N* Новосибирский государственный университет

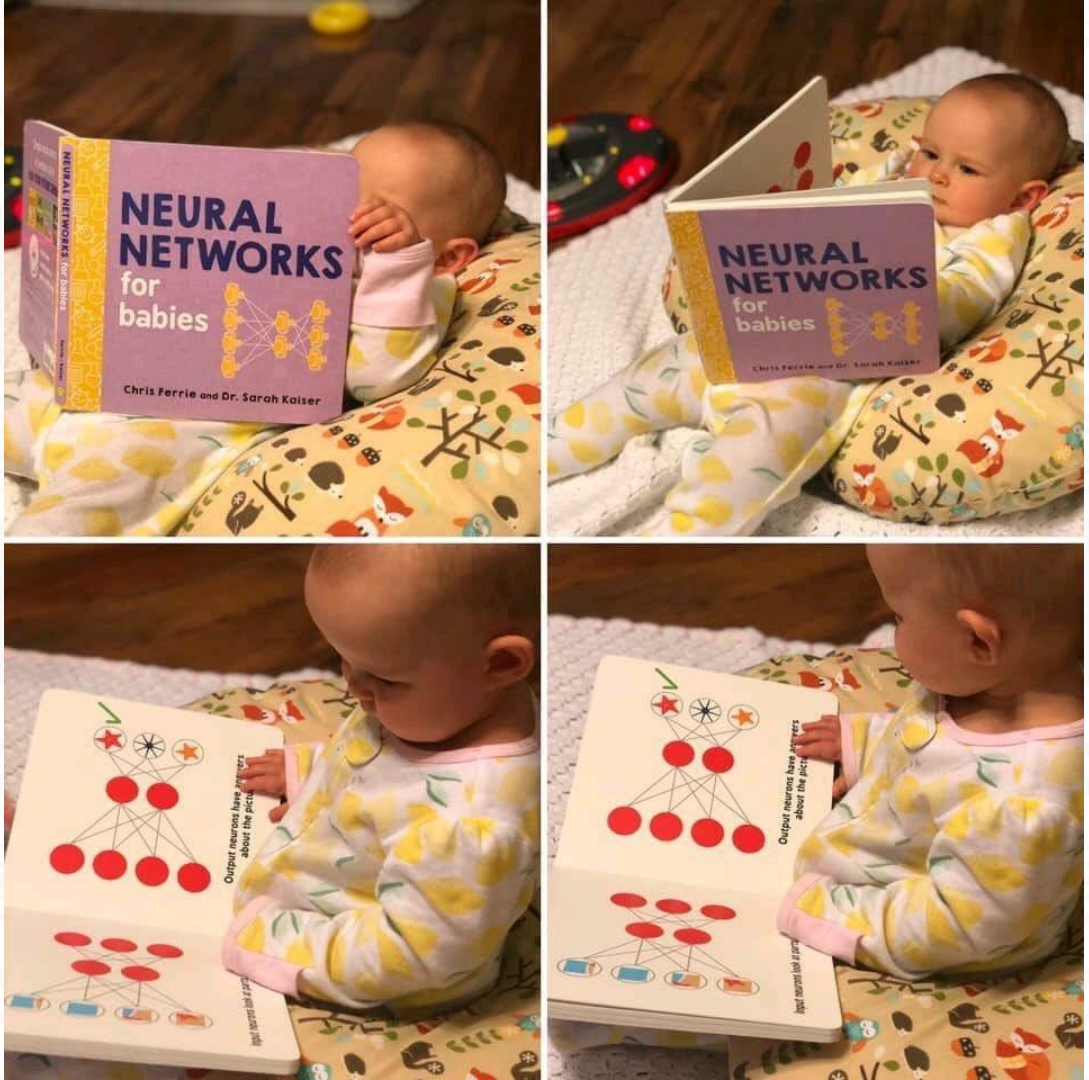
*НАСТОЯЩАЯ НАУКА

Нейронные сети

Кугаевских А.В.
К.т.н., доцент кафедры КТ НГУ

ЛИТЕРАТУРА

- Хайкин С. Нейронные сети: полный курс, 2-е изд., 2006
- Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение, 2-е изд., 2018


NEURAL NETWORKS
for babies
Chris Ferrie and Dr. Sarah Kaiser
NEURAL NETWORKS
for babies
Output neurons have features about the pictures
Output neurons have features about the pictures


ПИТОНИСТАМ МАЛО ПЛАТЯТ...
СТАНОВИСЬ ДАТА САЕНТИСТОМ КАК Я
ТЫ ЧЁ ПЁС, Я МАТЕМАТИК
ЭТО ЗАЧЕМ?
БУДЕШЬ РАЗРАБАТЫВАТЬ ИСКУССТВЕННЫЙ ИНТЕЛЕКТ ЗА 300К/СЕК
ТАК ТЫ ЖЕ ПРОСТО РАНДОМНО ПОДБИРАЕШЬ КОЭФФИЦИЕНТЫ ПОКА КРОСС-ВАЛИДАЦИЯ НЕ ДАСТ НОРМАЛЬНЫЙ РЕЗУЛЬТАТ


T-Mobile
2:38 PM
100%
“Cortana, get me today’s movie times”
Who’s Cortana?
“Oops I meant Siri”
Who is Cortana?
“Please get me the movie times”
tap to edit
Maybe you should ask Cortana for the movie times
pikabu.ru

МАШИННОЕ ОБУЧЕНИЕ

- Говорят, что компьютерная программа **обучается** на **опыте E** относительно некоторого класса **задач T** и **меры качества P**, если качество на задачах из T, измеренное с помощью P, возрастает с ростом опыта E

ЗАДАЧА

- Классификация
  - Классификация при отсутствии некоторых данных
- Регрессия
- Машинный перевод
- Структурный вывод
- Обнаружение аномалий
- Синтез и выборка
- Шумоподавление
- Кластеризация

* ОБУЧЕНИЕ

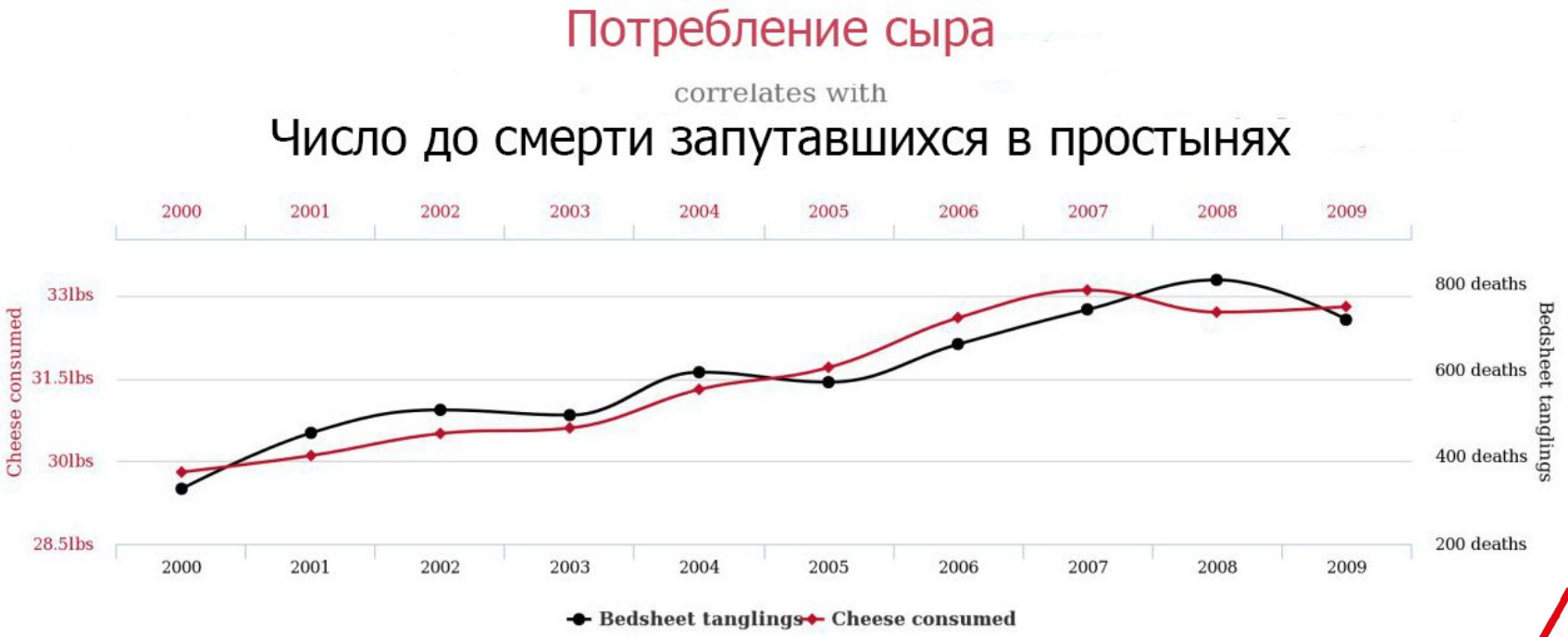

И ЭТО твоя система машинного обучения?
Ага! Высыпаешь данные в эту большую кучу линейной алгебры, а потом с другой стороны собираешь ответы.
А если ответы неверные?
просто перемешай кучу, пока они не станут выглядеть правильно.
DATA
ANSWERS

ПРИЧИНЫ НЕУДАЧ

- Ошибочная цель (неточная, неправильная)
- Ложные корреляции
- Накопление шума
- Технологические ошибки, неправильные запросы
- Данные не полны и/или загрязнены
- Не интерпретируемые модели
- Невоспроизводимые результаты
- Нет реальных данных
- Ошибки в архитектуре

* ЗАНЯТНЫЕ СОВПАДЕНИЯ

* ЗАНЯТНЫЕ СОВПАДЕНИЯ

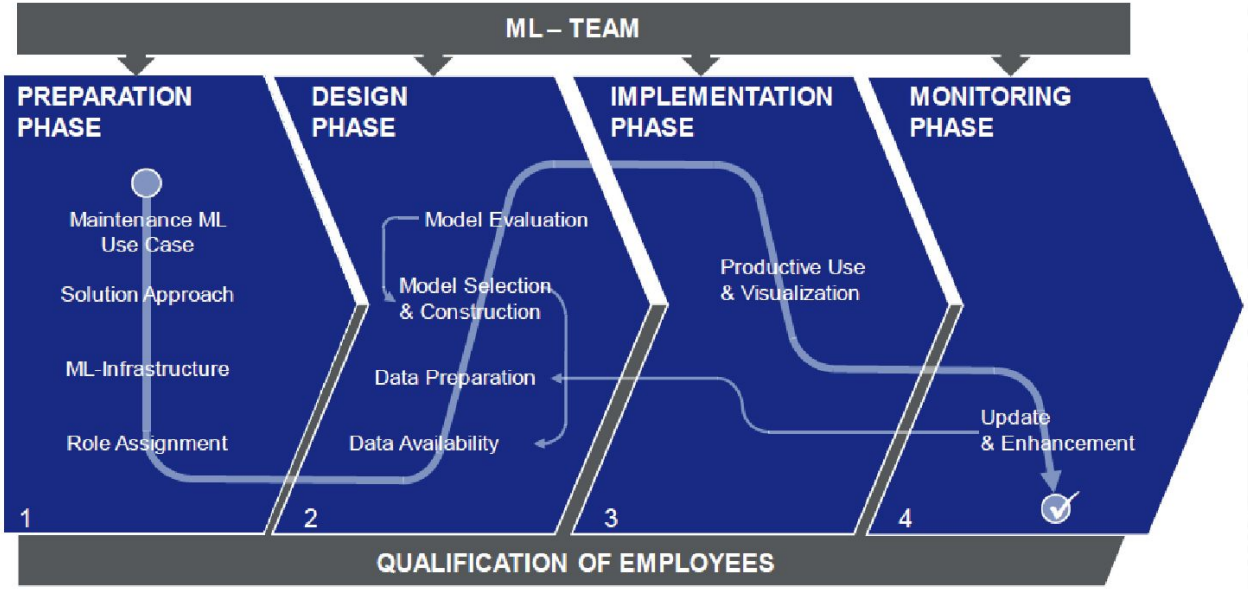


РАБОТА СПЕЦИАЛИСТА ПО ДАННЫМ

* CRISP-DM

Business Understanding/ Бизнес-анализ	Data Understanding/ Анализ данных	Data Preparation/ Подготовка данных	Modeling/ Моделирование	Evaluation/ Оценка решения	Deployment/ Внедрение
Determine Business Objectives/ **Определение бизнес-целей**	Collect Initial Data/ **Сбор данных**	Select Data/ **Выборка данных**	Select Modeling Techniques/ **Выбор алгоритмов**	Evaluate Results/ **Оценка результатов**	Plan Deployment/ **Внедрение**
Assess Situation/ **Оценка текущей ситуации**	Describe Data/ **Описание данных**	Clean Data/ **Очистка данных**	Generate Test Design/ **Подготовка плана тестирования**	Review Process/ **Оценка процесса**	Plan Monitoring and Maintenance/ **Планирование мониторинга и поддержки**
Determine Data Mining Goals/ **Определение целей аналитики**	Explore Data/ **Изучение данных**	Construct Data/ **Генерация данных**	Build Model/ **Обучение моделей**	Determine Next Steps/ **Определение следующих шагов**	Produce Final Report/ **Подготовка отчета**
Produkt Project Plan/ **Подготовка плана проекта**	Verify Data Quality/ **Проверка качества данных**	Integrate Data/ **Интеграция данных**	Assess Model/ **Оценка качества моделей**		Review Project/ **Ревью проекта**
		Format Data/ **Форматирование данных**			

ПРОЦЕСС РАЗРАБОТКИ РЕШЕНИЙ В МО

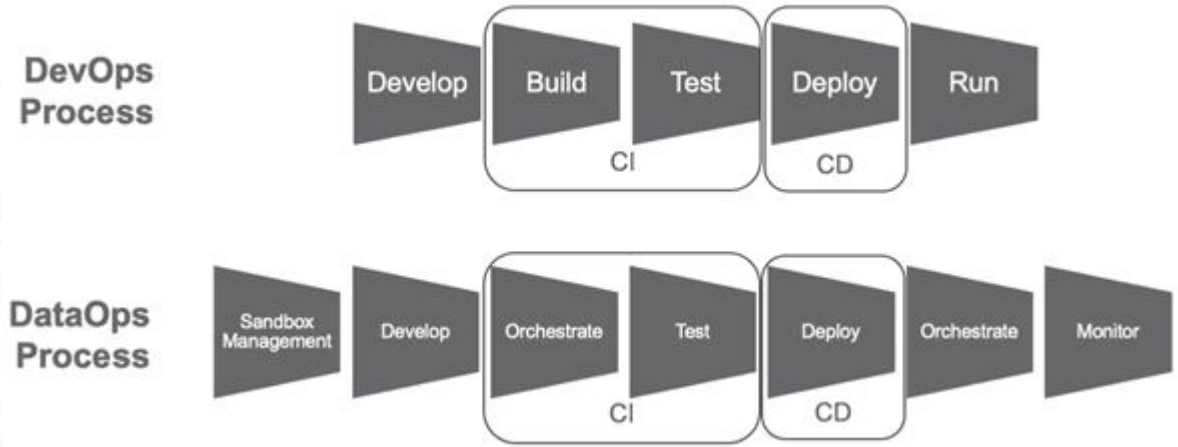


ГИБКИЕ ПРАКТИКИ

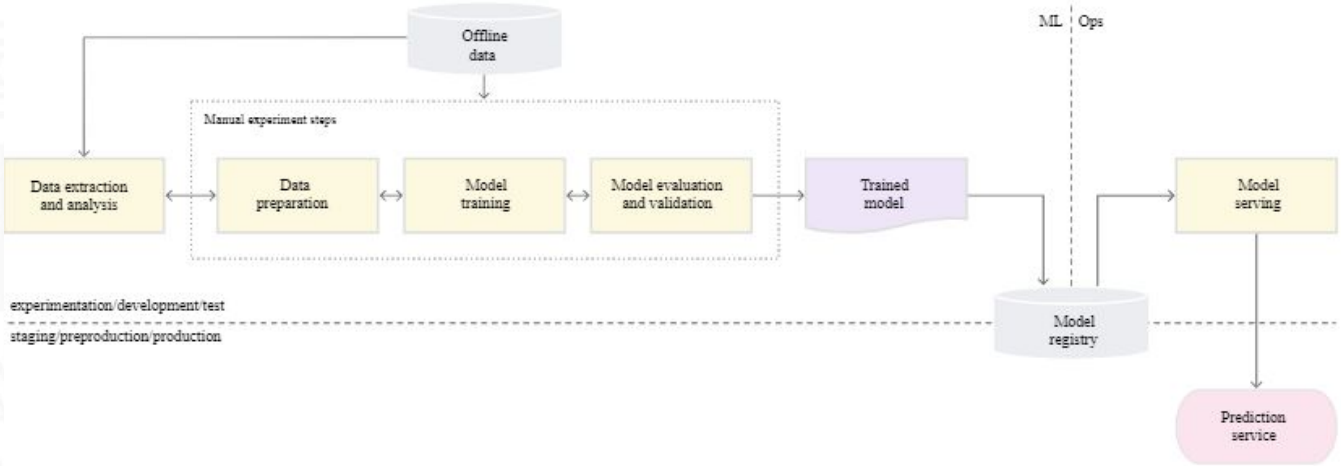
- DevOps
- DataOps
- ModelOps
  - MLOps

КОНВЕЙЕР ДАННЫХ

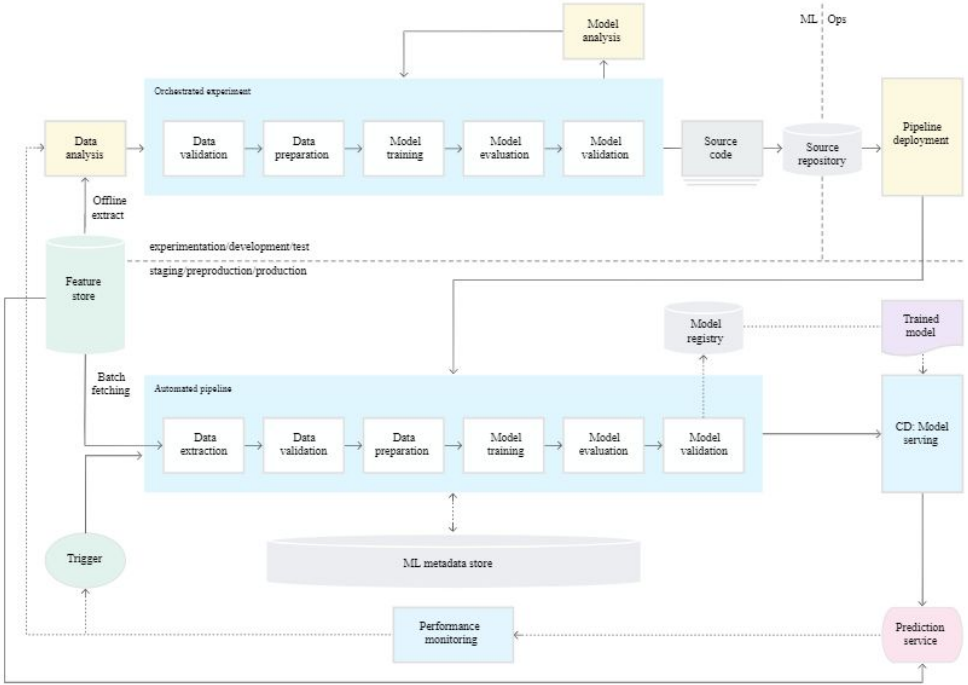
* DATAOPS


DevOps Process
Develop
Build
Test
Deploy
Run
CI
CD
DataOps Process
Sandbox Management
Develop
Orchestrate
Test
Deploy
Orchestrate
Monitor
CI
CD

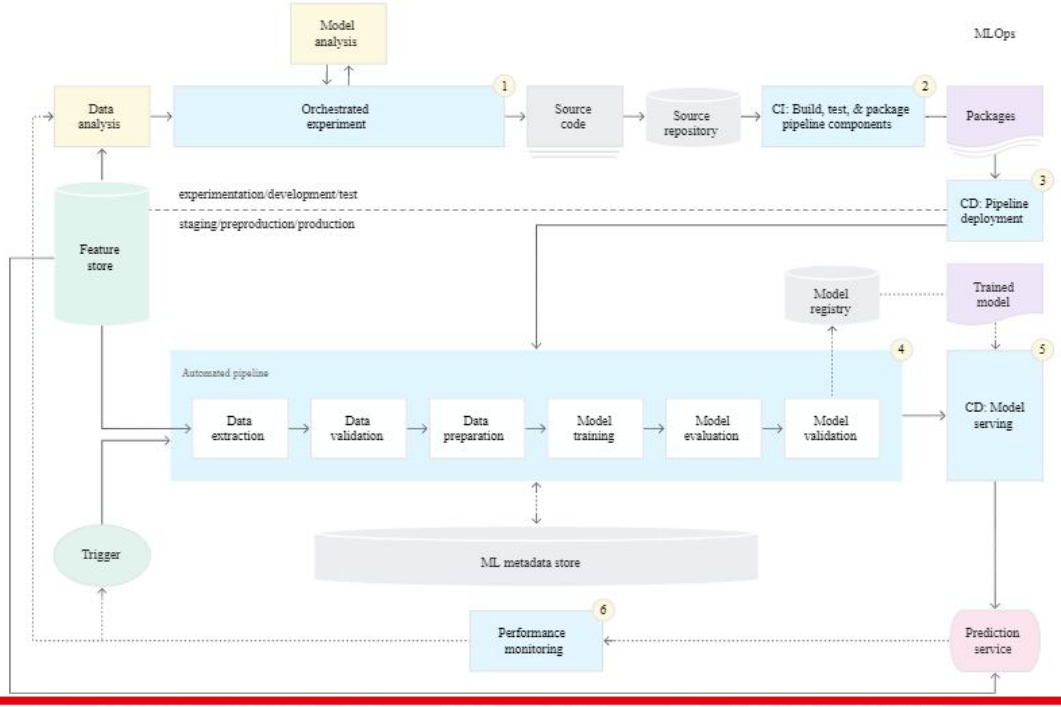
* MLOPS LEVEL 0



* MLOPS LEVEL 1



* MLOPS LEVEL 2



ПРОБЛЕМЫ С НАБОРОМ ДАННЫХ

- Входные данные должны иметь смысл
- Ошибка в коде загрузчика
- Ошибки в разметке входных данных
- Слишком много шума
- Порядок данных
- Несбалансированность классов
- Малая обучающая выборка

*ДАННЫЕ

- **Извлекайте все данные,** которые можно извлечь, но руководствуйтесь здравым смыслом.
- Оцените временной горизонт, полноту и корректность данных
- Можно ли доверять Вашим данным?
- Оцените сбалансировать данных по классам
- Достаточность размера выборки
- Избегайте синтетических данных

ПРОКЛЯТИЕ РАЗМЕРНОСТИ

- Размерность пространства решения определяется количеством признаков и их увеличение приводит к экспоненциальному росту данных.
- Это в свою очередь ведет к увеличению требуемых вычислительных ресурсов (как по памяти, так и по процессорному времени) и к риску возникновения мультиколлинеарности и переобучения

МУЛЬТИКОЛЛИНЕАРНОСТЬ

- тесная корреляционная взаимосвязь между отбираемыми для анализа признаками, совместно воздействующими на общий результат

ПРОЕКТИРОВАНИЕ ПРИЗНАКОВ

- **Инженерия признаков (feature extraction and feature engineering) – превращение данных, специфических для предметной области, в понятные для модели векторы**
- **Преобразование признаков (feature transformation) – трансформация данных для повышения точности алгоритма**
- **Отбор признаков (feature selection) – отсечение ненужных признаков**

ПРИЗНАКИ

- **Исходные**
- **Производные**
  - Агрегированные – показатели, определенные по группе (сумма, среднее, минимум, максимум)
  - Индикаторы – наличие или отсутствие характеристики
  - Отношения – взаимосвязь между двумя или более значениями данных
  - Отображения – преобразование непрерывных в категориальные

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ

- тексты – это токенизация
- изображения – извлечение краев и цветовые пятна
- дата и время– полезно вычленить выходные и праздники, дни недели
- местоположение (адрес или координаты) - извлечь плотность, средний доход по району

ОТБОР ПРИЗНАКОВ

- Знание предметной области
- Описательная статистика
- Матрица корреляций признаков – с высокой степенью корреляции подумать над удалением
- Важность – самые неважные можно удалить, на самые важные посмотреть внимательнее
- Оценить распределение - выбросы

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

- Непрерывные признаки
  - Количество
  - процент пропусков
  - минимум
  - первый квартиль ($x_{0.25}$)
  - среднее значение (μ)
  - Медиана
  - третий квартиль ($x_{0.75}$)
  - Максимум
  - стандартное отклонение (σ)
  - мощность (количество различных значений)
- Категориальные признаки
  - Количество
  - процент пропусков
  - Мощность
  - Мода
  - частота моды
  - процент моды
  - вторая мода
  - частота второй моды
  - процент второй моды

ВАЖНОСТЬ ПРИЗНАКА

- Критерий Пирсона
- Прирост информации
  - Критерий Гини
  - Gain_ratio из алгоритма C4.5

КРИТЕРИЙ РАЗБИЕНИЯ

-

КРИТЕРИЙ РАЗБИЕНИЯ

-

КРИТЕРИЙ РАЗБИЕНИЯ

- Оценка потенциальной информации, получаемой при разбиении множества T на n подмножеств. Необходим для учета атрибутов с уникальными значениями.

$$split_{info(X)} = -\sum_{i=1}^{n}\left(\frac{|T_i|}{|T|} * \log_2\left(\frac{|T_i|}{|T|}\right)\right)$$

$$Gain\_ratio(X) = \frac{Info(T) - Info_X(T)}{split_{info(X)}}$$

ПРОБЛЕМЫ С ПРИЗНАКАМИ

- Неоткалиброванные признаки
- Слишком сильная аугментация
- Применение предобработки только для одной из выборок
- Долговечность признака
- Пропуски
- Нерегулярная мощность
- Выбросы – значения, которые лежат далеко от центра распределения признака

ПРОПУСКИ

- Если доля пропущенных значений выше 60%, такой признак стоит игнорировать
- Иногда сам факт отсутствия данных может быть полезен

ЗАПОЛНЕНИЕ ПРОПУСКОВ

- **Не заполнять пропуски нулями!**
- Не применять восстановление к признакам, имеющим **более 30%** пропусков
- среднее значение или медиана, для категориальных – мода
- линейная или логистическая регрессия

МОЩНОСТЬ

- Если равна 1, полезной информации нет, удалить
- Если мощность значительно меньше количества экземпляров, можно изменить тип признака с непрерывного на категориальный

ВЫБРОСЫ

- Недопустимые – шум
- Допустимые - правильные значения, которые сильно отличаются от остальных значений признака
- Если разрыв между третьим квартилем ($x_{0.75}$) и максимальным значением заметно больше, чем разрыв между медианой и третьим квартилем, это говорит о том, что максимальное значение является выбросом.

ПОРОГИ ОТСЕЧЕНИЯ

- Задать вручную
- нижнее значение - $(x_{0.25} - 1.5 * (x_{0.75} - x_{0.25}))$ и верхнее - $(x_{0.75} + 1.5 * (x_{0.75} - x_{0.25}))$
- Для нормального распределения можно взять $\mu \pm 2\sigma$

НОРМАЛИЗАЦИЯ

- К диапазону $[low, high]$, но чувствительна к выбросам

$$a'_i = \frac{a_i - \min(a)}{\max(a) - \min(a)} * (high - low) + low$$

- Стандартизация – к диапазону $[-1,1]$

$$a'_i = \frac{a_i - \mu}{\sigma}$$

- Для отличного от нормального – среднее меняем на медиану, стандартное отклонение – на интерквартильный размах

ТЕХНИКИ ОТБОРА

- Обертка – процедура поиска, которая включает обучение и оценку модели. Начинаем с пустого множества и добавляем в него по одному признаку при условии, что он улучшает качество модели.
- Фильтрация. Набор признаков более общий, чем набор, полученный из обёртки, что приводит к меньшей способности предсказания, чем у обёртки. Однако набор признаков не зависит от модели.
  - Алгоритм Relief и его производные

N* Новосибирский государственный университет

*НАСТОЯЩАЯ НАУКА

СПАСИБО ЗА ВНИМАНИЕ!

a-kugaevskikh@yandex.ru