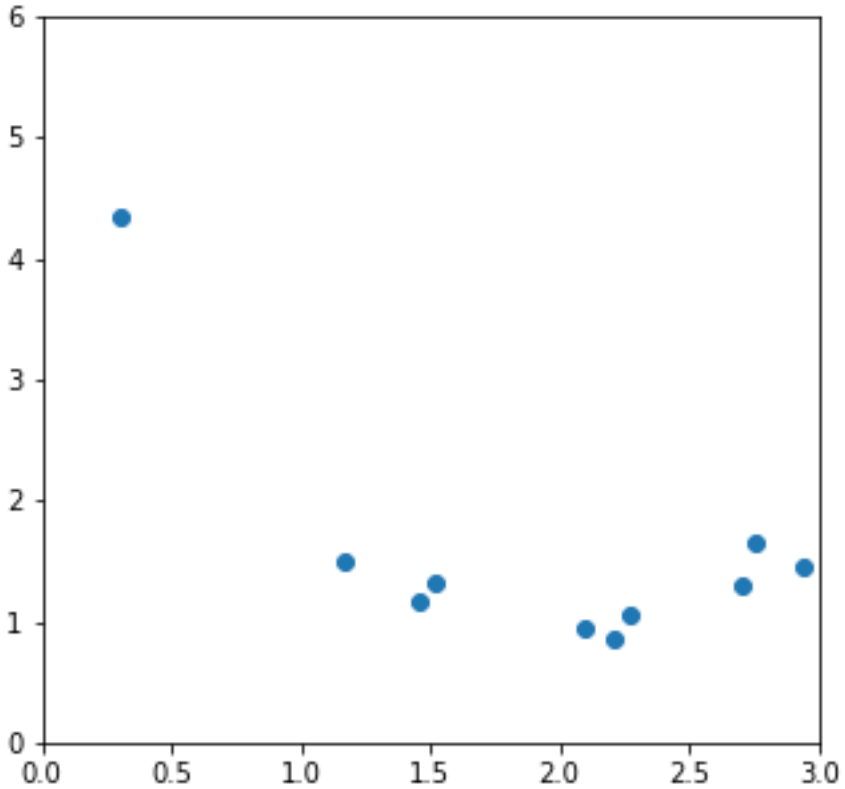
Недообучение и переобучение

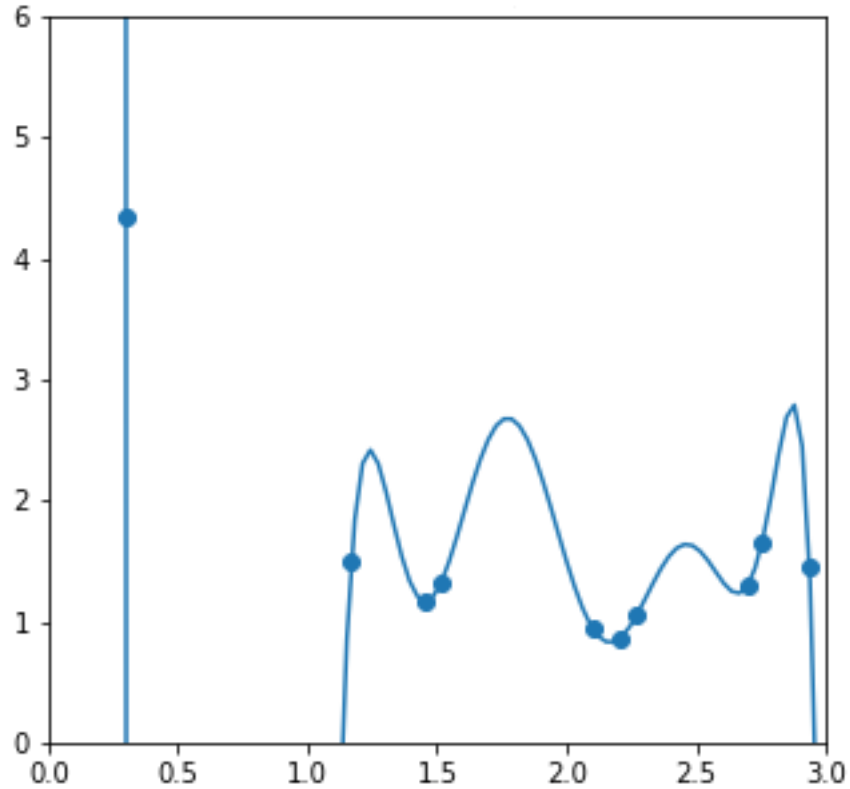
Разделение на обучающую, валидационную и тестовую выборки

Кросс-валидация

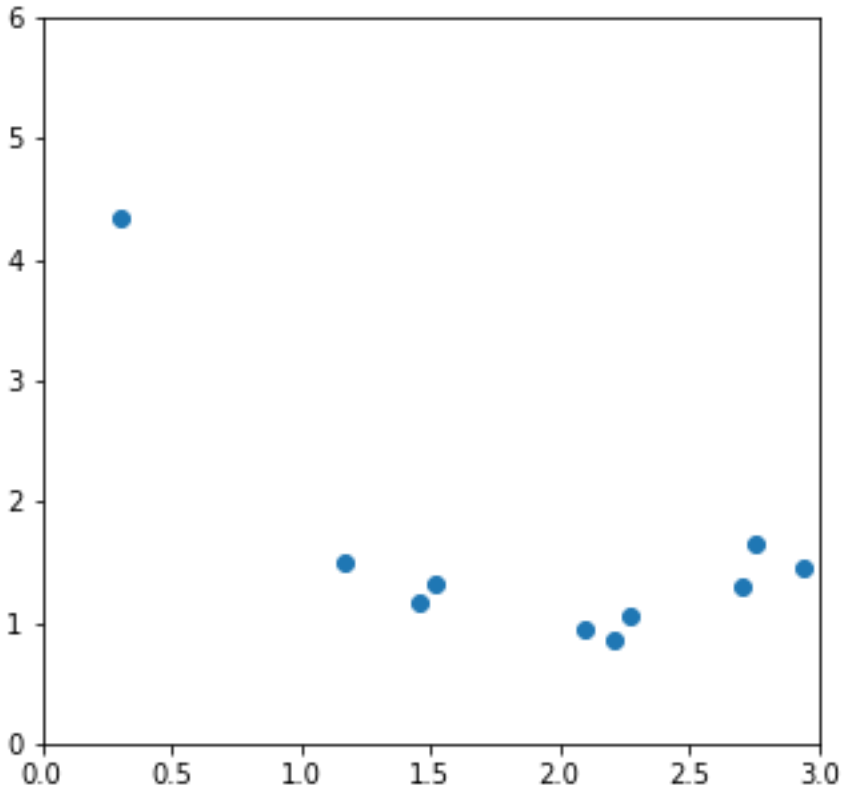
Метрики

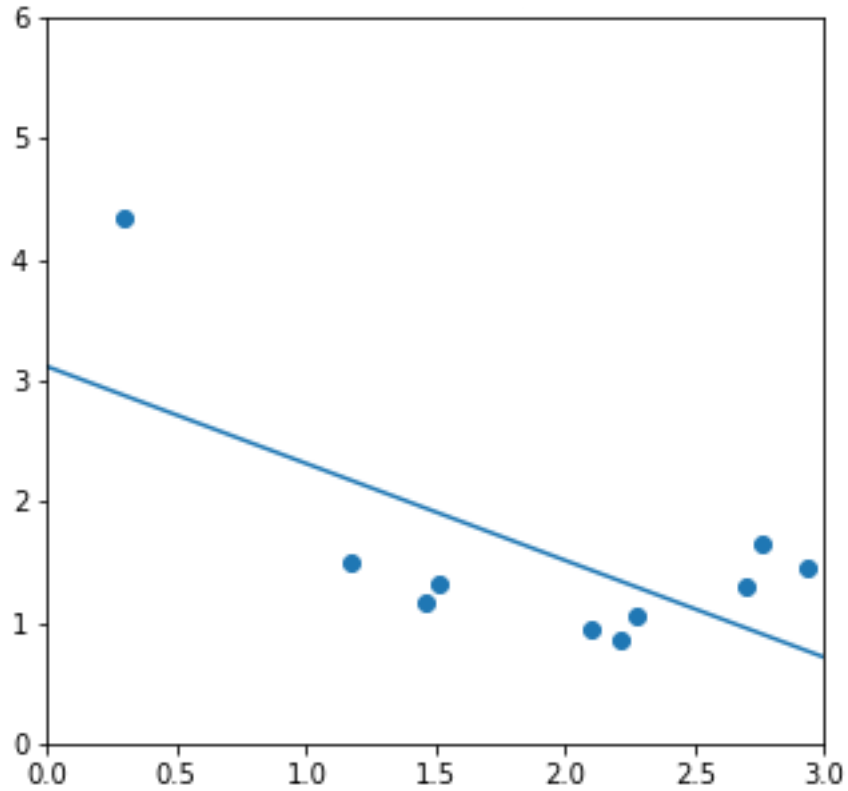
Выбираем класс функции



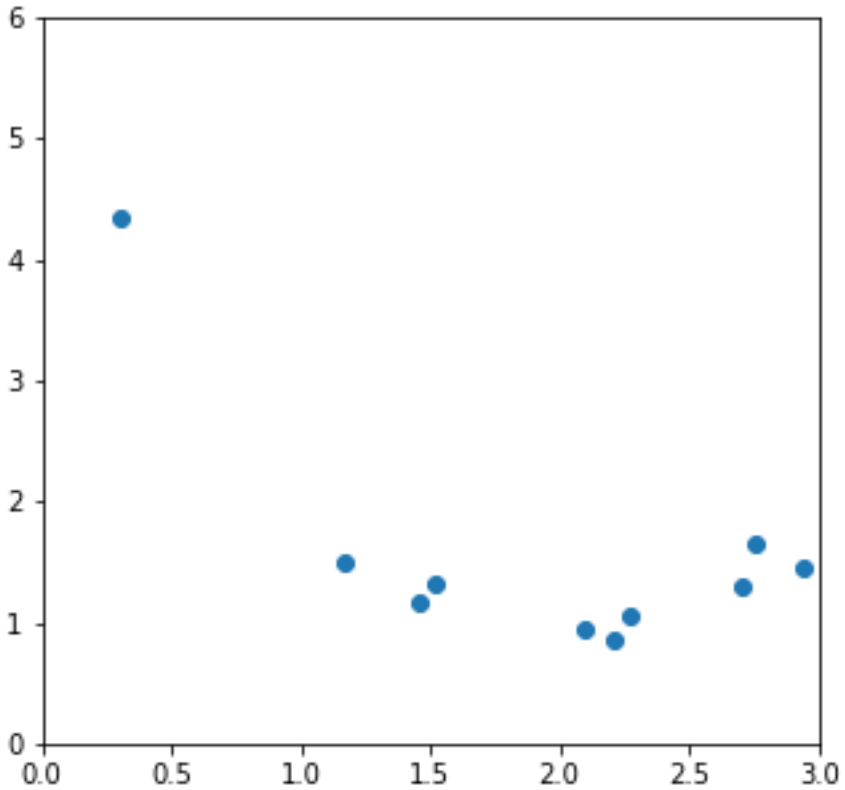


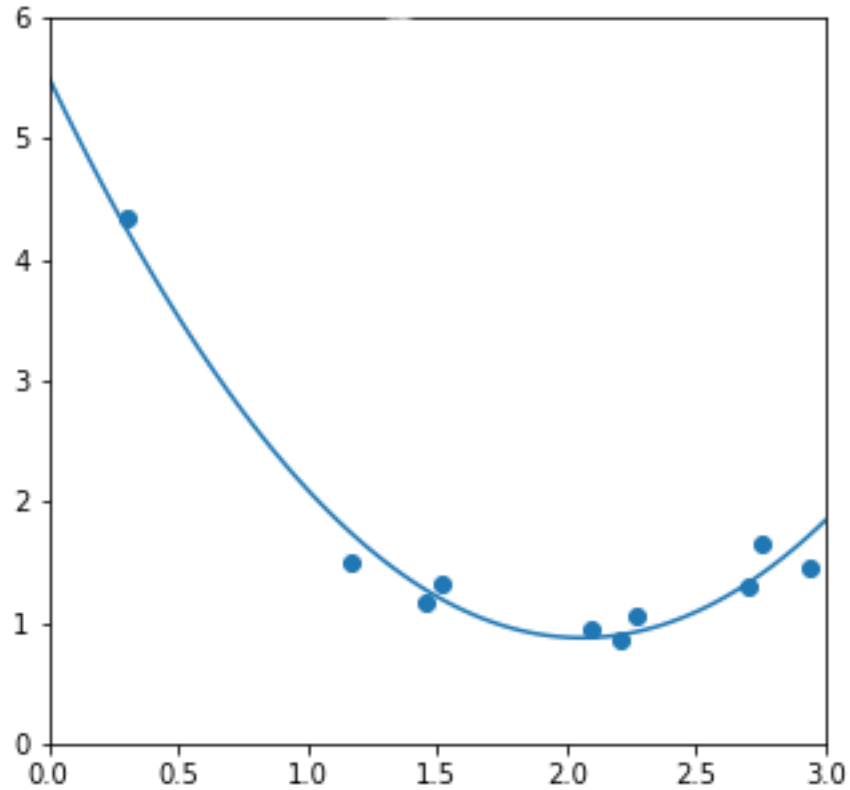
Выбираем класс функции



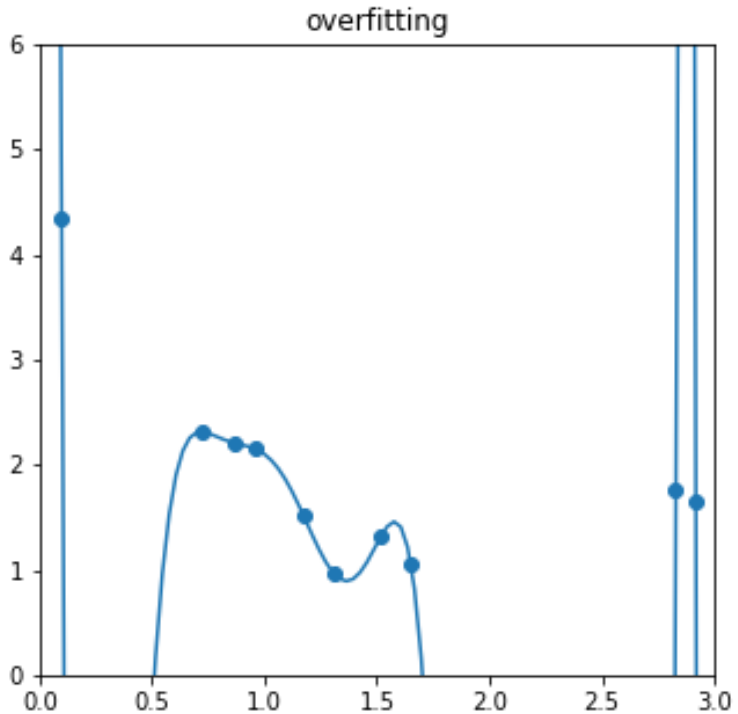


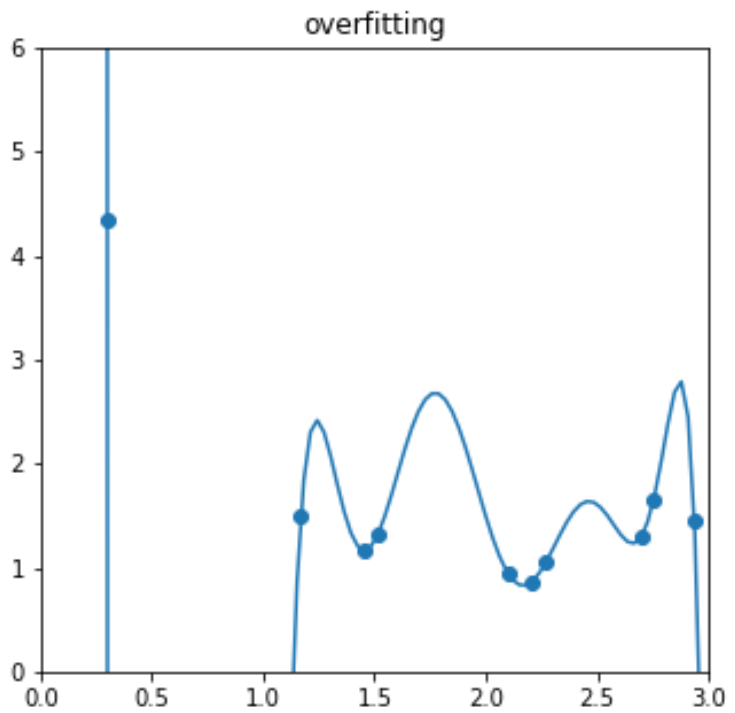
Выбираем класс функции

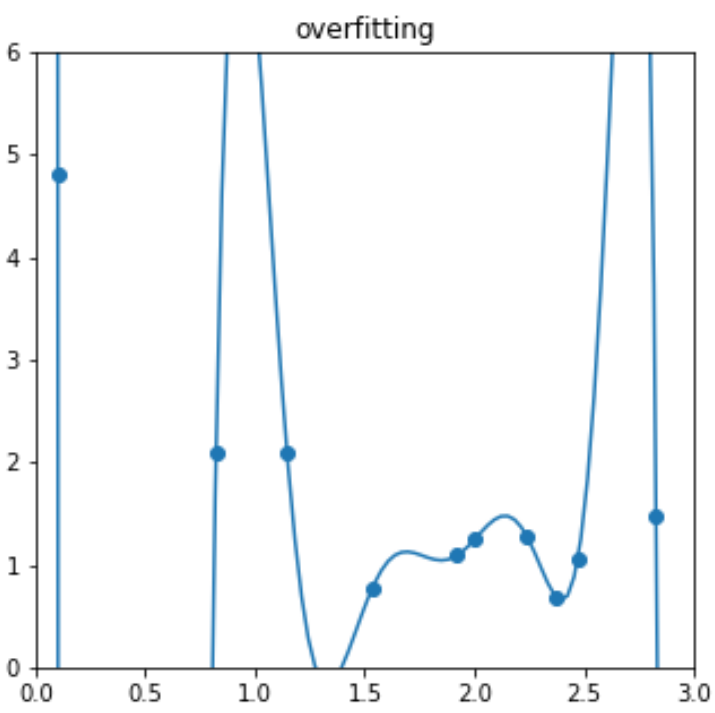




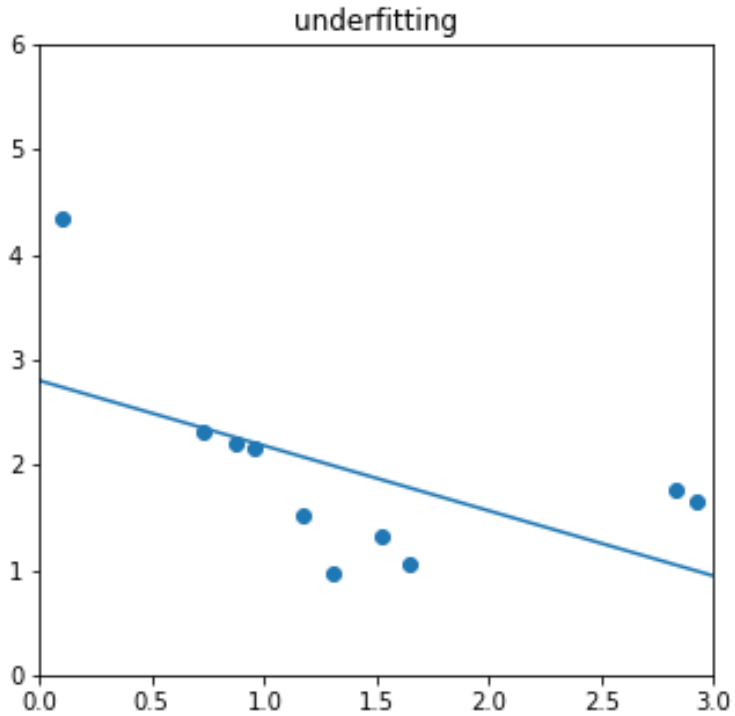
Разброс (Variance), переобучение

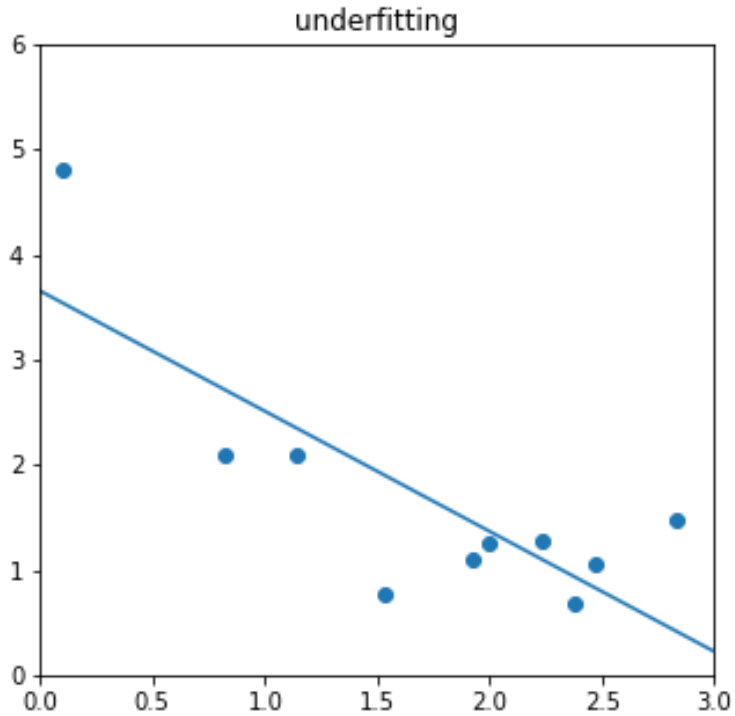


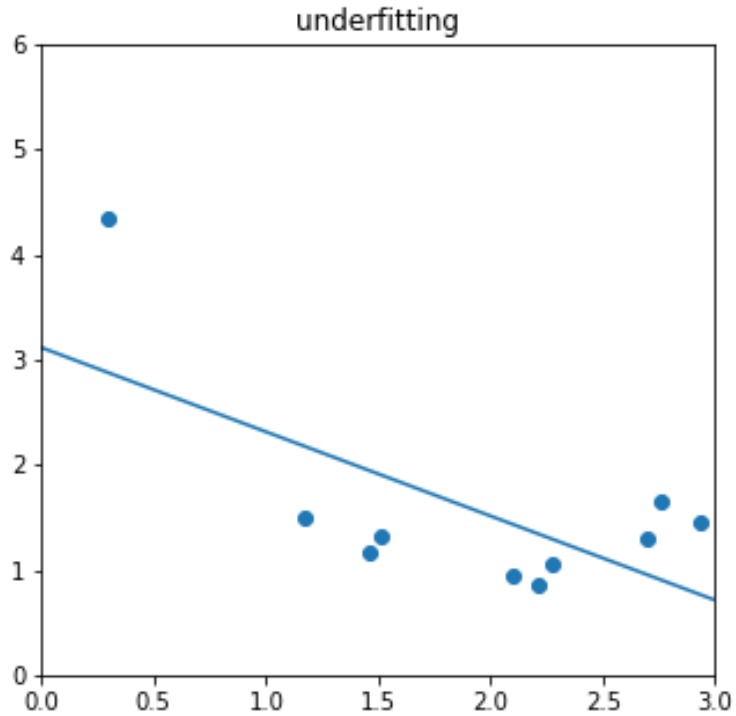




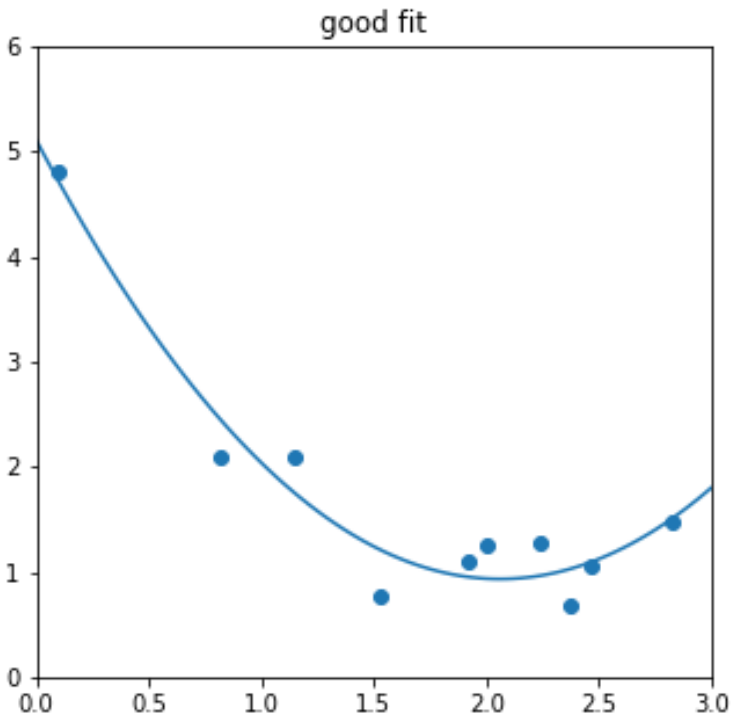
Смещение (Bias), недообучение

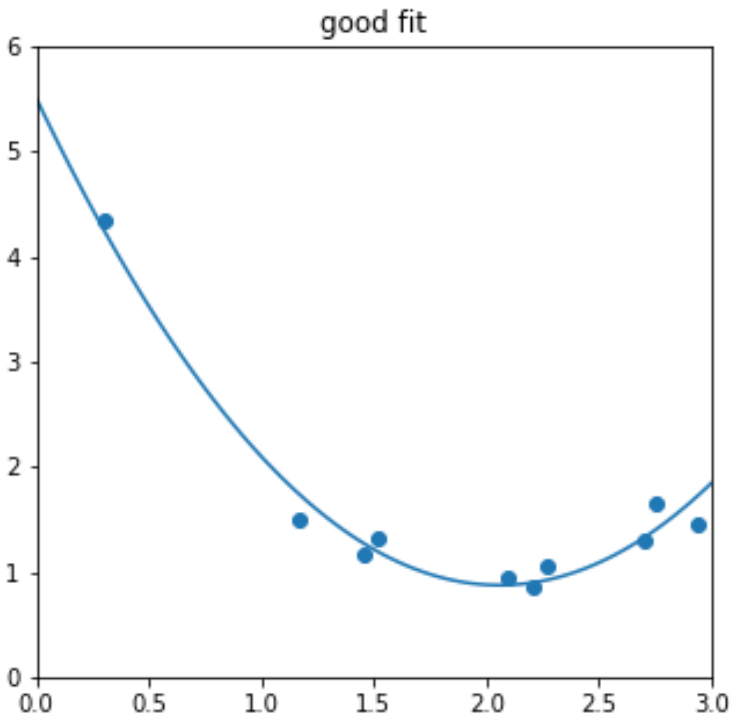


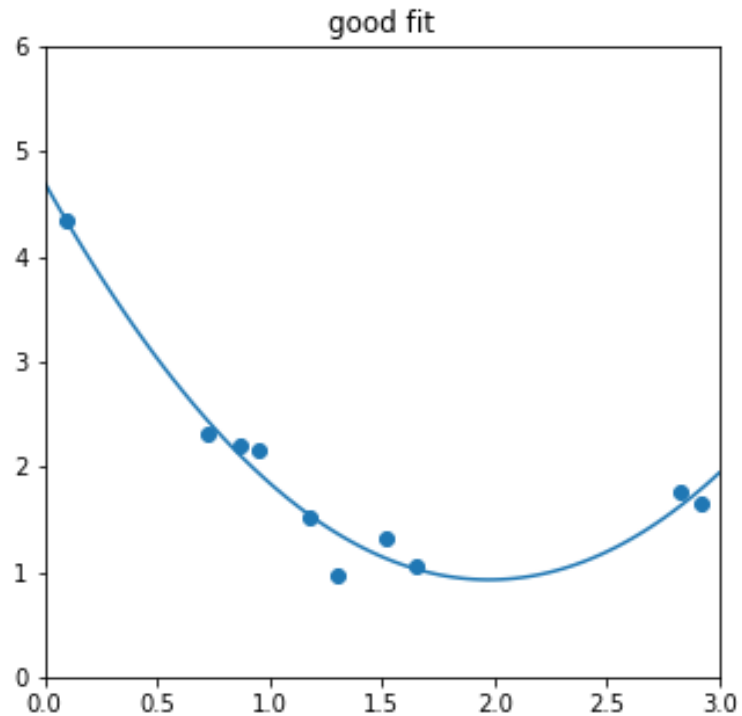




Оптимальные разброс и смещение







Переобучение (overfitting) и недообучение (underfitting)

- Модель с высоким смещением и низким разбросом (дисперсией) является недообученной моделью. Она недостаточно точно отражает статистические взаимосвязи в наших данных.
- Модель с высоким разбросом (дисперсией) и низким смещением является переобученной моделью, поскольку она улавливает взаимосвязи, слишком специфичные для конкретных данных, на которых мы ее обучаем. Эти взаимосвязи могут отсутствовать в общем распределении и, вероятно, являются ложными.

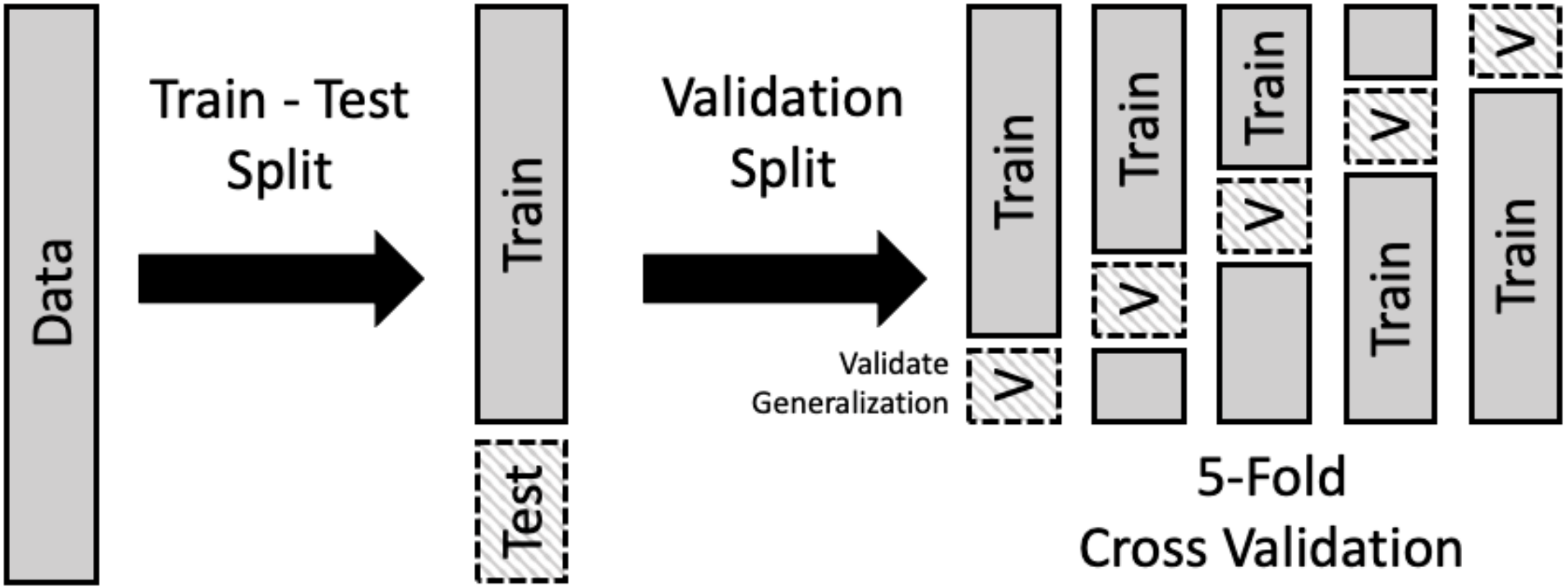
Разделение датасета (Dataset splitting)

- **Тренировочный набор данных**
  - Используется для обучения модели, позволяя ей изучать закономерности и взаимосвязи.
  - Как правило, составляет наибольшую часть данных (например, 60-70%).
- **Валидационный набор данных**
  - Используется для настройки параметров модели и предотвращения переобучения.
  - Помогает оценить производительность модели во время обучения, но не используется для непосредственного обучения модели.
- **Тестовый набор данных**
  - Полностью отдельный набор данных, используемый для окончательной оценки модели.
  - Обеспечивает объективную оценку производительности модели на новых данных.

Кросс валидация

- **Цель:** Подход, позволяющий максимально эффективно использовать данные для обучения и валидации.
- **Процесс:** Набор данных разбивается на k «складок». Каждая складка используется в качестве валидационного набора, а оставшиеся k-1 складки — для обучения.
- **Распространенный подход:** k-складочная перекрестная валидация (например, 5-складочная), где результаты усредняются по всем складкам.
- **Преимущество**: Обеспечивает более надежную оценку, снижая риск переобучения и недообучения.

Кросс-валидация



Source: https://learningds.org/ch/16/ms_cv.html

Данные в машинном обучении

- Данные — основа. Модели полагаются на данные для обучения и прогнозирования.
- Качество имеет значение. Высококачественные данные = точные модели; некачественные данные = неточные, предвзятые результаты.
- Более эффективные решения. Чистые данные приводят к полезным выводам, которые способствуют принятию более эффективных решений.
- Снижает предвзятость. Репрезентативные данные помогают обеспечить справедливые и этичные результаты работы ИИ.
- Вывод: «Хорошие данные = хорошие модели».

Вызовы при работе с данными

1. Проблемы качества и целостности

- **Отсутствующие данные:** Данные часто неполны из-за незарегистрированных значений или ошибок во время сбора.
- **Шумные данные:** Данные со случайными ошибками или шумом, часто требующие методов очистки или шумоподавления.
- **Выбросы:** Экстремальные значения, которые могут искажать модели, особенно в чувствительных алгоритмах.
- **Дублирующиеся данные:** Множественные записи для одной и той же сущности могут исказить модель, придавая чрезмерный вес определенной информации.
- **Неправильные метки:** Неправильная маркировка в тестовых наборах данных может сбить с толку модели и снизить точность.

2. Проблемы распределения данных

- **Несбалансированные данные:** Непропорциональное представление классов, что может смещать модели в сторону преобладающего класса.
- **Асимметричное распределение данных:** Сильно асимметричные признаки могут влиять на точность модели, особенно для алгоритмов, предполагающих нормальное распределение.
- **Перекрытие границ классов:** Плохо определенные границы между классами затрудняют классификацию.
- **Временной дрейф (дрейф концепций):** Статистические свойства данных меняются со временем, что приводит к устареванию моделей.

3. Проблемы с признаками данных

- **Нерелевантные признаки:** Неинформативные признаки добавляют шум в процесс обучения модели.
- **Высокая размерность (проклятие размерности):** Слишком много признаков увеличивает вычислительную сложность и риск переобучения.
- **Взаимодействие признаков:** Сложные взаимосвязи между признаками трудно выявить.
- **Перекрытие границ классов:** Перекрывающиеся признаки затрудняют различение классов моделями.

4. Проблемы форматов представления данных

- **Несогласованное форматирование:** Различия в форматах данных (например, дат или единиц измерения) приводят к ошибкам обработки.
- **Гетерогенные источники данных:** Данные из разных источников с различными форматами и структурами усложняют интеграцию.
- **Неструктурированные данные:** Текст, изображения или аудио требуют специальной обработки для машинного обучения.

5. Безопасность данных

- **Вопросы конфиденциальности:** Конфиденциальная информация должна быть анонимизирована или обрабатываться с осторожностью.
- **Законы о защите данных:** Необходимо соблюдение таких правил, как GDPR или HIPAA.
- **Проблемы безопасности:** Утечки или нарушения безопасности данных могут поставить под угрозу конфиденциальные наборы данных.

6. Вызовы при разметке данных

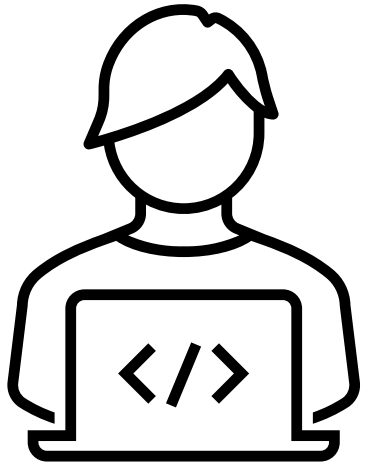
- **Недостаточная разметка:** Небольшие или неправильно размеченные наборы данных снижают качество модели.
- **Дорогие процессы разметки:** Разметка вручную, особенно в специализированных областях (например, в медицине), обходится дорого.
- **Неоднозначность классов:** Неоднозначные метки затрудняют точное обучение моделей.

Что такое предобработка данных?

Ресурсы для изучения ML

- Scikit learn: https://scikit-learn.org/stable/
- Kaggle: https://www.kaggle.com/
- Machine Learning Mastery: https://machinelearningmastery.com/
- Books: https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md

Практика



Ресурсы для изучения ML

Лабораторные работы