ML System Design Interview Framework

Page • 1 backlink • uni

Могут как спрашивать про систему, которую идет разрабатывать кандидат, так и про случайную систему

Шаги

- уточнить требования (онлайн/оффлайн, трейдоффы, наличие железа, предыдущий опыт)
- перекладывать требования на условия ML-задачи
- процесс подготовки данных
- процесс разработки модели
- выбор метрик оценки качества
- процесс деплоя и обслуживания
- процесс мониторинга и инфраструктура

> На каждом шаге задаем наводящие запросы и в зависимости от ответов заостряем внимание на самом важном (о том, что, возможно, прямо сейчас болит у интервьюера)

При **уточнении требований** важно понимать

- зачем вообще система разрабатывается (деньги, реклама, прочее)
- ключевые возможности
- данные (опираемся на собственный опыт как пользователя таким систем)
- ограничения (наличие мощностей, ориентировочная мощность конечного устройства, необходимость в совершенствовании модели со временем)
- масштаб системы (число пользователей и айтемов)
- необходимая производительность

При **перекладывании на ML**

- какая ML-задача близка к продуктовой

- что приходит на вход и какой выход (аналогичные сетапы можно рассмотреть по-разному: человек → вероятность каждого события или человек + событие → вероятность)

При обзоре **работы с данными**

- с какими источниками работаем
- какие будут использованы хранилища или какие сейчас используются
- какие типы хранимых данных, схема
- какой проводить feature engineering
- как гарантировать приватность (учитывается даже просмотр сотрудниками) и отсутствие bias

При **разработке модели**

- переход: бейзлайн → простая модель → сложная модель → ансамбли
- сколько данных нужно модели для обучения
- с какой скоростью модель будет обучаться/дообучаться
- ограничения по железу
- нужно ли непрерывное дообучение

> Парадигма тетраэдра throughput-accuracy-training time-latency – к одному приближаемся, от других отдаляемся

- релевантные знания из курса машинного обучения, особенно про несбалансированность в данных и про недообучение/переобучение
- откуда брать разметку: большие модели для разметки или люди

При **оценке модели**

- какие использовать оффлайн/онлайн-метрики (может быт как что-то более академическое, так и более продуктовое, даже деньги)
- возможен ли bias в сторону определенных социальных групп

При рассуждении про **деплой и обслуживание**

- облачная/устройственная раскатка
- сжатие модели
- тестирование на проде
- пайплайн предсказания и его связь с другими частями (например, процессом обучения)

При обсуждении **процесса мониторинга** (?)

Баннер с рекламой в мобильном приложении

Цель – максимизация выручки

Упрощенная модель: показываем один раз, картинка, стоимость по клику фиксирована

Уточнение фичей: можно показывать несколько раз + есть "крестик", клик по которому является отрицательным сигналом

...