Машинное обучение: запрос и валидация данных

Эмели Драль

Проектная работа

Весь объем работы можно разделить на три стадии:

- **Предпроектное исследование**
- Работа над проектом
- Работа после окончания проекта

Запрос и валидация данных

1. Мотивация
2. Запрос на данные
3. Валидация данных
4. Результаты анализа

Мотивация

Мотивация

Мотивация

? Для чего требуется анализ выгрузки (sample) перед стартом проекта?

Мотивация

Валидация данных

- Уточнение запроса (или процесса) на выгрузку полного набора данных
- Анализ структуры данных, интерпретация
- Проверка на то, что данные удовлетворяют ожиданиям и ограничениям
- Проверка полноты и достаточности
- Оценка сложности процесса выгрузки

Запрос на данные

Запрос на данные

Формулировка запроса

? Для чего требуется детализированный запрос на данные?

Запрос на данные

Содержание запроса

- Объем выгрузки
- Формат данных
- Описание структуры и содержания данных
- Содержание выгрузки
- Персональные и чувствительные данные
- Предобработка данных
- Способ передачи данных

Запрос на данные

Объем выгрузки

- небольшой объем, но репрезентативная выгрузка
- последовательные данных vs случайные строки

Запрос на данные

Объем выгрузки

- 1 неделя / 1 день / 1 час
- стоит учесть частоту логирования, объем данных

Запрос на данные

Формат данных

- предпочтительно выгружать данные в том же формате, что и полный набор данных в дальнейшем
- в идеальном случае, в том же формате, в котором данные доступны в production

Запрос на данные

Описание данных

- для данных, которые передаются в виде отдельных таблиц или дампа базы требуется запросить концептуальную схему
- для любых данных требуется запросить описание таблиц и полей

Запрос на данные

Содержание выгрузки

- Целевая функция
- Переменные
- Внешние данные

Запрос на данные

Персональные данные

- Не запрашиваем
- Анонимизация
- для чувствительных данных: пост-обработка, логарифмирование

Запрос на данные

Предобработка данных

- лучше всего выгружать сырые данных без постобработки
- если по объективным причинам это невозможно, требуется также запросить описание процесса обработки

Запрос на данные

Передача данных

? Какие способы безопасной передачи данных вы можете предложить?

Валидация данных

Валидация данных

Валидация данных

? Какие риски мы хотим снизить в процессе валидации данных?

Валидация данных

План анализа

- Ревью структуры данных
- Анализ целевой функции (целевых событий)
- Анализа переменных
- Статистический (разведочный) анализ
- Визуализация данных
- Построение baseline моделей (если применимо)

Валидация данных

Ревью структуры

- Соответствуют ли данные предоставленной схеме?
- Получается ли объединить данные из нескольких таблиц по предоставленной схеме?
- Есть пропуски, коллизии в ID?
- Удается ли отличить нулевые значения от пропущенных, специалных?

Валидация данных

Анализ целевых событий

- Наличие целевой функции в наборе данных
- Распределение, баланс
- Постобратботка

Валидация данных

Анализ переменных

- Значения переменных, интерпретация
- Типы переменных
- Области значений
- Распределения
- Выбросы, шумы, ошибки
- Основные закономерности (корреляции, материальный баланс, энергетический баланс и пр.)

Валидация данных

Визуализация данных

- Визуализация целевой функции
- Распределения по классам, по сегментам
- Парные распределения

Инструменты:

- Matplotlib
- Plotly, Dash
- Seaborn
- Pandas Profiling
- Facets

etc

Валидация данных

Baseline modeling

- Константные модели
- Простые модели
- Leave-one-out

Результаты анализа

Результаты анализа данных

Валидация данных

? Стоит ли инвестировать ресурсы в подготовку отчета о результатах анализа данных?

Результаты анализа данных

Executive summary

- Краткая информация о результатах проделанной работы
- Основные выводы, риски
- Содержание отчета

Результаты анализа данных

Общая информация о наборе данных

- Источники данных, версии выгрузки и другая метаинформация
- Размеры, dateranges
- Визуализация, основные статистики
- (опционально) результаты моделирования

Результаты анализа данных

Вопросы и комментарии

- Вопросы должны требовать минимального дополнительного контекста
- Лучше всего привести весь контекст (визуализации, выгрузки данных, примеры) непосредственно рядом с вопросом

Результаты анализа данных

Основные риски

- Достаточность данных (целевые событий)
- Сезонность, цикличность, underrepresented segments
- Непротиворечивость данных
- Смещения, ошибки
- Важные источники и наличие сигнала

Результаты анализа данных

Выводы

- Пригодность набора данных для моделирования
- Готовность получить полную выгрузку в аналогичном формате
- Соответствие данных ожиданиям, сформулированным ранее
- Какие действия следует предпринять для снижения рисков, описанных выше?

Запрос и валидация данных

1. Мотивация
2. Запрос на данные
3. Валидация данных
4. Результаты анализа

Машинное обучение:
запрос и валидация данных

Спасибо!

Эмели Драль