Линейные модели: введение

Н. Поваров, И. Куралёнок

СПб,
2020

Что нужно для понимания*

- Теория вероятностей и математическая статистика
- Линейная алгебра
- Язык программирования

Как отчитываться

- Будет экзамен, возможно письменный
- Возможно тесты перед лекцией

Цель

- Уметь сформулировать задачу в терминах ML
- Найти подходящий класс решающих алгоритмов по формулировке
- Ориентироваться в области и знать “где посмотреть” существующие решения
- Понимать границы применимости

Что будет в модуле

- Общая картина дисциплины
- Линейные модели
- Интерпретация линейных моделей

Чего не будет в модуле

- Time series
- Полноценного Data Mining

Что почитать?

- Википедия
- R. Tibshirani, J. Friedman “Introduction to Statistical Learning”
- T. Hastie, R. Tibshirani, J. Friedman “The elements of Statistical Learning” **
- Труды конференций: ICML, NIPS, CIKM, KDD, etc. **

Машинное обучение: определение

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

ru.wikipedia.org

Машинное обучение: определение

Machine learning — the ability of a machine to improve its performance based on previous results.

Webster

Машинное обучение: определение

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Tom M. Mitchell

Машинное обучение в картинках

История

- 50-70гг — базы знаний, полнотекстовый поиск, распознавание образов, нейронные сети
- 70-80гг — ID3 деревья, разумные практические результаты, VC-оценки
- 80-90гг — первые конференции, много практического применения, активное применение кластеризации в анализе
- 90-00гг — повторное сэмплирование в ML, SVM, применение в IR, ML != DM, LASSO, bootstrap, bagging, boosting
- 00-10гг — Compressed sensing и прочие восстановления сигналов, царство деревьев, развитие ансамблей, . . .
- 10-20гг — Deep Learning, Convolutional, Recurrent, GANN, Transformers

Основные понятия

- Область работы = Universe = Γ.
- Решающая функция = Decision Function = $F_0 \in F$ – класс решающих функций.
- Опыт = Data Set = $\mathrm{D} = X \times Y$
- Целевая функция = Target = $\mathrm{T}(\mathrm{y}, \mathrm{F}(\mathrm{x}))$

Задача обучения

В ML оптимизация часто проводится в одних условиях, а эксплуатация в других.

$$\underset{F,B:\, F_0 = B(F)}{\operatorname{argmax}} \quad A(\Gamma, F_0)$$

A — цели эксплуатации (например деньги) на всей области работы

B — способ оптимизации, который реализуем

Классификация машинного обучения

ML можно делить по:

- виду целевой функции;
- способу получения опыта;
- классу решающих функций.

Классификация машинного обучения

ML можно делить по:

- **виду целевой функции;**
- способу получения опыта;
- классу решающих функций.

Классификация машинного обучения: цель

- С учителем
  - классификация (classification);
  - регрессия (regression);
  - отношение порядка (learning to rank);
  - обучение метрики (metric learning).
- Без учителя:
  - кластеризация (cluster analysis);
  - уменьшение размерности (dimensionality reduction);
  - обучение отображению (representation learning).
- Смешанные

Классификация машинного обучения: цель

- С учителем
  - классификация (classification);
  - регрессия (regression);
  - отношение порядка (learning to rank);
  - обучение метрики (metric learning).
- Без учителя:
  - кластеризация (cluster analysis);
  - уменьшение размерности (dimensionality reduction);
  - обучение отображению (representation learning).
- Смешанные

Обучение с учителем: два примера

- Классификация:
  - $y \in \{-1,\ 1\},\ \ F\colon \Gamma \longrightarrow [0,1]$
  - $y \in \{1,\ \ m\},\ F\colon \Gamma \longrightarrow [0,1]^m$
  - $y \in \{0, 1,\ \ m\},\ \ F\colon \Gamma \longrightarrow [0,1]^{m+1}$
- Аппроксимация: $y \in \mathbb{R}, F\colon \Gamma \longrightarrow \mathbb{R}$

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

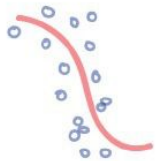
Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»



Без учителя

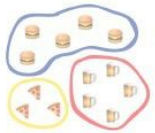
Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»



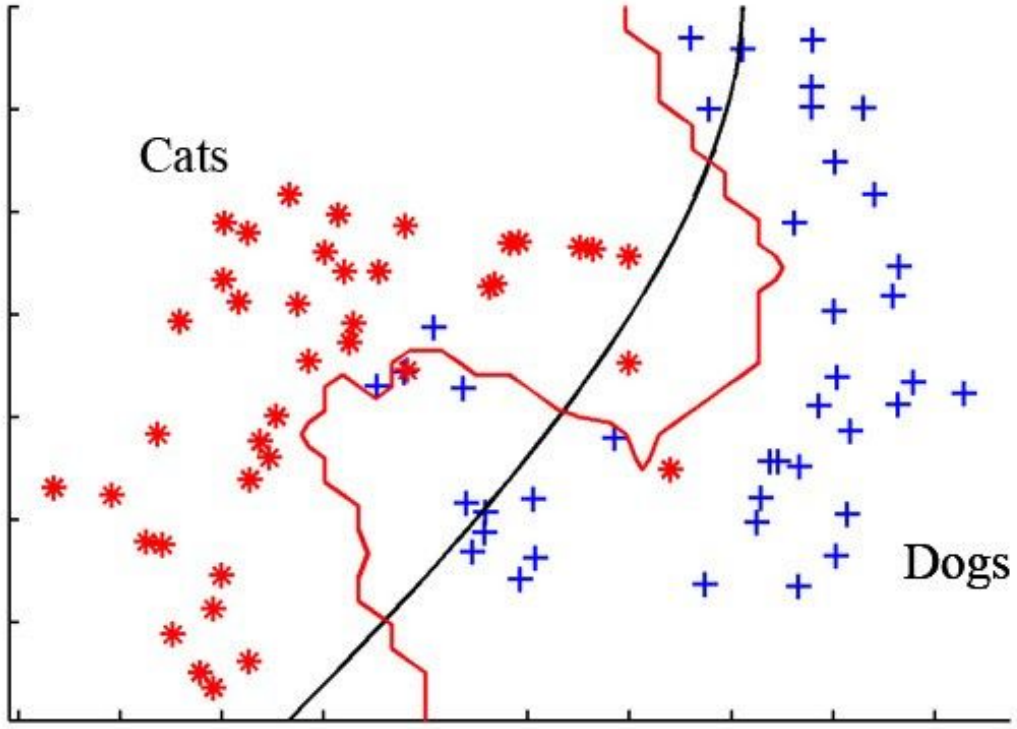
Ассоциация

«Найди какие вещи я часто ношу вместе»

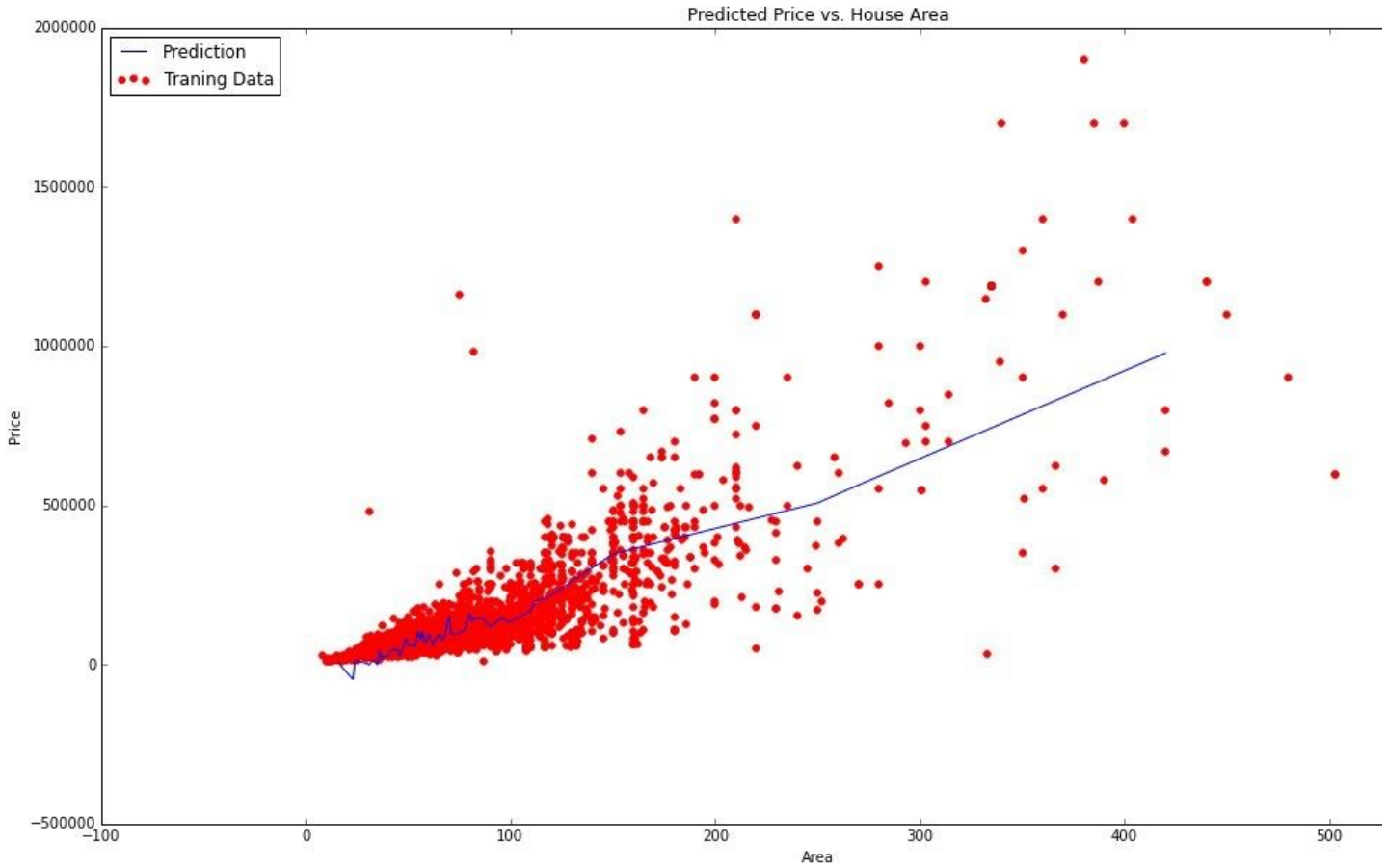
Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Классификация в картинках



Регрессия в картинках



Классификация машинного обучения

ML можно делить по:

- виду целевой функции;
- **способу получения опыта;**
- классу решающих функций.

Классификация машинного обучения: опыт

- Transductive learning
- Обычное обучение
- Активное обучение (active learning)
- Обучение с бюджетом (budget learning)
- Интерактивное обучение (online learning)
- Многорукие бандиты (multi-armed bandits)
- Обучение с подкреплением (reinforcement learning)

Обычное обучение

- Определяем генеральную совокупность Γ
- Фиксируем множество примеров X
- Обучаемся на доступных примерах

$$F_0 = B(F, D)$$

Классификация машинного обучения: опыт

- Transductive learning
- Обычное обучение
- Активное обучение (active learning)
- Обучение с бюджетом (budget learning)
- Интерактивное обучение (online learning)
- Многорукие бандиты (multi-armed bandits)
- Обучение с подкреплением (reinforcement learning)

Классификация машинного обучения: опыт

- Transductive learning
- Обычное обучение
- Активное обучение (active learning)
- Обучение с бюджетом (budget learning)
- Интерактивное обучение (online learning)
- Многорукие бандиты (multi-armed bandits)
- Обучение с подкреплением (reinforcement learning)

Классификация машинного обучения: опыт

- Transductive learning
- Обычное обучение
- Активное обучение (active learning)
- Обучение с бюджетом (budget learning)
- Интерактивное обучение (online learning)
- Многорукие бандиты (multi-armed bandits)
- Обучение с подкреплением (reinforcement learning)

Классификация машинного обучения

ML можно делить по:

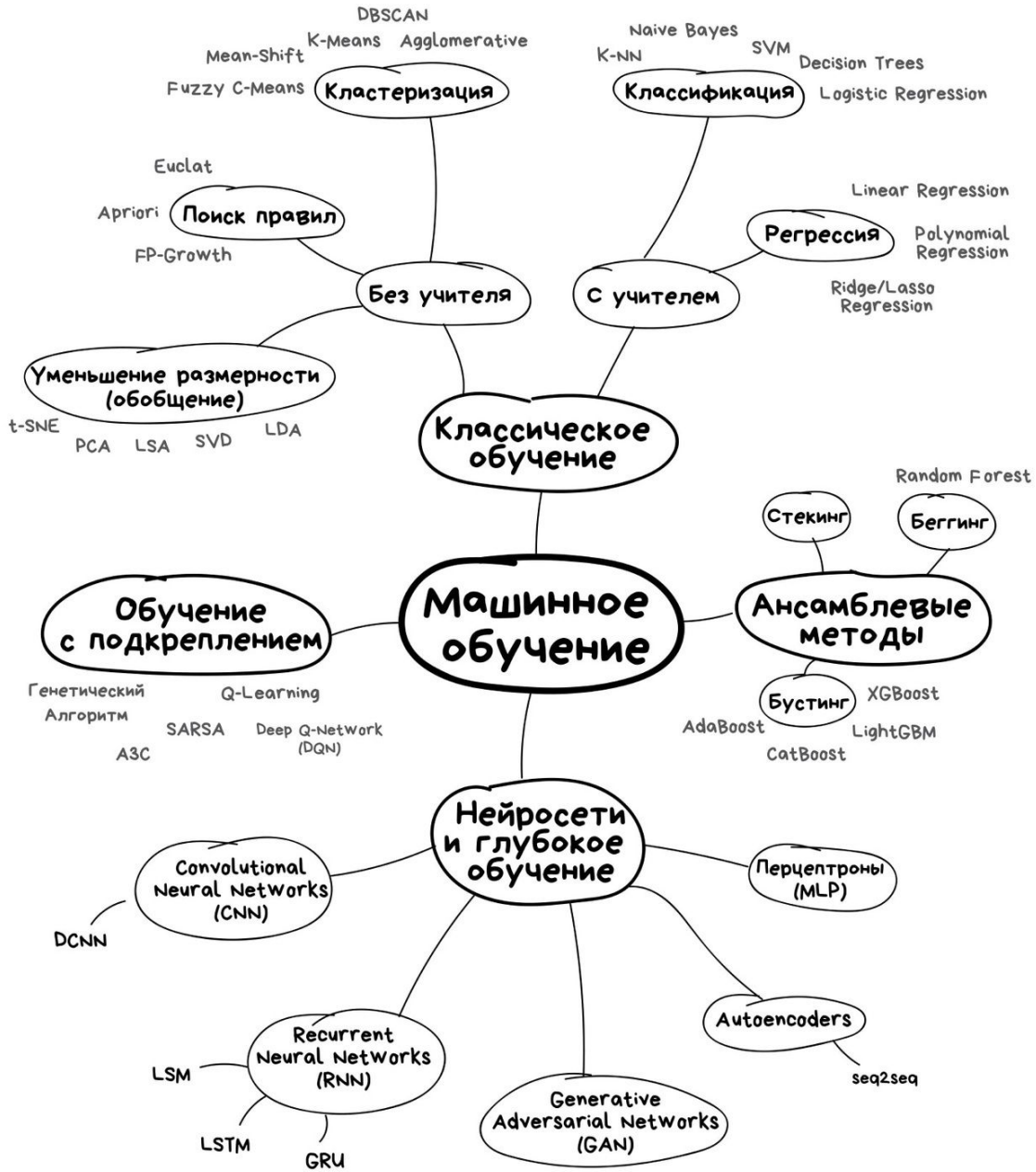
- виду целевой функции;
- способу получения опыта;
- **классу решающих функций.**

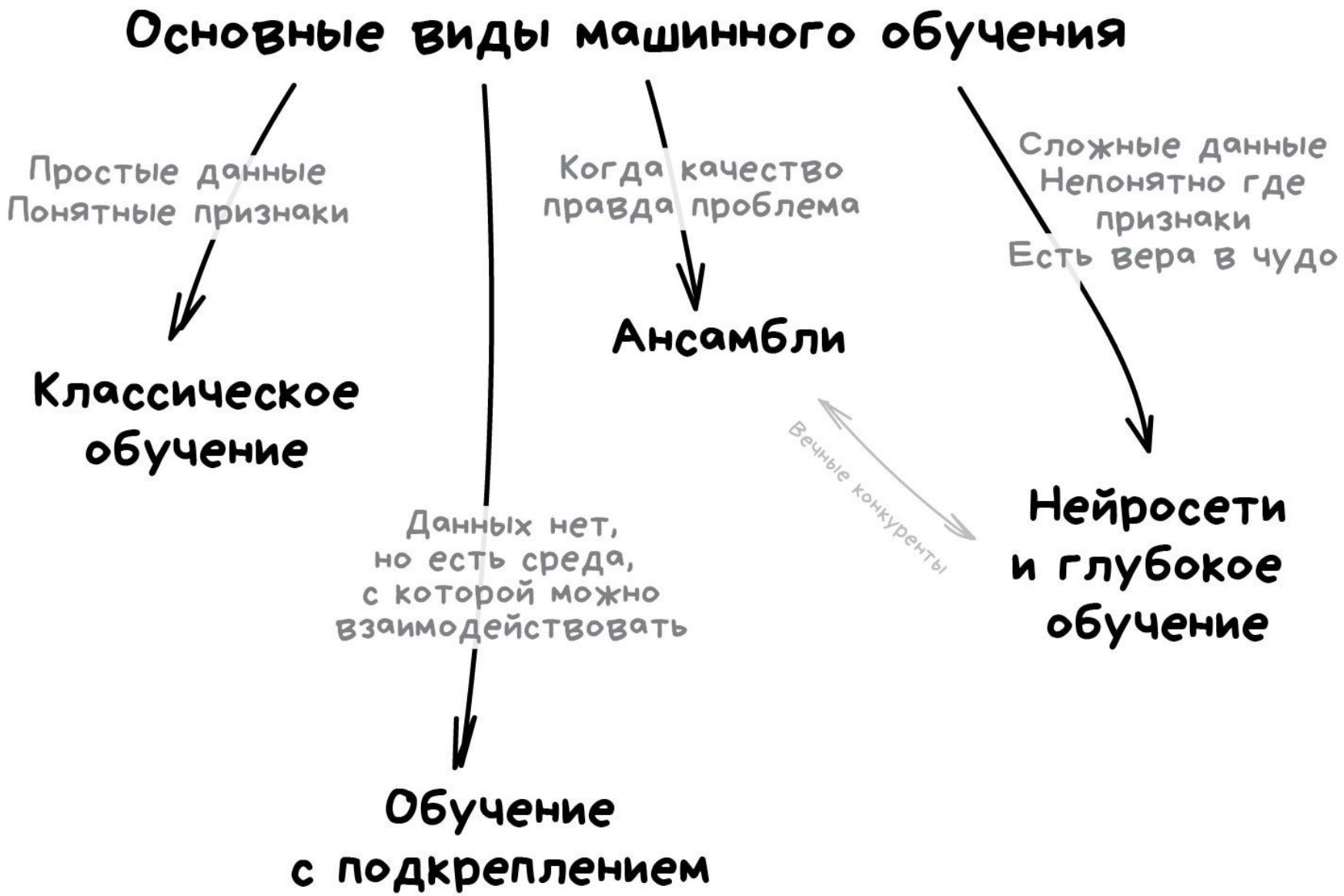
Основные классы решающих функций

- Instance based learning (kNN)
- Линейные решения
- Нейронные сети (ANN)
- Деревья решений
- Параметрические семейства
- …
- Ансамбли

Основные классы решающих функций

- Instance based learning (kNN)
- Линейные решения
- Нейронные сети (ANN)
- Деревья решений
- Параметрические семейства
- ...
- Ансамбли


Машинное обучение
Классическое обучение
Без учителя
Кластеризация
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Поиск правил
Euclat
Apriori
FP-Growth
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
С учителем
Классификация
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Обучение с подкреплением
Генетический Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
XGBoost
AdaBoost
LightGBM
CatBoost
Нейросети и глубокое обучение
Convolutional Neural Networks (CNN)
DCNN
Перцептроны (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq


Основные виды машинного обучения
Простые данные
Понятные признаки
Классическое обучение
Данных нет,
но есть среда,
с которой можно
взаимодействовать
Обучение
с подкреплением
Когда качество
правда проблема
Ансамбли
Вечные конкуренты
Сложные данные
Непонятно где
признаки
Есть вера в чудо
Нейросети
и глубокое
обучение

Итого во введении

- Определение
- История
- Классификация методов

Задача

Давайте попробуем отделить «плохих» студентов от «хороших»

Формальная постановка

Предскажем оценку ближайшей сессии

План работ

- Датасет
- Обучение
- Анализ результатов

План работ

- Датасет
- Обучение
- Анализ результатов

Датасет

- Готовый
- Запросим у ВУЗов
- Сделаем сами

Готовый датасет

+ Минимум усилий
+ Проверен
+ Легко хвастаться результатом

- Применимость
- Нет возможности повлиять

«Запросим» у ВУЗ-а

+ «Реальные» данные
+ Есть влияние
+ Может примениться

- Возможность повлиять ограничена
- Долго

Сделай сам

+ «Реальные» данные
+ Влияние полное
+ Может примениться

- Не похвастаться результатом
- Не всегда есть возможность

Наш датасет

- Результат опроса
- 51 человек
- 23 вопроса
- Давность: 2 года*

Векторизация

Векторизация: перевод представления о предмете в векторное выражение.

Компоненты полученного в результате векторизации вектора будем называть **факторами***

и **фичами** ☹

Студент -> вектор -> факторы

- Пол
- Город рождения
- Город учёбы в школе
- Город учёбы в институте
- Рост
- Год рождения
- Месяц рождения
- Братья/сестры
- Школьный балл
- Номер школы
- Тип школы
- Школьная математика
- Олимпиады
- Олимпиады по математике
- Победы в олимпиадах

Студент -> вектор -> факторы

- Пол (0,1)
- Город рождения (A-AH)
- Город учёбы в школе (A-AH)
- Город учёбы в институте (A)
- Рост (1-39)
- Год рождения (1-25)
- Месяц рождения (1-12)
- Братья/сестры (0, 1)
- Школьный балл (-1-5)
- Школа (A-AN)
- Тип школы (текст)
- Школьная математика (-1- 5)
- Олимпиады (0,1)
- Олимпиады по математике (0,1)
- Победы в олимпиадах (-1-1)

Адаптация факторов

- С городами беда -> сделаем фактор “понаехали”
- Сортирующиеся факторы размапим от балды
- В рядах возьмём min
- В пиве возьмём max
- В СШБ возьмём min
- Соцсети сложим
- Бинаризуем мобильные OS
- ...

План работ

- Датасет
- Обучение
- Анализ результатов

Решающая функция

- $$h_{\hat{\beta}}(x) = \hat{\beta}^T x$$

$$\hat{\beta}, x \in \mathbb{R}^n$$

- Простая
- Универсальная
- Легко интерпретируемая

Целевая функция

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{(x_i, y_i) \in L} \left\| h_\beta(x_i) - y_i \right\|$$

- Простая
- Универсальная
- Легко интерпретируемая

Решение

•

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{(x_i, y_i) \in L} \left\| x_i{}^T \beta - y_i \right\| =$$

$$= \underset{\beta}{\operatorname{argmin}} \sum_{(x_i, y_i) \in L} \left\| X^T \beta - y \right\|^2$$

$$\boldsymbol{\hat{\beta} = (XX^T)^{-1} Xy}$$

Результат

Среднее	4,22
Пол	-1,98
Города	-0,51
Рост	0,07
Год рождения	0,09
Месяц	0,03
Братья/сёстры	-0,63
Школьный балл	0,09
Школа 239	-1,29
Гимназия	-0,42
Оценка по математике	-0,20
Олимпиады	0,54
Олимпиады по математике	-0,42
Победы	-0,09
Путь	0,00
Общага	-0,25
Ряд	-0,04
Прогулы	0,49
Автоматы	0,48
Социальщина	0,00
iOs	-0,89
Интернет	-0,62
Спиннер	1,22
Пиво	-0,21

План работ

- Датасет
- Обучение
- Анализ результатов

Интерпретация результата

- Пол = -1,98 — ну, шовинизм не в моде
- Школа 239 = -1,29 — обычный расслабон
- Спиннер = 1,22 — рулит

Вопросы?

Интерпретация результата

- Пол = -1,98 — ну, шовинизм не в моде
- Школа 239 = -1,29 — обычный расслабон
- Спиннер = 1,22 — рулит

Это всё фигня!!!

Нормализация факторов

- Хотим min = 0, max =1
- Или матожидание 0 и дисперсия 1
- Или хотим от -1 до 1

Результат II

Среднее	4,22
Пол	-1,95
Города	-0,44
Рост	2,75
Год рождения	2,25
Месяц	0,30
Братья/сёстры	-0,59
Школьный балл	0,48
Школа 239	-1,27
Гимназия	-0,38
Оценка по математике	-0,99
Олимпиады	0,55
Олимпиады по математике	-0,44
Победы	-0,06
Путь	-0,57
Общага	-0,25
Ряд	-0,29
Прогулы	0,90
Автоматы	0,47
Социальщина	0,72
iOs	-0,85
Интернет	-1,13
Спиннер	1,17
Пиво	-0,91

Интерпретация результата II

- Пол = -1,95 — ну, всё ещё не в моде
- Школа 239 = -1,27 — обычный расслабон
- Спиннер = 1,17 — рулит
- Рост и год рождения = >2 — жгут оба

Вопросы?

Интерпретация результата II

- Пол = -1,95 — ну, всё ещё не в моде
- Школа 239 = -1,27 — обычный расслабон
- Спиннер = 1,17 — рулит
- Рост и год рождения = >2 — жгут оба

Но и это фигня!

Оценка результата

Можно поделить множество на две части и обучить на одной, а оценить на другой:

$$DS = L \cup T, \qquad L \cap T = \emptyset$$

+ Расскажет о качестве предсказания
+ Можно посмотреть на качество на L и на T

- Использует меньше данных в обучении
- Если исходное множество непоказательно, то всё равно всё плохо

Результат III

Ошибка на Learn	Ошибка на Test
2,55	1,50
1,92	1,27
4,38	1,33

Стабильность решения

Поделим несколько* раз и посмотрим как меняются компоненты решающей функции.

- Стабильные компоненты заслуживают веры
- Если всё нестабильно, то плохо
- Выкидываем лишнее

* несколько это 1 000 или 10 000

Результат IV

Оценка за экзамен
=
0.24 x Балл по математике
—
0.31 x Средний школьный балл
+
0.12 x Рост
+
4.42

Что ещё можно попробовать:

- Другая нормализация
- Удаление студентов из обучения
- Новые факторы
- Не такой жадный способ фильтрации
- Больше данных

Итого в примере:

- Всё равно сделали фигню
- Посмотрели как исходная задача формулируется в техническую
- Разобрали способы сделать датасет
- Обучили
- Попытались поинтерпретировать

Вопросы?