Практические задачи анализа данных

Лекция 1. Введение. Простые методы анализа данных

Московский авиационный институт
«МАИ»

1 сентября 2021 г.

План курса

1. Введение. Простые методы анализа данных
2. Понятие среднего. Оценка вероятности и плотности распределения. Прогноз визита покупателей
3. Визуализации
4. Ансамбли в машинном обучении
5. Интерпретация моделей машинного обучения
6. Настройка на функционал качества. Калибровка вероятностей
7. Задача с несбалансированными классами
8. Uplift моделирование
9. Анализ социальных сетей

Курс 1. Анализ данных

1. Что такое анализ данных и машинное обучение
2. Постановка основных задач
3. Метрические методы
4. Контроль качества: классификация и регрессия
5. Линейная и логистическая регрессии
6. Логические методы
7. Ансамбли
8. Градиентный бустинг

+ 8 лекций «Введение в нейронные сети»

План курса

Курс 2. Статистический анализ данных

1. Вводная часть
2. Обучение без учителя
   - Кластеризация
   - Поиск ассоциативных правил
   - Одноклассовая классификация (поиск аномалий)
3. Оценка качества алгоритмов
   - Смещение и разброс
   - Логистическая функция потерь
   - Коэффициент Джини
   - Функционалы качества в задаче многоклассовой классификации
   - Функционалы качества в задаче кластеризации
4. Работа с данными
   - Подготовка данных (Data preparation)
   - Отбор признаков (Feature selection)
   - Создание новых признаков (Feature engineering)

План курса

5. AutoML
   - Цели, содержание, история, сравнение фрейморков
   - Оптимизация гиперпараметров
   - LightAutoML от Сбербанк
6. Статистика в анализе данных
   - Бинарная классификация в классической статистике
   - Бинарная классификация в байесовской статистике
   - Наивный байесовский классификатор
   - Классификация в гауссовском модели
   - Робастный линейный классификатор
7. Методы снижения размерности
   - Метод главных компонент (PCA, Kernel PCA)
   - Нелинейное сокращение размерности
   - t-distributed stochastic neighbor embedding (t-SNE)
8. Gaussian Mixture Models. EM-алгоритм
9. Метод опорных векторов

+ лекции NLP и Рекомендательные системы

Откуда что берётся

Откуда что берётся

Sklearn — документация и код примеров

- https://scikit-learn.org/stable/index.html
- https://scikit-learn.org/stable/modules/cross_validation.html

Полезные библиотеки и ссылки

- Awesome Python https://github.com/vinta/awesome-python
- Компьютерные науки https://github.com/papers-we-love/papers-we-love
- Awesome open source libraries for ML https://github.com/EthicalML/awesome-production-machine-learning
- Awesome Big Data https://github.com/0xnr/awesome-bigdata
- ML & DL Tutorials https://github.com/ujjwalkarn/Machine-Learning-Tutorials
- Awesome deep learning https://github.com/ChristosChristofidis/awesome-deep-learning
- Awesome Pytorch https://github.com/bharathgs/Awesome-pytorch-list
- 500 ML Projects with code
- Open Source Projects https://github.com/ml-tooling/best-of-ml-python

Откуда что берётся

Kaggle

- обзор сообщества https://www.kaggle.com/dwin183287/kagglers-seen-by-continents
- наборы данных https://www.kaggle.com/datasets
- постановки бизнес-задач https://www.kaggle.com/c/acea-water-prediction
- обучение https://www.kaggle.com/learn
- практические советы https://www.kaggle.com/c/tabular-playground-series-apr-2021/discussion/231738
- полезный код
- РГР https://www.kaggle.com/c/iplmsac
- лучшие решения https://farid.one/kaggle-solutions/
- доступ к вычислительным ресурсам

Откуда что берётся

Соревнования

- Kaggle https://www.kaggle.com/competitions
- Общий список https://ods.ai/competitions
- Пример соревнования https://emergencydatahack.ru/

Отличие от бизнес-задач

- сформулирована постановка задачи
- нет бизнес-метрик
- большая часть этапа сбора и подготовки данных уже сделана
- данные часто анонимизированы
- возможность пользоваться чужими идеями и кодом
- целью является победа в соревновании, а не запуск решения в производство
- как правило строятся слишком сложные модели

Откуда что берётся

Данные

- 15 популярных наборов данных
- Поисковая система https://datasetsearch.research.google.com/
- ML datasets https://www.datasetlist.com/
- UCI Repository https://archive.ics.uci.edu/ml/datasets.php
- Sklearn https://scikit-learn.org/stable/datasets.html
- Проблема описания данных https://arxiv.org/abs/1803.09010
- Сборник ссылок https://www.kdnuggets.com/datasets/index.html
- Сборник ссылок https://github.com/awesomedata/awesome-public-datasets
- R Datasets
  - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html
  - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Учебные курсы

- Шпаргалки
- Список Ютуб-каналов
  https://github.com/benthecoder/yt-channels-DS-AI-ML-CS
- Awesome Computer Science courses
  https://github.com/prakhar1989/awesome-courses
- Мурмландия ФизТеха
  https://mipt-stats.gitlab.io/courses/ad_fivt/lecture1.pdf
- Corsera https://ru.coursera.org/search?query=machine+learning
- Stepik https://stepik.org/catalog
- КалТех https://work.caltech.edu/telecourse.html
- Эндрю Ын
  - https://ru.coursera.org/learn/machine-learning
  - https://youtu.be/jGwO_UgTS7I
- https://bloomberg.github.io/foml/
- Derek Kane. Data Science Lectures
- Karpov Course
- NLP https://lena-voita.github.io/nlp_course.html

Откуда что берётся

Учебные курсы

- Data mining in action https://github.com/data-mining-in-action
  https://www.youtube.com/channel/UCop3CelRVvrchG5lsPyxvHg/videos
- Deep Learning School
  https://www.youtube.com/channel/UCFTNoZYjkg-3LZTHrHfV1nQ
- МФТИ
  https://www.youtube.com/playlist?list=PL4_hYwCyhAvasRqzz4w562ce0esEwS0Mt
- Deep Bayes https://deepbayes.ru/
- Проекты студентов Стэнфорда по ML http://cs229.stanford.edu/projects.html

Откуда что берётся

Университеты

- ВШЭ http://wiki.cs.hse.ru/
- МФТИ https://youtube.com/playlist?list=PLk4h7dmY2eYHHTyfLyrl7HmP-H3mMAW08
- СПбГУ https://www.youtube.com/channel/UCdfMlHaF7spha_q8iM_LZHg/playlists
- ВМК
  - http://www.machinelearning.ru/
  - https://github.com/Dyakonov
- Киев https://github.com/fbeilstein/machine_learning
- ВМиИТ https://www.youtube.com/channel/UCcY6LFZNgZHR2skk4K_-PKw/featured

Откуда что берётся

Сообщества

- Open Data Scince
  - https://ods.ai/
  - https://mlcourse.ai/
  - учебные курсы (треки) https://ods.ai/tracks/mts-recsys-df2020
- Индия ML https://www.analyticsvidhya.com/blog/
- Мировой электронный журнал https://medium.com/
- Хабр
  - обучение https://habr.com/ru/post/321216/
  - новости https://habr.com/ru/company/cloud4y/blog/346968/
  - заметки из жизни https://habr.com/ru/post/295954/

Откуда что берётся

Блоги

- Кантора Andrew Ng https://read.deeplearning.ai/the-batch/
- Google AI Blog https://ai.googleblog.com/
- Системный блок https://sysblok.ru/
- Визуализации https://flowingdata.com/about/
- Domino Data Lab https://blog.dominodatalab.com/
- Transformers https://www.topbots.com/
- Личные блоги
  - Анализ малых данных https://dyakonov.org/
  - Рецензии статей https://andlukyane.com/blog/
  - Lilian Weng, Applied AI Research https://lilianweng.github.io/lil-log/
  - https://calculatedcontent.com/
  - http://jakevdp.github.io/

Литература

- https://ur.ru1lib.org/ (Z-library)
- Andrew Ng. Machine Learning Yearning https://habr.com/ru/post/419757/
- В. Майер-Шенбергер, К. Кукьер. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. Манн, Иванов и Фербер. 2014
- Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. https://web.stanford.edu/ hastie/ElemStatLearn/
- Серия книг https://machinelearningmastery.com/products/
- Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. –– М.: ДМК Пресс, 2016
- С. Рашка, В. Мирджалили. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow 2, 3-е изд. — СПб, Диалектика, 2020
- L. P. Coelho, W. Richert, M. Brucher. Building Machine Learning Systems with Python, 2018. Packt Publishing

Литература

- M. J. Zaki, W. Meira. Data Mining and Machine Learning Fundamental Concepts and Algorithms. Cambridge University Press. 2020
- Э. Гласснер. Глубокое обучение без математики. Основы. Практика. ДМК Пресс, 2019
- Хайкин С. Нейронные сети: полный курс, 2-е изд. —- М.: ООО «И.Д. Вильямс», 2016
- Charu C. Aggarwal. Data Mining: The Textbook. Springer, 2015
- I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. MIT Press, 2016
  https://www.deeplearningbook.org/
  Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение — М.: ДМК Пресс, 2018
- Франсуа Шолле. Глубокое обучение на Python. — СПб.: Питер, 2018

Откуда что берётся

Магистерские программы (совместные)

- ШАД от Яндекс https://yandexdataschool.ru/
  + МФТИ
  https://mipt.ru/education/chairs/da/education/masters/
  + ВШЭ https://www.hse.ru/ma/datasci/
- Mail.ru https://data.mail.ru/
  + Баумана https://park.mail.ru/pages/about/
- ВШЭ+СБЕР https://www.hse.ru/ma/fintech/
- Ozon https://ozonmasters.ru/
- Сколтех
  https://www.skoltech.ru/en/education/msc-programs/ds/
- МФТИ + X5 Retail Group https://mipt.x5.ru/
- МАИ + Avito
  https://mai.ru/press/news/detail.php?ID=88083
- МФТИ + СБЕР https://fpmi.mipt.ru/master/mach-learn/
- МИСиС https://data.misis.ru/

Откуда что берётся

Статьи, конференции

- Сервис для поиска книг, статей, докладов на конференциях https://dl.acm.org/
- Сборник ссылок https://mlpapers.org/
- Конференция KDD https://kdd.org/conferences
- Журналы
  https://analyticsindiamag.com/10-essential-academic-journals-data-scientists/
- Nature https://www.nature.com/sdata/articles?type=analysis
- Альманах искусственный интеллект (аналитический сборник) https://aireport.ru/
- наш журнал по ML http://jmlda.org/ru/journal

Откуда что берётся

Работа, разное

- что такое DS
  https://hdsr.mitpress.mit.edu/pub/gg6swfqh/release/1
- про карьеру и найм в DS
  https://ods.ai/tracks/ds-hiring-df2021
- сборник вопросов на собеседовании
  https://github.com/DopplerHQ/awesome-interview-questions
- опросник-тест
  164 Data Science Interview Questions and Answers
- тест https://www.gangboard.com/blog/data-science-interview-questions-and-answers
- Собеседования
  https://www.youtube.com/playlist?list=PLiSyTNGp2j5M1hMB4JEY34IxGR5tt6k_n

Мелкое мошенничество в AI/ML

Среди наиболее очевидных примеров из области машинного обучения:

- пробовать новый блестящий алгоритм на нескольких десятках сидов, а в статье сообщить только о лучших
- Хорошенько поработать над гиперпараметрами своего подхода, а для базовой линии использовать значения по умолчанию
- Специально выбирать примеры, где модель хорошо выглядит
- Специально выбирать для тестирования наборы данных, на которых модель доказала свои преимущества
- Придумывать новые постановки задач, новые наборы данных, новые цели, чтобы одержать победу на пустом игровом поле

Откуда что берётся

Мелкое мошенничество в AI/ML

- Заявить во введении, что работа — «многообещающий первый шаг», хотя вы прекрасно понимаете, что никто и никогда не будет её развивать
- Поняв, что основные идеи доклада не совсем верны, всё равно отправить его на конференцию, ведь время не должно быть потрачено впустую

«Большинство исследователей в той или иной степени являются карьерными исследователями, мотивированными властью и престижем, которые вознаграждают тех, кто преуспевает в академической системе, а не идеалистическим стремлением к научной истине»

Кризис воспроизводимости

Простые методы анализа данных

Простые методы в АД

Простые правила

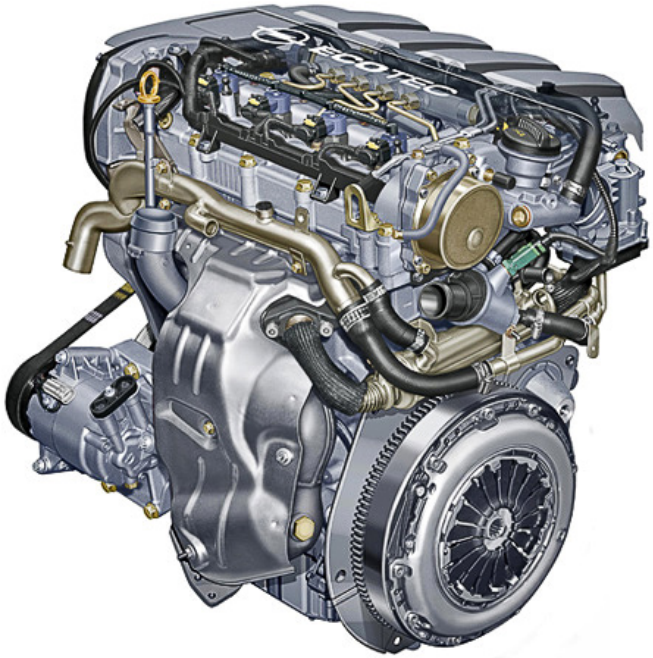
- Сначала надо «посмотреть на задачу», понять, какое значение чему соответствует, изобразить это на графиках, т.е. провести разведочный анализ данных (EDA)
- У реальной задачи часто есть очень простое и эффектное решение
- Решение прикладных задач требует практики

Цели EDA

- найти «волшебные признаки»
- понять, как «меняются признаки» при изменении времени или категорий объектов
- использовать контекст (экспертные знания, нашу интуицию и т.п.)
- выявить информационные утечки
- построить простые бенчмарки (алгоритмы в несколько строчек кода, которые надёжны, интерпретируемы и т.п.)

«Ford Classification Challenge» (2008)

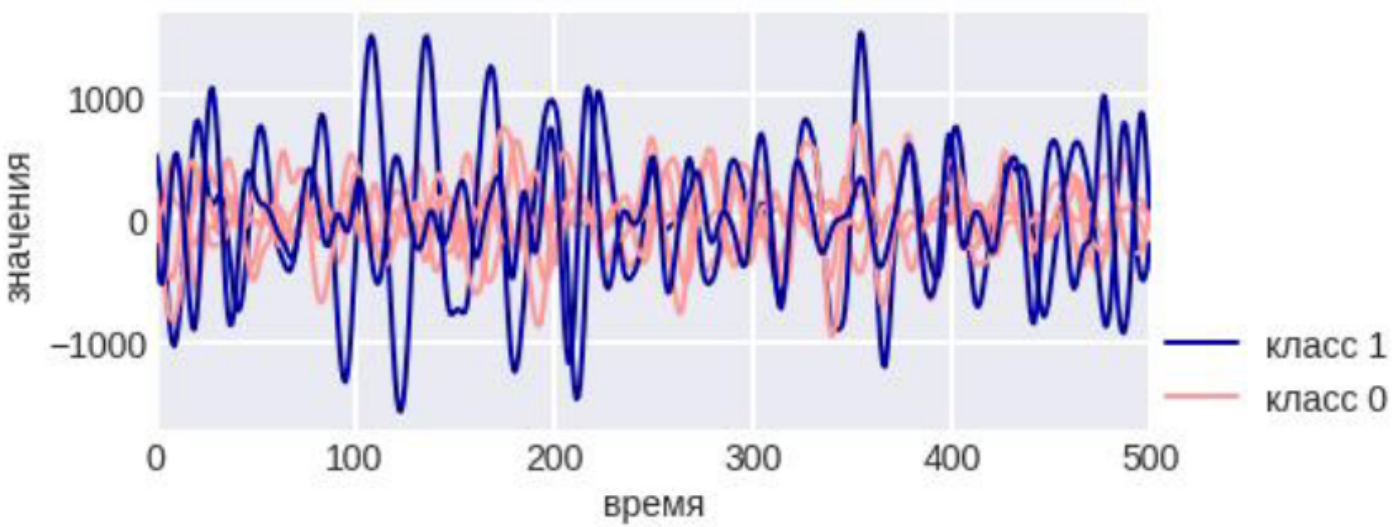
Диагностика двигателя по сигналам датчиков



http://www.timeseriesclassification.com/description.php?Dataset=FordA

Простые методы в АД

Требуется классифицировать работу двигателя как исправную (0) или неисправную (1)

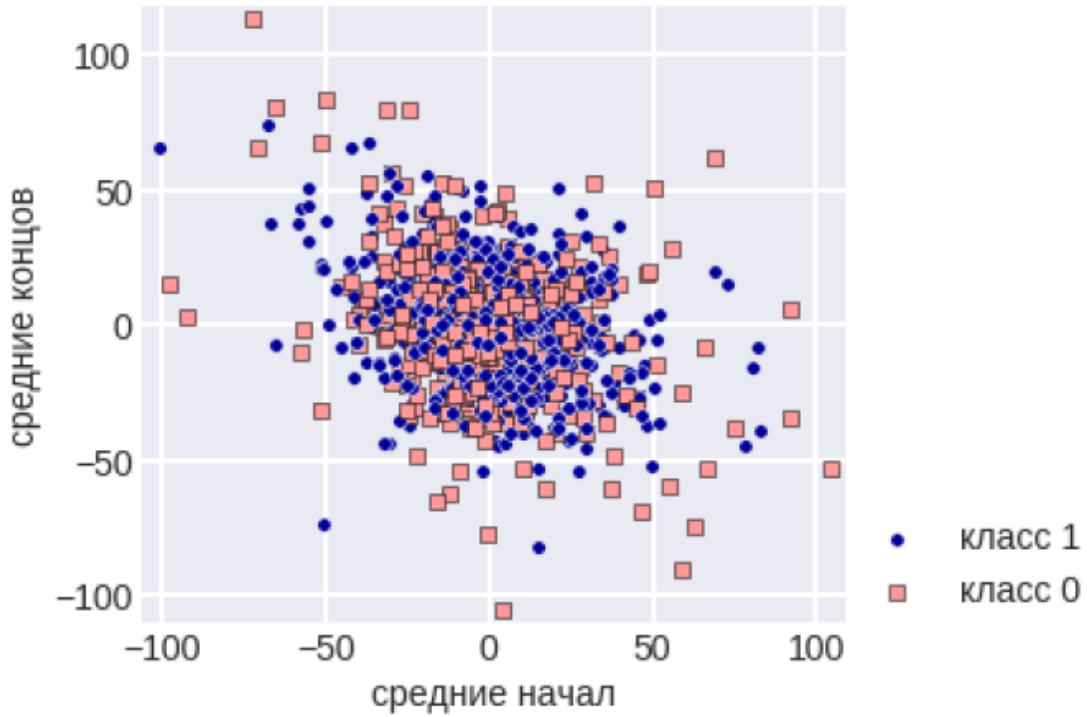


Размеры данных: 3271×500

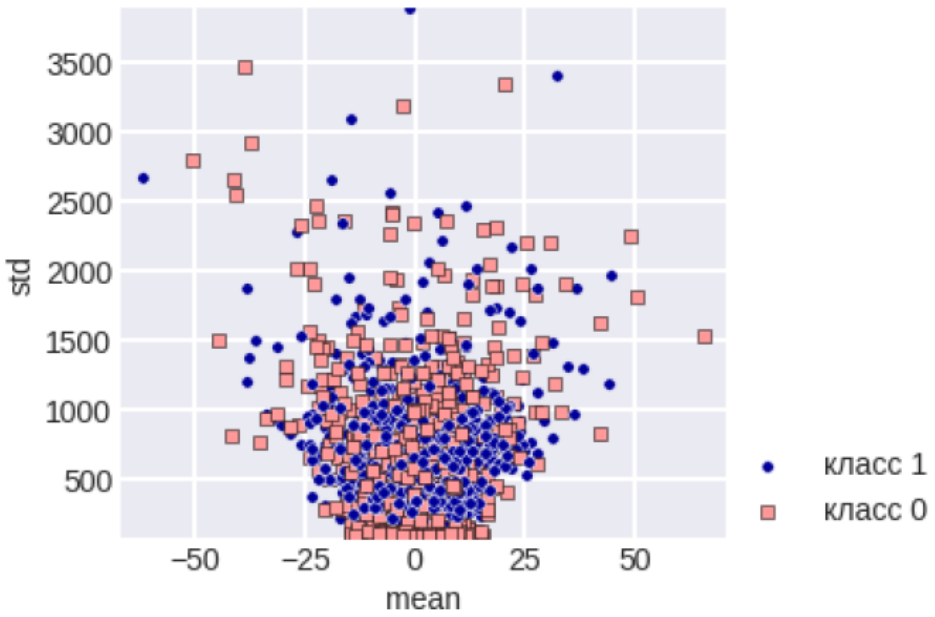
Простые методы в АД

Особенности данных

- Неоднородность
  (средние значения в начале сигнала не коррелируют со средними в конце)
- Непериодичность
- Отсутствие заметных паттернов



Пытаемся найти хороший признак

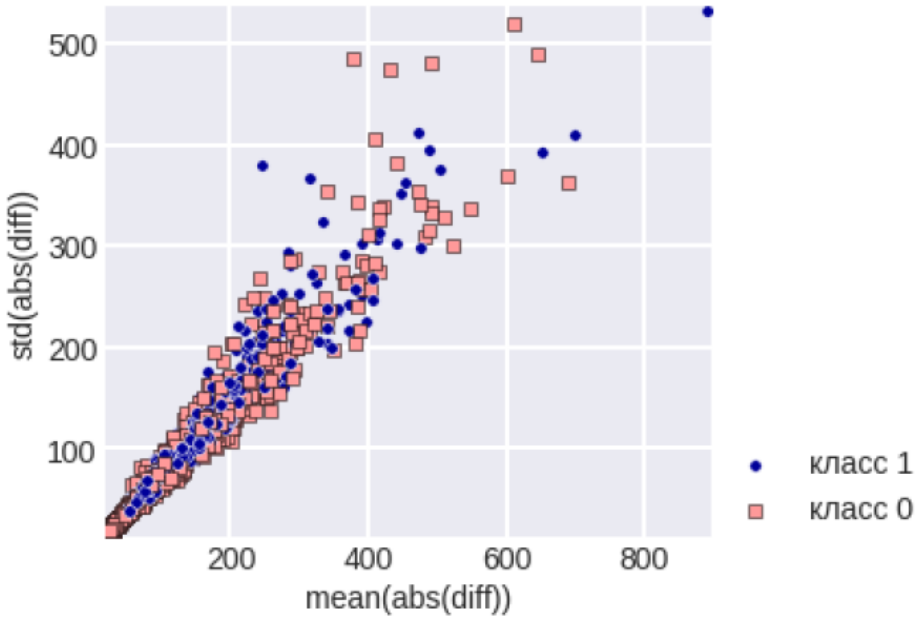


Пытаемся найти хороший признак

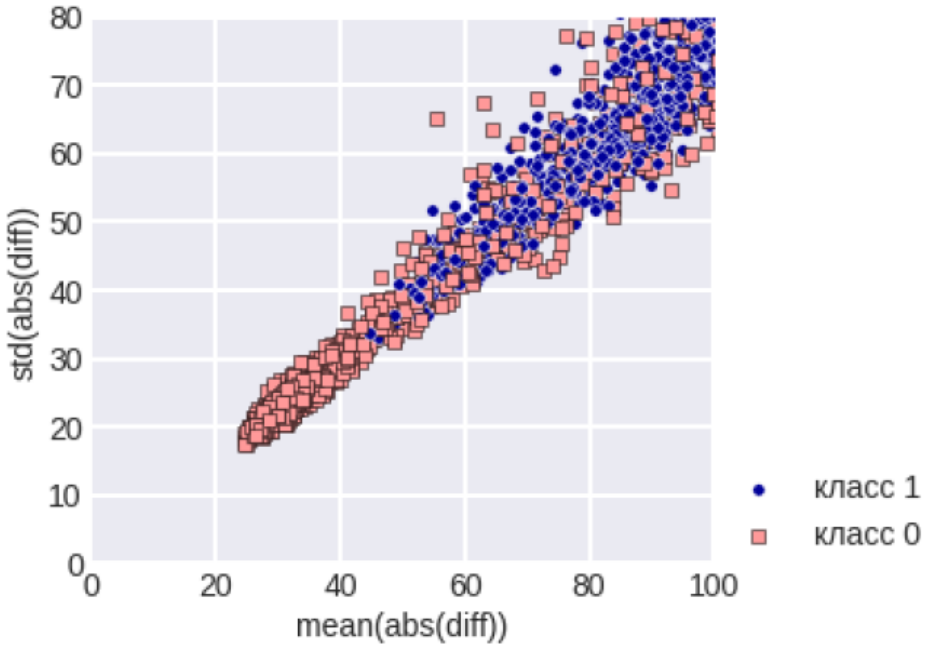
Если посмотреть на максимальные и минимальные значения сигнала, то можно предложить такую эвристику: «если максимальное значение сигнала меньше 350, то это сигнал класса ‘0‘», которая правильно классифицирует 622 сигнала обучения (из 3271)

Простые методы в АД

Пытаемся найти хороший признак



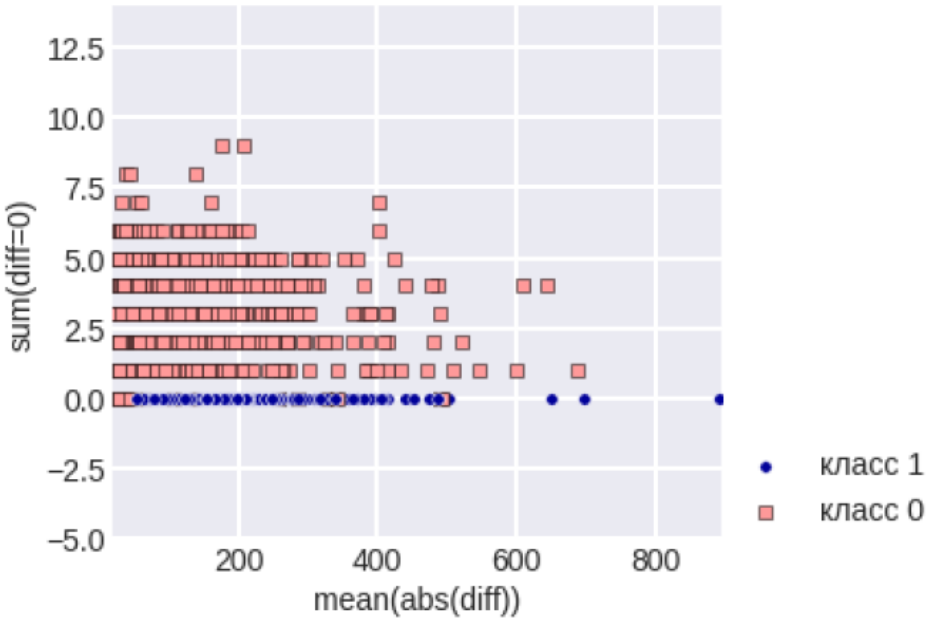
Пытаемся найти хороший признак



Увеличили изображение...

класс 0 $\Leftrightarrow$ маленькие разности

Пытаемся найти хороший признак



Максимальные значения сигналов и количества повторов соседних значений $(u_i = u_{i+1})$ в сигналах

Считаем, сколько раз значения в сигнале повторяются:

$$\frac{1}{n-1}\sum_{i=1}^{n-1} \mathrm{I}[u_{i+1} - u_i = 0] = \frac{1}{n-1}\sum_{i=1}^{n-1} \mathrm{I}[u_{i+1} = u_i]$$

Обобщение

Раз уж мы «нащупали» такой неплохой признак, попробуем его обобщить

Первое естественное обобщение — число незначительно отличающихся соседних точек:

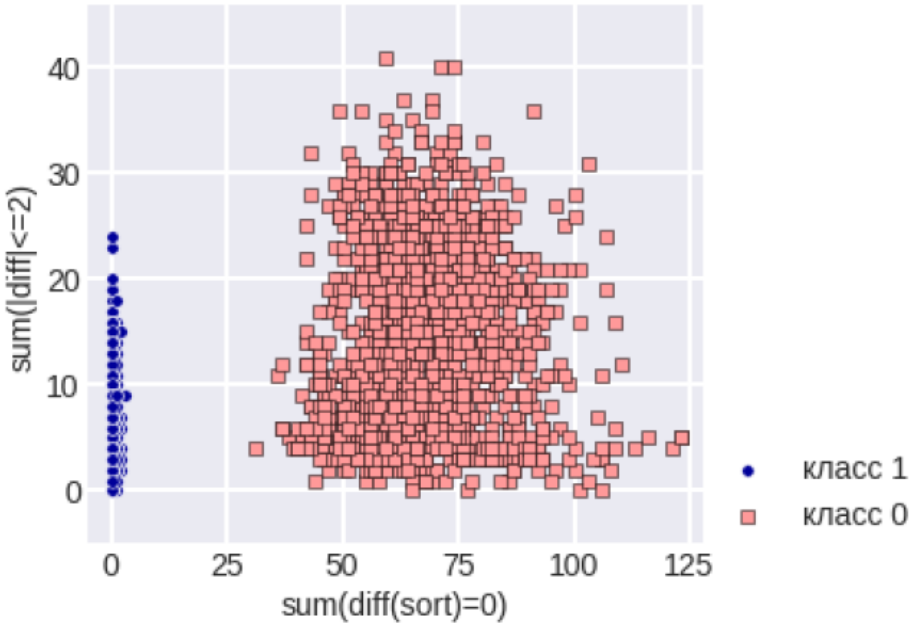
$$\frac{1}{n-1}\sum_{i=1}^{n-1}\mathrm{I}\left[|u_{i+1}-u_i|<\varepsilon\right]$$

Второе — считать совпадения не соседних значений, а вообще все совпадения в сигнале:

$$\frac{1}{n-1}\sum_{i=1}^{n-1}\mathrm{I}\left[u_{i+1}^{\mathrm{sorted}}=u_i^{\mathrm{sorted}}\right]$$

Простые методы в АД

Второе обобщение «работает», причём на 100% и на тестовой выборке



Решение задачи

```
2·(np.sum(np.diff(np.sort(train.values, axis=1), axis=1) == 0, axis=1) < 20) – 1
```

34 символа (MatLab)

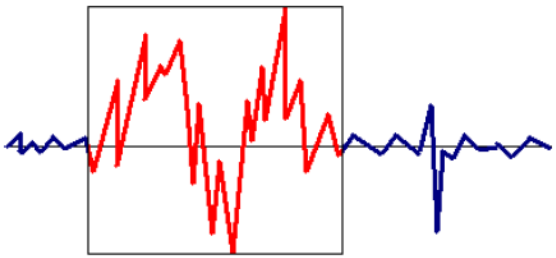
```
2·((sum(diff(sort(X'))==0)<20)'–1)
```

Итог

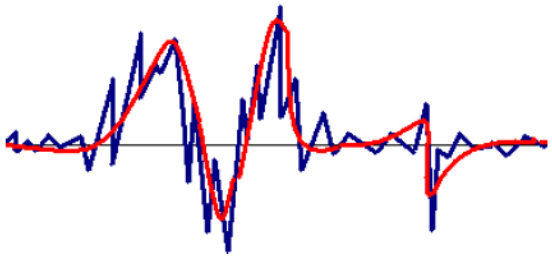
Поиск хороших признаков для сигналов в виде

Операторы первого типа

Подсигнал

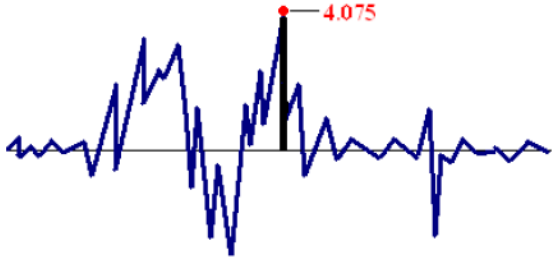


Сглаживание

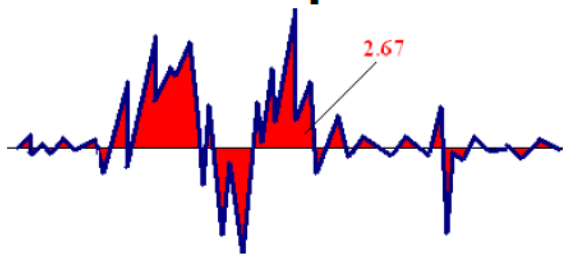


Операторы второго типа

Максимум



Интеграл



Утечки в данных

Задача определения реакции пользователя на рассылку: откликнется (1) или нет (0). Упрощённо обучающая таблица выглядела так:

клиент			услуга			статистика		
пол	id	регион	цена	скидка	категория	сколько предложений	сколько успешных	y
M	113	12	1100	0.2	17	5	2	0
Ж	078	13	1100	0.2	17	5	1	0
M	112	09	1200	0.0	16	7	0	0
M	111	07	1200	0.0	16	9	2	1

Первое что надо сделать, когда есть информация о действиях клиента — вычленить отдельных и взглянуть на них. В этой задаче клиент идентифицировался уникальной парой (id, регион)

Простые методы в АД

Упорядочим подтаблицу, содержащую информацию об одном клиенте по признаку «сколько предложений», то получим

пол	id	регион	цена	скидка	категория	сколько предложений	сколько успешных	y
	113	09				0	0	0
	113	09				1	0	0
	113	09				2	0	1
	113	09				3	1	1
	113	09				4	2	0
	113	09				5	2	?

Видно, что если предложение было успешным ($y = 1$), то число успешных предложений увеличивалось на единицу в следующей строке $\Rightarrow$ y восстанавливается по числу предложений и числу успешных. Это следует просто из названий признаков

Простые методы в АД

Похожую утечку повторила компания WikiMart, когда устраивало своё соревнование. Тогда в данных был признак «число страниц в сессии», «номер страницы по порядку посещения», нужно было определить, является ли текущая страница последней в сессии пользователя...

Сдвиг в данных

Часто в данных наблюдается проблемы сдвига распределения (data shift, target shift, covariate shift):

Что такое Data Shift

https://habr.com/ru/company/yandex/blog/568672/

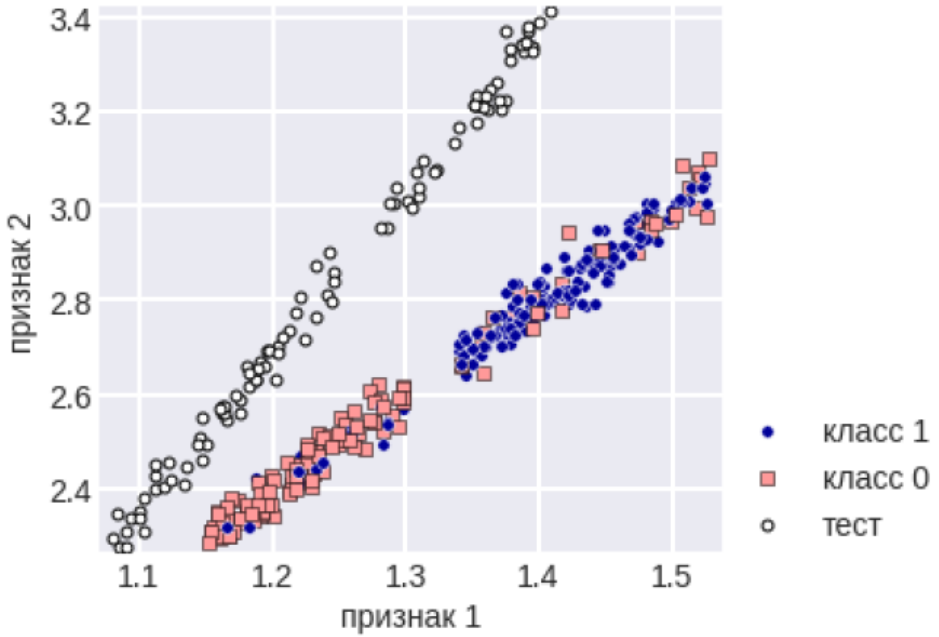
В начале 2000х проводилась серия соревнований «Brain Computer Interface», в которых предлагались задачи анализа сигналов головного мозга. Посмотрим на задачу бинарной классификации кортикограмм

Простые методы в АД

278 сигналов $\times$ 64 электрода $\times$ 3000 замеров (3 секунды с частотой 1000Гц)

- сигналы только с одного электрода
- хорошие признаки $\sim$ скорость изменения сигнала

Особенность — нестабильность признаков



два признака

$$\frac{1}{m-1}\sum_{i=1}^{m-1}|u_{i+1}-u_i| \qquad \frac{1}{m-1}\sum_{i=1}^{m-1}(u_{i+1}-u_i)^2$$

Простые методы в АД

В признаковых пространствах тест «смещается», но можно понять как:

тестовая выборка лежит в стороне, но по форме и некоторому зазору между двумя облаками точек она очень похожа на обучающую

Итоговое решение:

- сглаживание сигнала
- усреднение скачков по окрестности

$$\frac{1}{m-k}\sum_{i=1}^{m-k}\left(\max\{u_i,\ldots,u_i+k\}-\min\{u_i,\ldots,u_i+k\}\right)\geq\lambda$$

Порог λ выбран не с помощью скользящего контроля на обучении (как принято), а просто «по картинке»

kNN жив!

Не всегда задачи решаются сложными методами
В конкурсе Recommender System Challenge одна из задач: рекомендовать новые видео (для них нет статистики просмотров) для нового пользователя (для которого нет истории поведения, но есть информация о первом просмотренном видео) на ресурсе видеолекций

Простая идея: давайте синтезируем «хорошую» метрику на множестве видео-лекций и решим задачу обычным ближайшем соседом (к уже просмотренной лекции порекомендуем самые похожие из новинок)

Простые методы в АД

Легко придумать метрики для частей описания видео-лекции:

- Сравнение категорий видео (хэмингово расстояние — совпадают или нет)
- Сравнение авторских коллективов (косинусная мера сходства на характеристических векторах авторов)
- Сравнение языков (хэмингово расстояние — совпадают или нет)
- Сравнение названий (любая метрика над текстами)
- Сравнение описаний (любая метрика над текстами)
- и т.п.

Итоговую метрику в простейшем варианте можно искать как линейную комбинацию перечисленных базовых

https://bijournal.hse.ru/data/2012/05/29/1252471276/5.pdf

Ссылки

Лекция «Шаманство в анализе данных»
http://alexanderdyakonov.narod.ru/lpotdyakonov.pdf

Как бенчмарк попал в призы
https://dyakonov.org/2018/12/23/

Post scriptum

Делаем проект по машинному обучению на Python Оригинал
Перевод
https://habr.com/ru/company/nix/blog/425253/
https://habr.com/ru/company/nix/blog/425907/
https://habr.com/ru/company/nix/blog/426771/

Плохое качество кода (в том числе на Kaggle)
Название переменных

PyTorch и TensorFlow: отличия и сходства фреймворков
https://neurohive.io/ru/tutorial/pytorch-vs-tensorflow/

Александр Фонарев — Подводные камни Data Science проектов
https://youtu.be/v7ULptkbtDQ