

Введение в машинное обучение



Как мы видим,
тут всё очевидно

Программисты программируют!

Датасаенс!

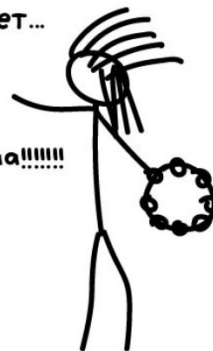
Профессия будущего!

Буквально через пять лет...

Экспоненциально!!!

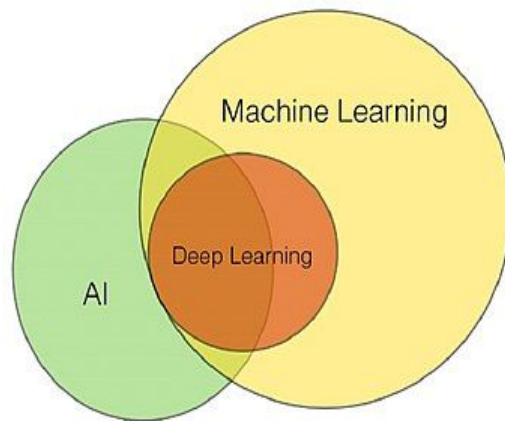
УМНЫЕ РОБОТЫ!

А-А-А-А-А-А-А-А-А-А-А-А-ааа!!!!!!!



Есть два типа статей про машинное обучение

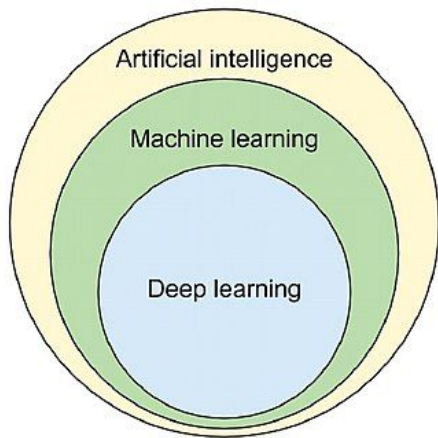
Что такое “машинное обучение”?



В машинном обучении модель, обучившись на некоторых данных, предсказывает результат по входным (новым) данным.

Чтобы обучить ML-модель, нужны:

- данные
- признаки
- сам алгоритм



ВИДЫ ML

Классическое

- признаки(features) выражены очевидно
- делится на “с учителем” и “без учителя”

Ансамбли

- набор классических ml-алгоритмов, обученных последовательно

Обучение с подкреплением

- задача -- не анализ данных, а выживание в среде (штрафы за ошибки)

Глубокое

- признаки выражены не очевидно (для человека)
- здесь нейросети

подробнее можно почитать [здесь](#)

Данные

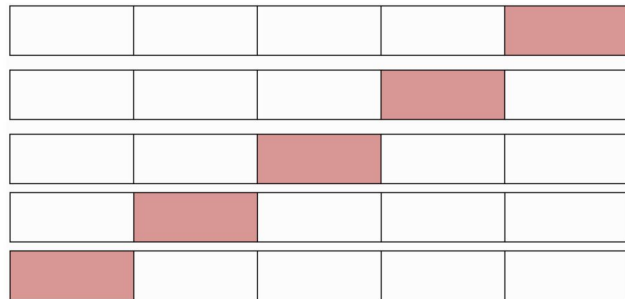
основная идея любого ML-алгоритма: даем обучающие данные >> модель учится >> тестируем на новых данных

Данные можно разделить на:

- обучающие и тестовые (train/test) - обычно делят на 80/20
- обучающие, валидацию, тест (train, validation, test)

Кросс-валидация:

- Делим выборку на k непересекающихся одинаковых частей;
- Обучаем k раз, каждый раз на $k-1$ части обучающей выборки;
- Тестируем на части выборки, которая не участвовала в обучении.



Важно, чтобы в процессе обучения модель не видела тестовые данные

Оценка резултата

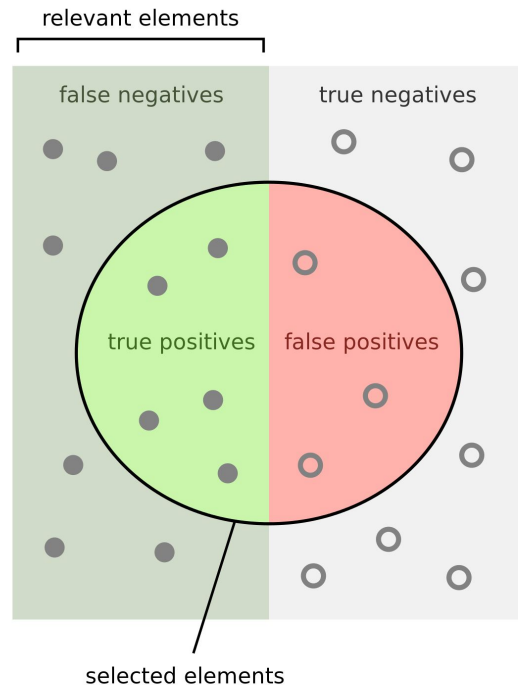
$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

$$\text{F1} = 2(\text{P} * \text{R}) / (\text{P} + \text{R})$$

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN



How many selected items are relevant?

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Классическое машинное обучение (Classic ML)

С учителем (supervised ML)

- машина учится на конкретных размеченных примерах, затем тестируется
- подтипы задач:
 - ◆ классификация — предсказание категории объекта
 - *спам-фильтр в почте*
 - ◆ регрессия — предсказание места на числовой прямой
 - *в какие часы меньше всего пробок*

Без учителя (unsupervised ML)

- Обучающие данные не размечены, машина учится сама выявлять закономерности. Затем тестируется
- подтипы задач:
 - ◆ Кластеризация
 - *тематическое моделирование*
 - ◆ Уменьшение размерности
 - *рекомендательные системы: определение вкусов пользователя*
 - ◆ Поиск правил
 - *какие товары пользователи покупают вместе*

Обучение с подкреплением (Reinforcement Learning)

Примеры:

- Самоуправляемые автомобили
- Роботы-пылесосы
- Игры (автоматическое прохождение Mario)
- Автоматическая торговля (ботнеты на биржах)

Идея:

минимизировать ошибки, пока выполняешь задание в среде

Ансамбли (Ensembles)

Примеры:

- Поисковые системы
- Компьютерное зрение
- Распознавание объектов

Идея:

Берем несколько алгоритмов, последовательно обучаем, особенно -- исправлять ошибки предыдущих

Глубокое обучение (Deep Learning)

Примеры:

- Распознавание изображений, объектов, перенос стиля
- Машинный перевод, генерация текстов
- Синтез речи
- Примерно все задачи, которые можно себе представить

Идея:

модель(нейросеть) сама находит нужные признаки в данных в удобном для себя виде (не всегда интерпретируемом человеку), прогоняя данные через свои слои, минимизируя функцию ошибки, и так обучается решать задачу

О некотором в ML подробнее

Классическое Обучение



Классическое ML: с учителем: задачи классификации: Наивный Байес

- неплохо решает задачи бинарной классификации
- сейчас используется реже
- подробнее про работу алгоритма

привет... 1829
валера ...1710
нет ... 1191
куда ... 1012
небо ... 985
огурцы ... 873
говорить...747
третий ... 739

нормальные письма

виагра ... 1552
казино ... 1492
100% ... 1320
кредит... 1184
скидка ... 985
нажми ... 873
free ... 747
доход ... 739

спам-письма

672 раза

«КОТИК»

13 раз

Простейший спам-фильтр
(использовались года до 2010)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

формула Байеса



не спам

Наивный Байес

Классическое ML: с учителем: задачи классификации: Деревья Решений

- данные структурируются по вопросам, на которые можно ответить «да» или «нет»
- сами по себе деревья не очень “умные”, но наборы деревьев эффективны и быстры
- [подробнее про работу деревьев](#)

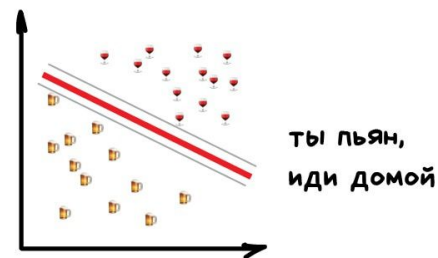
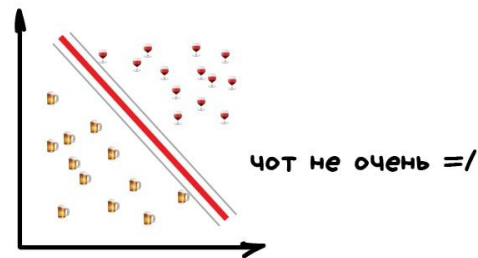
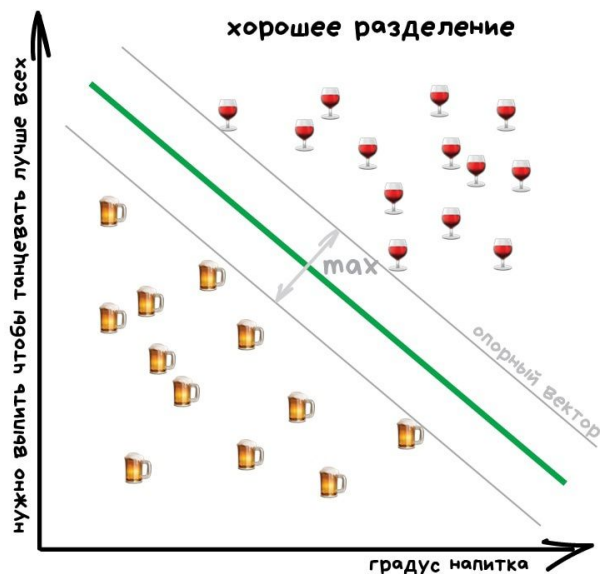
Давать ли кредит?



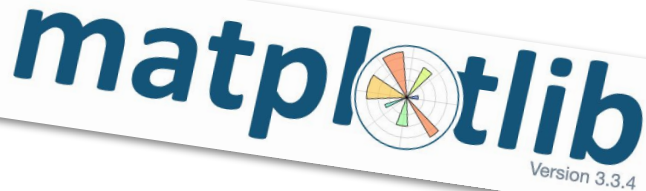
Классическое ML: с учителем: задачи классификации: метод опорных векторов (SVM)

Разделяем виды алкоголя

- разделяем два класса так, чтобы образовался наибольший “зазор” между ними
- [подробнее про SVM](#)



Основные библиотеки



matplotlib
Version 3.3.4



seaborn



Numpy



Scikit-learn



pandas

- [NumPy](#) (matrices, linear algebra, random numbers)
- [SciPy](#) (linear algebra, image optimization, signal and image processing)
- [Scikit-learn](#) (most popular ML algorithms + preprocessing + eval)
- [Pandas](#) (data manipulation)
- [NLTK](#) (NL data pre-processing)
- [matplotlib](#), [seaborn](#) (viz)

Полезные ссылки

Книги:

- [Hal Daume](#)
- [Joav Goldberg](#)

Статьи

- [обзорная про ML](#)
- [Про первый год работы ML-инженером](#)
- [Про best practices](#)
-

Разное

- [Датасеты+ тьюториалы](#)
- [Справочник](#)