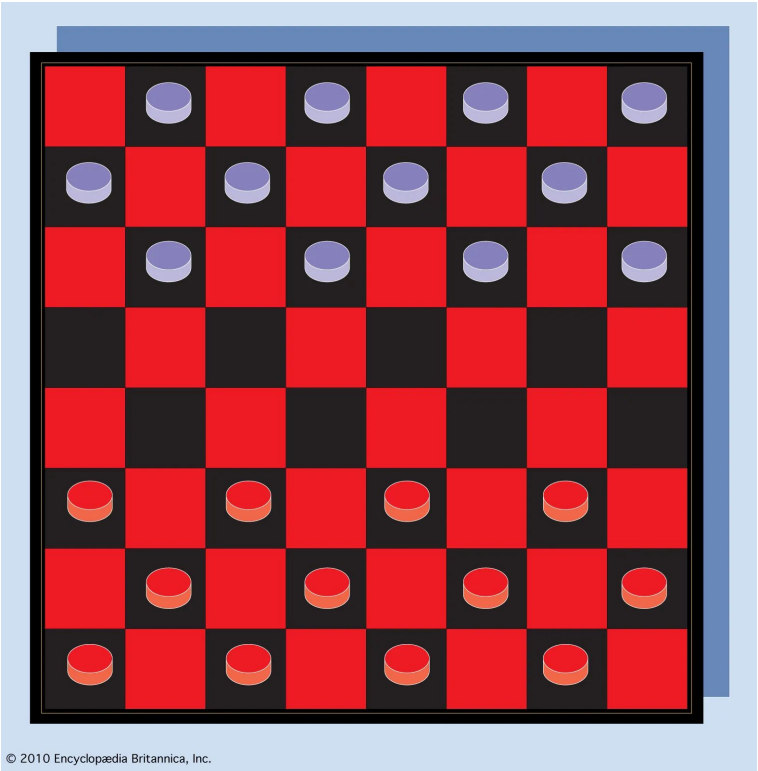
Машинное обучение

Введение

Что такое машинное обучение/Machine Learning

- A field of study that gives computers the ability to learn without being explicitly programmed (A. Samuel, 1959);
- Можно не придумывать правила, а позволить машине выучить их самостоятельно.

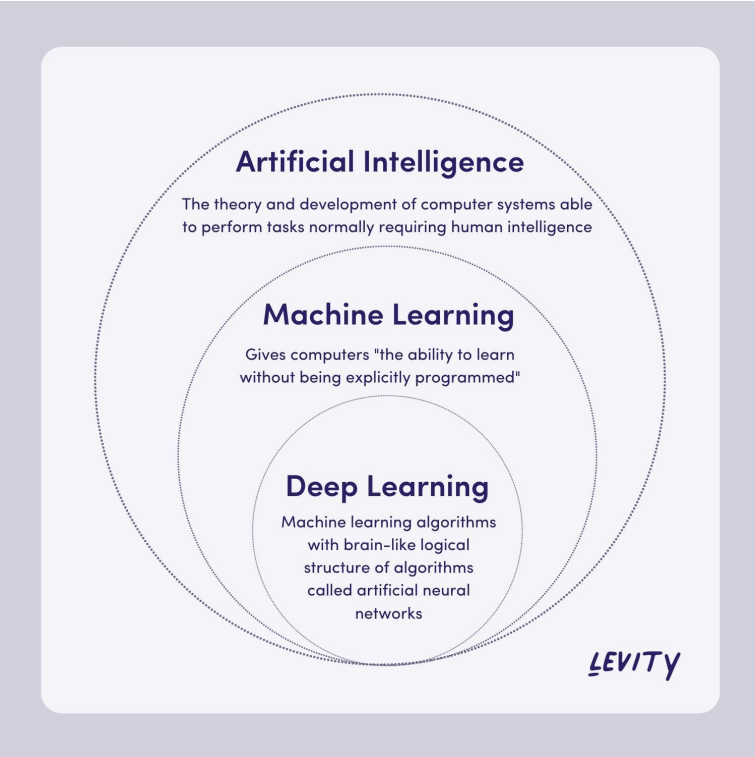


Цитата: A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," in IBM Journal of Research and Development, vol. 3, no. 3, pp. 210-229, July 1959, doi: 10.1147/rd.33.0210.

Картинка: Encyclopaedia Britannica

Machine Learning vs. Deep Learning

- Глубокое обучение (Deep Learning) - группа задач машинного обучения с использованием нейронных сетей.
- Нейросетям не нужен feature engineering: они сами вычленяют признаки, на которые надо обратить внимание;
- Нейросети хорошо работают с очень большими наборами данных;
- Нейросети сложнее интерпретировать.



Картинка: Levity.ai

Применение машинного обучения

- Рекомендательные системы
- Распознавание лиц и предметов (например, чтение рентгенограмм)
- Машинный перевод
- …

Про этот курс

Что будет на этом курсе?

- Основные задачи машинного обучения: классификация, регрессия, кластеризация, методы снижения размерности, ансамбли
- Оценка работы моделей, анализ ошибок
- Глубокое обучение: как работают нейросети, “классический Deep Learning” (CNN, RNN, LSTM), трансформеры

Оценка знаний

- Квизы:
  - Для оценки вашего понимания материала;
  - Большинство вопросов закрытые;
  - Можно проходить сколько угодно раз, но оцениваются только первые два;
  - Дедлайн в конце курса;
  - Засчитываться будет лучшее прохождение из двух первых, **поэтому, пожалуйста, пользуйтесь всегда одной почтой!**

Оценка знаний

- 2 лабораторные работы:
  - №1: Лабораторная на всё “классическое” машинное обучение;
  - №2: Лабораторная по глубокому обучению;
  - Дедлайн будет указан для каждой лабораторной отдельно, но не меньше двух недель.

Формула оценки

Квизы*0.2 + Лабораторная работа №1*0.4 + Лабораторная работа №2*0.4

Основные принципы сдачи работ

- ВСЕ задания курса выполняются самостоятельно;
- ВСЕ задания на программирование выполняются на питоне. По возможности, придерживайтесь нашего стека:
  - scikit-learn
  - pytorch
  - transformers

Допустимы другие ПОХОЖИЕ библиотеки - например, keras вместо pytorch, но их использование должно быть согласовано.

Методы машинного обучения

- Обучение с учителем (supervised learning)
- Обучение без учителя (unsupervised learning)
- Частичное обучение (semi-supervised learning)
- Обучение с подкреплением (reinforcement learning)

Обучение с учителем / supervised learning

Supervised learning

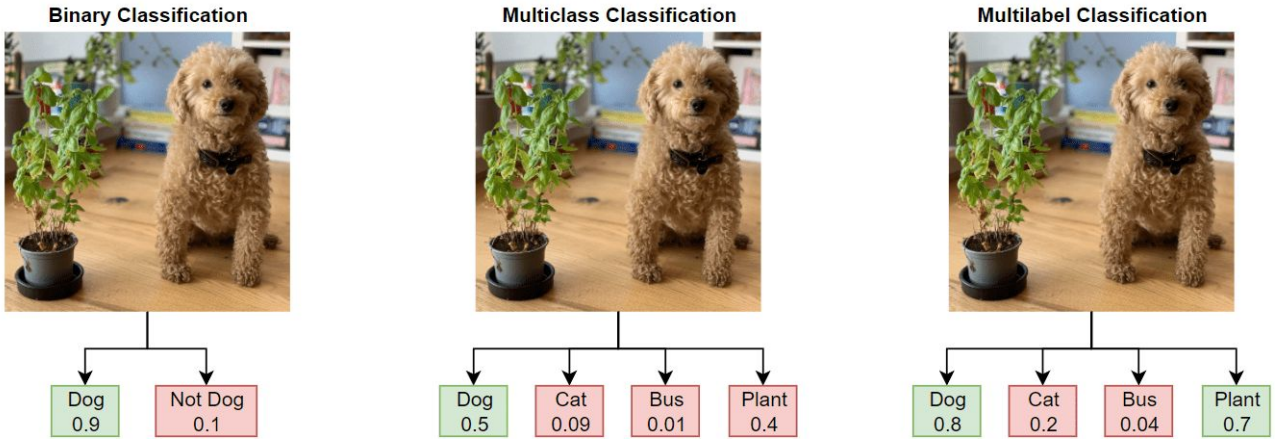
Есть входы X (набор признаков, источник) и выходы y (целевая переменная, таргет). Постройте функцию, максимально приближающую зависимость y от X.

Примеры:

- Регрессия: предскажите возраст человека по тексту (шкала от ~5 до ~100)
- Классификация: предскажите тон комментария по тексту (заданное количество тонов)

Классификация

Задача: построить алгоритм, который разделял бы выборку на заранее известные классы.



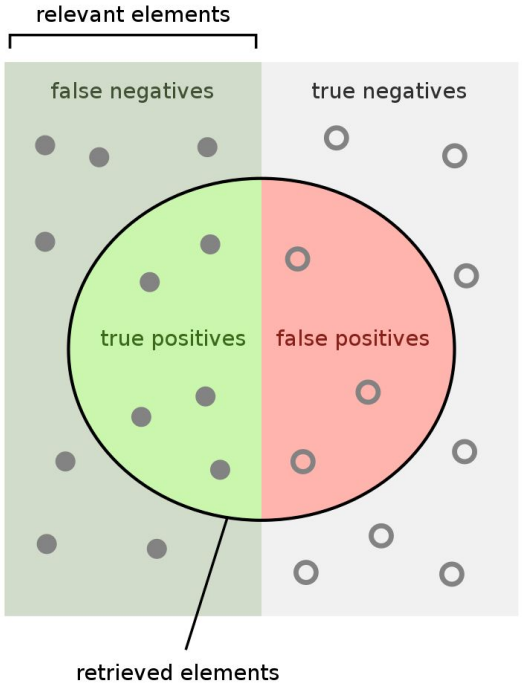
Картинка: www.mathworks.com

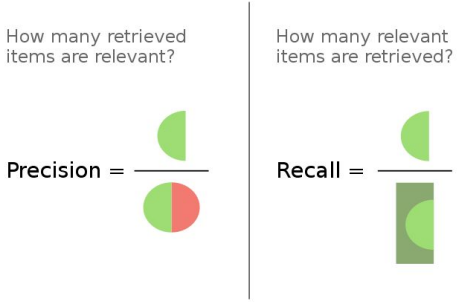
Классификация

Как оценивать качество? Пример для бинарной классификации (делим элементы на релевантные и нерелевантные):

- Точность (precision): сколько элементов, определенных моделью как релевантные, действительно релевантны?
- Полнота (recall): сколько действительно релевантных элементов найдено?
- F1-score: гармоническое среднее между точностью и полнотой.

$$F1 = \frac{2PR}{P + R}$$





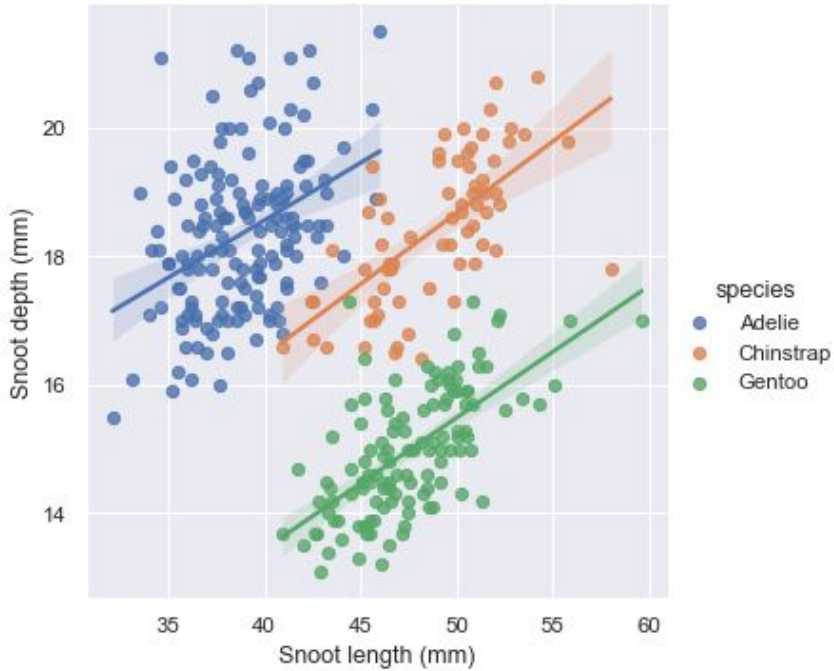
Картинка: Walber - Diagram representation of Precision and Recall, https://commons.wikimedia.org/w/index.php?curid=36926283

Регрессия

Задача: исследовать влияние независимых переменных на зависимую.

Как оценивать качество?

- Среднеквадратичная ошибка;
- Доля объясненной дисперсии;
- …

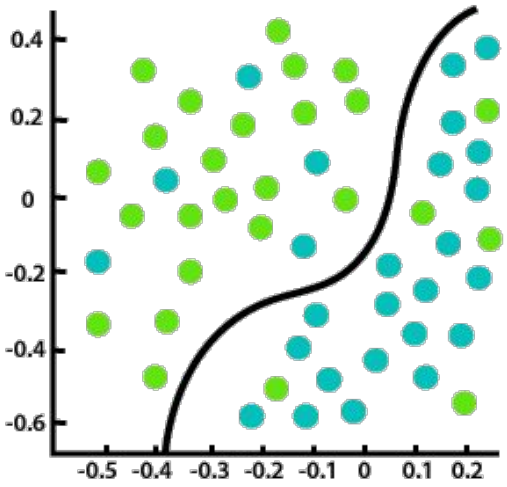


Картинка: Multiple linear regression in Seaborn, https://seaborn.pydata.org/examples/multiple_regression.html

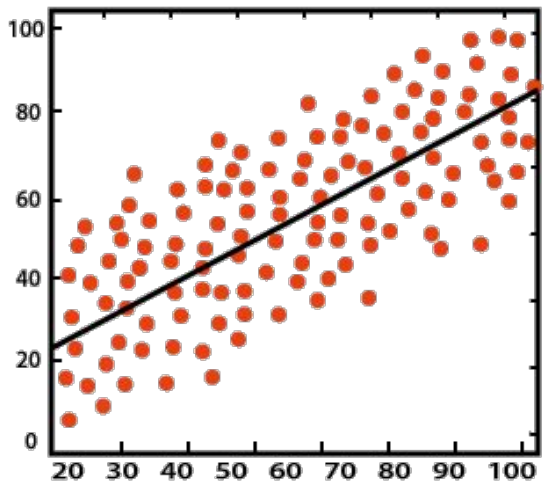
Supervised learning

Классификация: мы знаем, где зеленые точки, а где голубые. Мы хотим научиться отличать зеленые от голубых.

Регрессия: мы знаем, где красные точки. Мы хотим предсказать, где появятся следующие красные точки.



Classification

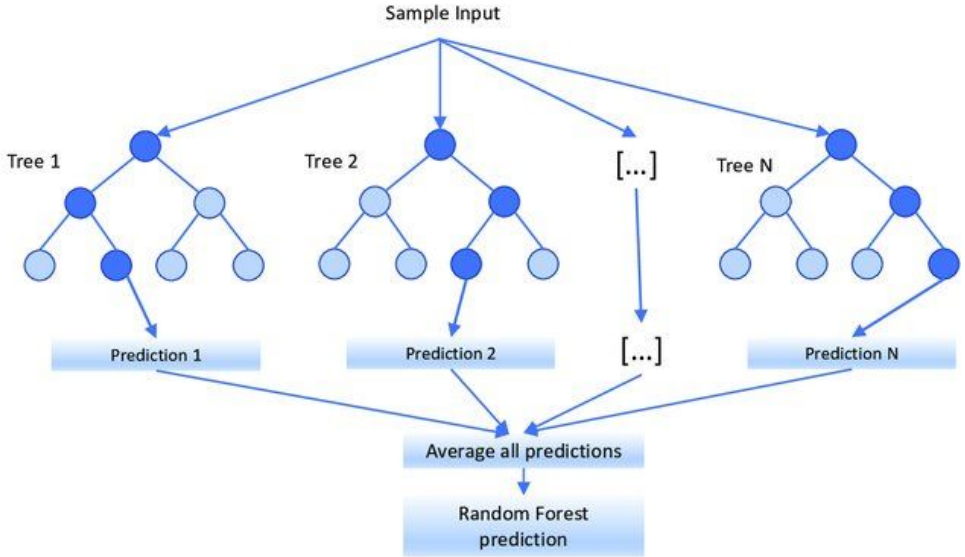


Regression

Картинка: www.javatpoint.com

Ансамбли

- Идея: чтобы улучшить качество предсказаний, можно вместо одной модели обучить несколько и скомбинировать их
- Примеры:
  - Рандомный лес
  - Градиентный бустинг
  - …
- Часто применяются для задач классификации



Источник: Segura, David & Khatib, Emil & Barco, Raquel. (2022). Dynamic Packet Duplication for Industrial URLLC. Sensors. 22. 587. 10.3390/s22020587.

Обучение без учителя / unsupervised learning

Unsupervised learning

Дан набор объектов одного рода. Обнаружьте зависимости между ними.

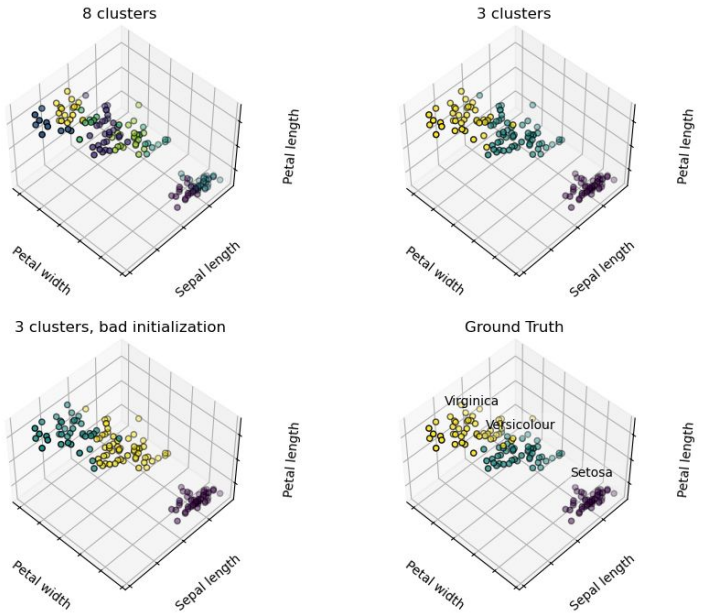
Важно: мы не пытаемся объяснить зависимую переменную! Просто ищем паттерны в данных.

Примеры:

- Кластеризация: поделите новости на тематические кластеры;
- Снижение размерности: подготовьте к классификации массив текстовых данных

Кластеризация

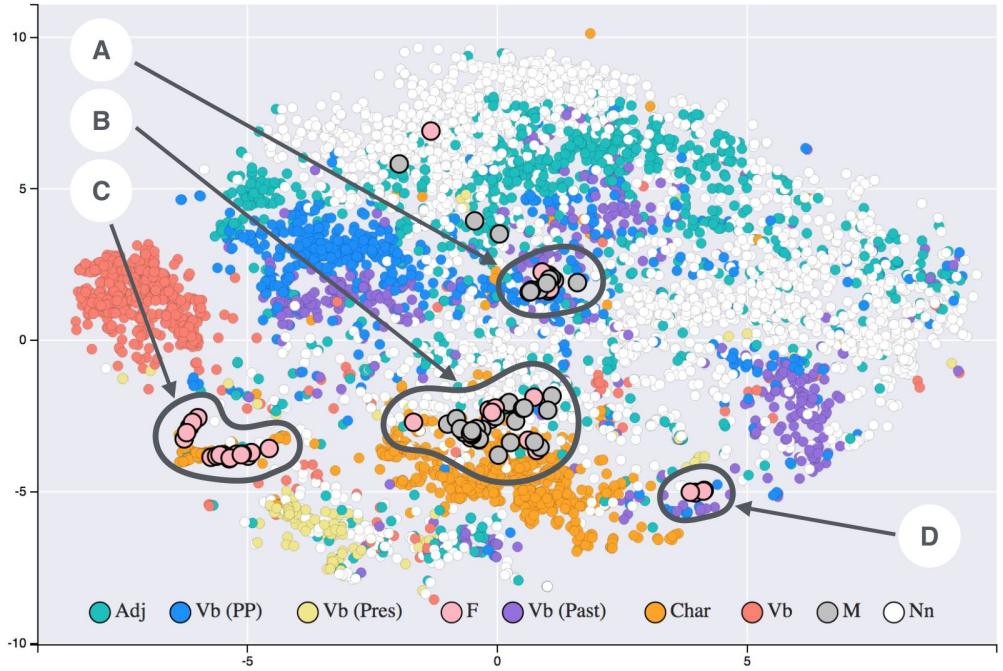
- Задача: поделить объекты на группы таким образом, чтобы объекты в каждой группе были больше похожи друг на друга, чем на объекты другой группы.
- Качество можно оценить, если истинные лейблы (и, соответственно, их количество) известны. Важно: нам НЕ нужны лейблы для обучения.



Источник: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_iris.html

Снижение размерности

- Задача: снизить размерность матрицы признаков, потеряв как можно меньше информации;
- Можно использовать для моделирования топиков;
- Также часто используется перед другой задачей (например, перед классификацией) или для визуализации данных.



Источник: Siobhán Grayson - T-SNE visualisation of word embeddings generated using 19th century literature, https://commons.wikimedia.org/w/index.php?curid=64541584

Классификация и кластеризация: в чём разница?

Задача: разделите новости по темам

Данные: корпус новостей, для каждой из которых мы знаем тему или набор тем.

Как учимся: даем модели признаки (вектора текстов) и зависимую переменную (класс).

Чего ждём: что модель научится разделять пространство признаков.

Как оцениваем: на отложенном тестовом множестве.

Данные: корпус новостей. Может быть, он вообще не размечен по темам, может быть, размечена только малая его часть.

Как учимся: даем модели признаки. Предполагаем, что в текстах может быть N разных тем. Даём это число модели.

Чего ждём: что модель найдет в пространстве признаков N отдельных групп.

Как оцениваем: можем оценить на малой размеченной части корпуса, если она есть.

Классификация и кластеризация: в чём разница?

Задача: разделите новости по темам

Данные: корпус новостей, для каждой из которых мы знаем тему или набор тем.

Как учимся: даем модели признаки (вектора текстов) и зависимую переменную (класс).

Чего ждём: что модель научится разделять пространство признаков.

Как оцениваем: на отложенном тестовом множестве.

Это **классификация.**

Данные: корпус новостей. Может быть, он вообще не размечен по темам, может быть, размечена только малая его часть.

Как учимся: даем модели признаки. Предполагаем, что в текстах может быть N разных тем. Даём это число модели.

Чего ждём: что модель найдет в пространстве признаков N отдельных групп.

Как оцениваем: можем оценить на малой размеченной части корпуса, если она есть.

Это **кластеризация.**

Другие методы: semi-supervised, reinforcement learning

Semi-supervised learning*

Одновременное обучение на аннотированных и неаннотированных данных.

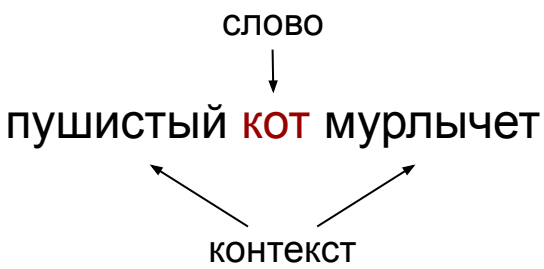
Примеры:

- Модель обучается на аннотированных данных, затем её предсказания используются для аннотирования остальной части выборки;
- Несколько моделей обучаются на аннотированных данных, их совместные предсказания используются для аннотирования остальной выборки.

Self-supervised learning*

Имея неаннотированные данные, достаньте из них объекты обучения самостоятельно.

- **Representation learning / обучение представлений**
  - Continuous Bag of Words: предскажите слово по контексту
  - Skip-gram: предскажите контекст по слову



Reinforcement learning*

Агент (модель) взаимодействует с неизвестной ему средой. Он должен выбрать такую стратегию поведения, которая максимизирует награду.

Пример: диалоговые системы

- Наблюдения: запрос от пользователя;
- Действия: генерация ответа;
- Награда: баллы за информативные, неодинаковые и грамматически правильные ответы.

Эксперименты в машинном обучении

Этапы эксперимента

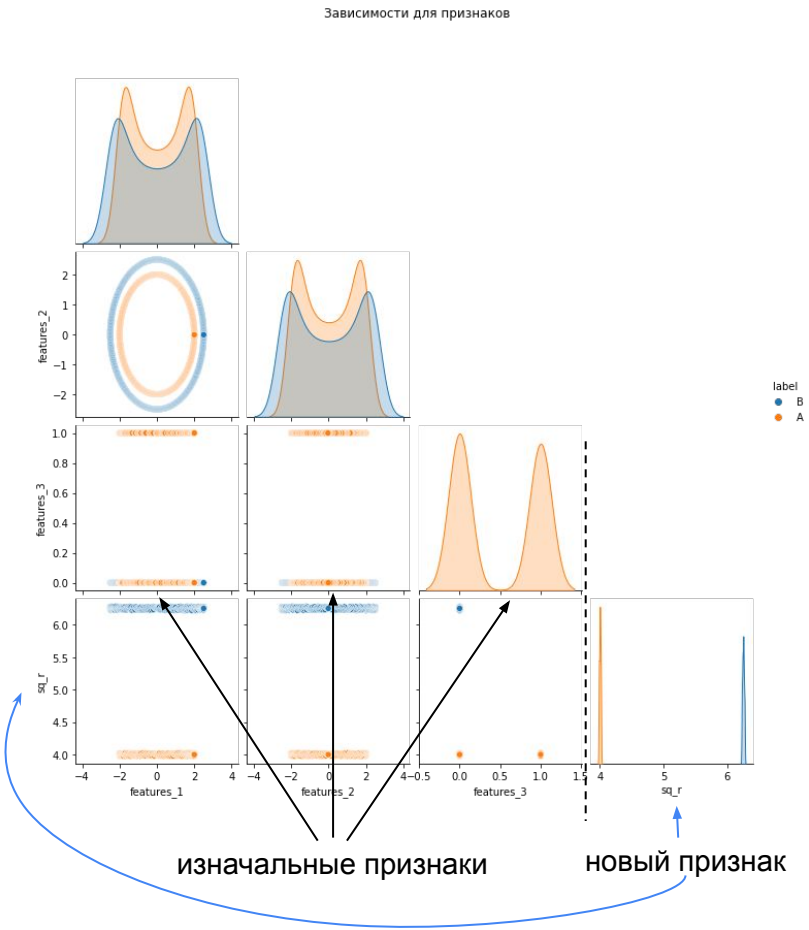
Подготовка данных

- Опциональный препроцессинг: очистка, лемматизация, удаление стоп-слов…
- Удаление пропусков (а также удаление выбросов или сохранение их для отдельной обработки)
- Токенизация
- Опционально: feature engineering/feature selection
- Разбиение на тренировочный и тестовый сет

Этапы эксперимента

Подготовка данных

- Опционально: feature engineering/feature selection*
  - Как правило, не выполняется для векторов слов;
  - Имеет смысл избавляться от признаков, которые, например, никак не могут объяснить зависимую переменную, либо пар признаков, сильно коррелирующих друг с другом;
  - Если признаки линейно неразделимы, можно попробовать их преобразовать (см. картинку).



Картинка: https://habr.com/ru/companies/ruvds/articles/680498/

Этапы эксперимента

- Выбор оптимальной модели для вашей задачи
  - Обратите внимание, по крайней мере, на следующие вопросы:
    - Какую задачу вы решаете?
    - К какому типу относятся ваши переменные?
    - Что вам важно: интерпретируемость или предсказательная способность?
- Обучение модели на обучающей выборке
- Тестирование модели на тестовой выборке

Этапы эксперимента

Оценка качества работы модели

- Автоматическими метриками (например, F-score для классификации, R^2 для регрессии)
- Анализ ошибок: посмотреть примеры, в которых модель ошибается. Есть ли объяснения этим ошибкам?
- Если работу модели оценивают люди, продумайте критерии оценки.

Пример ручного анализа ошибок

Допустим, вы анализируете работу модели для оценки CEFR-уровня текстов (A1, A2, B1 и так далее).

Вы используете относительные количества слов из определенных лексических минимумов (это списки слов, которые студент должен знать на определенном уровне освоения языка) в качестве независимых переменных.

Вы размечаете тексты по уровням согласно уровням учебников, из которых берете данные.

На тестовой выборке вы видите следующие результаты:

Предложение	**y_true**	**y_pred**
Антон изучает архитектуру.	A1	A1
В мультфильме *кот* ест пирожок.	A1	A2
Колбаса вкусная.	A1	A1

- *Курсивом* отмечены слова уровня A1;
- Подчеркнуты слова уровня A2;
- y_true - истинные уровни текстов;
- y_pred - предсказания модели.

Что можно сказать о причине ошибки модели в данном случае? Стоит ли её переучивать? Смог бы человек с точностью определять CEFR-уровни коротких текстов?

Про практические занятия и scikit-learn

Практические задания

Основная питоновая библиотека для классического ML - scikit-learn, он же sklearn.

- Проект начат в 2007м году; в 2010м первая публичная версия была выпущена исследователями из INRIA;
- Устанавливается командой pip install scikit-learn; в Google Colab предустановлен.

User Guide: https://scikit-learn.org/stable/user_guide.html. Рекомендуется как теоретическое пособие; практическое применение инструментов лучше смотреть отдельно на их страницах.

Пример:

1. Читаем про стохастический градиентный спуск в руководстве: https://scikit-learn.org/stable/modules/sgd.html
2. Смотрим описание класса и практическое применение классификатора на его странице: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html
3. Смотрим примеры внизу страницы.

Ещё есть словарь: https://scikit-learn.org/stable/glossary.html

Чего мы не будем делать

Здесь описаны некоторые частые ошибки в машинном обучении: https://scikit-learn.org/stable/common_pitfalls.html. Среди них, например:

1. Непоследовательный препроцессинг: например, вы используете некие инструменты для препроцессинга на тренировочных данных, но не на тестовых;
2. Утечка данных: информация о тестовом множестве каким-то образом “протекает” (leaks) в тренировочные данные, например, потому что:
    a. Тестовое множество не было выделено заранее;
    b. Метод .fit был вызван на тестовых данных. Тестовые данные только трансформируются, мы на них не обучаемся!

А что вы хотите?

Расскажите, пожалуйста, про ваши ожидания от этого курса: https://forms.yandex.ru/u/68a626ad5056902c42fd95db. Ваши ответы помогут мне познакомиться с вами, а также при необходимости скорректировать содержание занятий.

Практика

1. https://colab.research.google.com/drive/1YCzPA7p1fUbaNkmaiLtYNBKklZSAbsFy?usp=sharing
2. https://colab.research.google.com/drive/1LXNORdouvMCyuIuCic7w07Wwkpx4IlGL?usp=sharing

Источники

- Учебник Яндекса по машинному обучению: https://education.yandex.ru/handbook/ml/
- Scikit-learn’s User Guide: https://scikit-learn.org/stable/user_guide.html
- Курс по МЛ для математиков Ильи Щурова: https://github.com/ischurov/math-ml-hse-2018/tree/master
- Python Data Science Handbook: https://jakevdp.github.io/PythonDataScienceHandbook/