Дизайн систем машинного обучения

8. Диагностика ошибок и отказов

План курса

1) ~~Практическое применение машинного обучения~~
2) ~~Основы проектирования ML-систем~~
3) ~~Обучающие данные~~
4) ~~Подготовка и отбор признаков~~
5) ~~Выбор модели, разработка и обучение модели~~
6) ~~Оценка качества модели~~
7) ~~Развертывание~~
8) **Диагностика ошибок и отказов ML-систем — Вы находитесь здесь**
9) Мониторинг и обучение на потоковых данных
10) Жизненный цикл модели
11) Отслеживание экспериментов и версионирование моделей
12) Сложные модели: временные ряды, модели над графами
13) Непредвзятость, безопасность, управление моделями
14) ML инфраструктура и платформы
15) Интеграция ML-систем в бизнес-процессы

Естественные метки и петля обратной связи

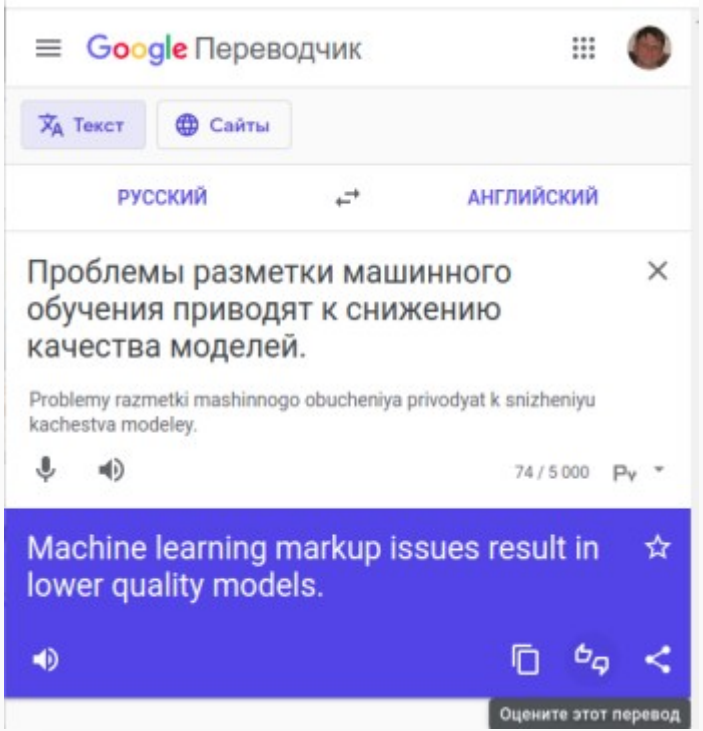
- Естественные метки (Natural Labels):
  метки, доступные непосредственно из входных данных
  (обычно вскоре после события, которое оценивала модель)
- Отложенные метки (Delayed labels):
  метки, доступные с большой задержкой
- Петля обратной связи (Feedback Loop):
  влияние выхода модели на входные данные

Естественные метки

- Система может полностью или частично оценивать качество своей работы из доступных ей данных
- Например:
  - Расчетное время прибытия
  - Прогноз спроса на поездки в такси
  - Цена акций
  - % кликов по рекламе
  - Рекомендации рекомендательной системы

Если нет естественных меток - создайте

- Добавьте возможность оценивать
  - Например, гугл переводчик
  - Поездка в такси
- По косвенным признакам
  - Больше не ищут
  - Чаще пользуются сервисом
  - Смотрят больше страниц
  - См про прокси-метрики →



Проблемы естественных меток

- Неполная выборка
  - Естественные метки могут быть доступны не для всех точек данных
- Смещение выборки (Selection Bias)
  - Например, мы не получим оценку о хороших поездках в такси

Отложенные метки (Delayed labels)

- Например:
  - погашение кредита
  - инвестиционные рекомендации
- Долго — это сколько?
  - Минуты, часы: Reddit / Twitter / TikTok
  - Недели, месяцы: мошеннические схемы с финансами
- Долго — если за время получения меток изменились:
  - Распределение входных данных
  - Распределение целевой переменной
  - Форма зависимости целевой переменной от входных данных

Временной горизонт естественных меток

- Перешел по рекламе
- Походил по страницам
- Купил
- Когда фиксируем событие?
- Раньше: ложноотрицательные
- Позже: ложноположительные, отложенная обратная связь

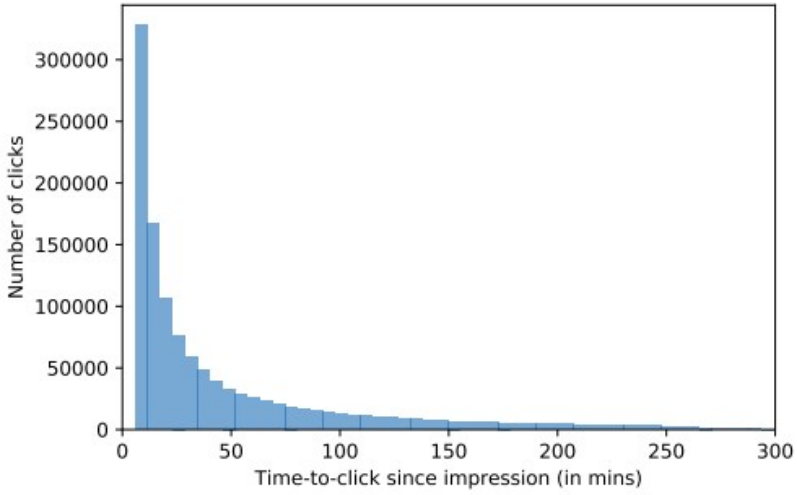


Figure 4: Distribution of time-to-click delay for training ads (longer than 5mins). Corresponds to the distribution after correcting for the CDF of the censoring distribution.

Addressing Delayed Feedback for Continuous Training with Neural Networks in CTR prediction

Отказы ML-систем

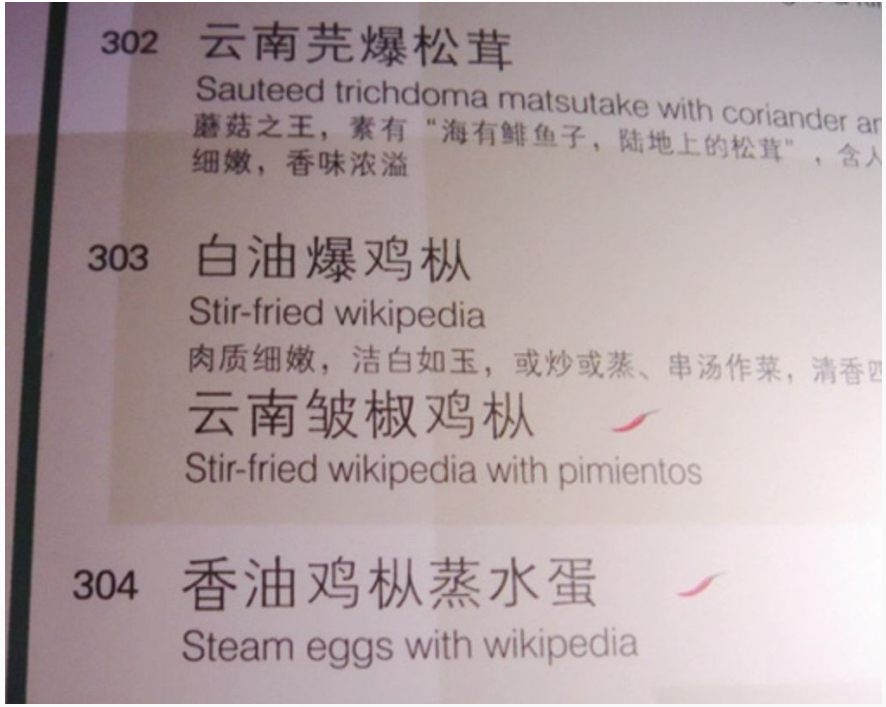
- Требования к системе формализуют через метрики
- Операционные метрики, например:
  - Средняя задержка, пропускная способность, % доступности
- ML-метрики, например:
  - Accuracy, F1, Recall, BLEU
- SLA (Service Level Agreement) →
  - SLI (Service Level Indicator) →
  - SLO (Service Level Objective) →

Пример: онлайн-переводчик

- Ввели текст, не получили перевода:
  - Операционный отказ . Фиксируем по журналам систем
- Ввели текст, получили неправильный перевод
  - ML-отказ? Необязательно.
    Мы ожидали какой-то процент неправильных переводов.
  - Отказ — если их слишком много (сколько?)

Жареная википедия

- Операционные ошибки:
  - Крик, шум, алерты
- ML ошибается тихо



Chip Huyen, sc329s.stanford.edu

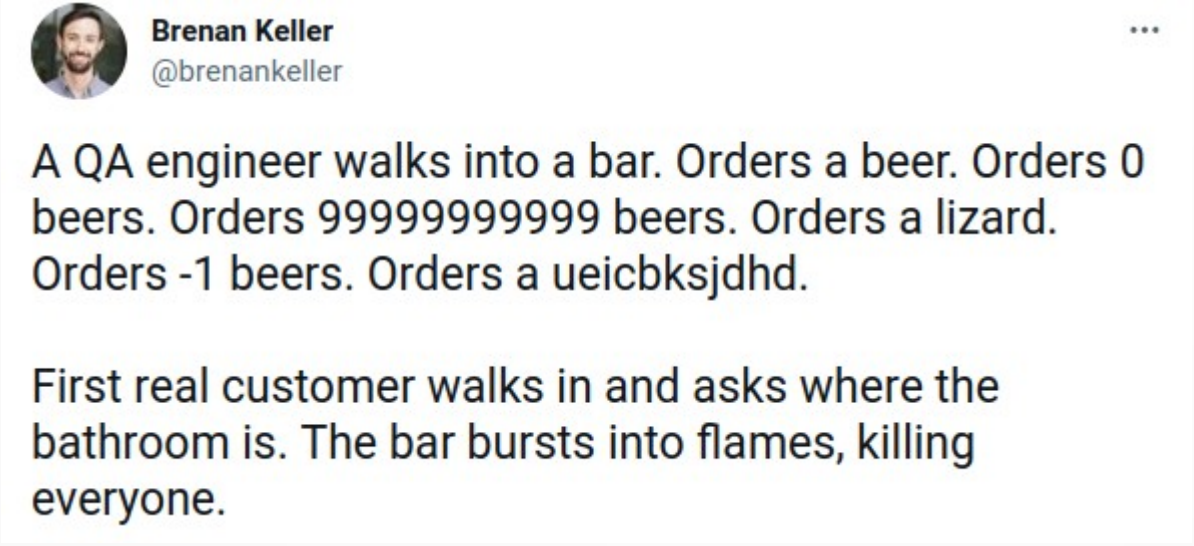
Кто тут сладкая булочка?

https://medium.com/@livewithai/the-significance-of-edge-cases-and-the-cost-of-imperfection-as-it-pertains-to-ai-adoption-dc1cebeef72c

Причины операционных отказов

- Проблемы зависимостей
- Проблемы деплоя
- Аппаратные отказы
- Сетевые проблемы (недоступность или перегрузка каналов)
- Оценочно 60% ML отказов - операционные (слайды, видео)

Если бы мы проектировали бары



Tweet from Brenan Keller

Специфичные для ML отказы

- Реальные данные отличаются от обучающих данных
- Крайние случаи
- Вырожденная обратная связь
  - Echo chamber
  - Filter Bubbles
  - Программы учатся обыгрывать другие программы
  - Пользователь может выбрать только из того, что ему показали

Отличия от обучающих данных

- Разница распределений тестовых и рабочих данных
  - Модель работает на реальных данных хуже чем, на тестовых
- Сдвиг распределений рабочих данных
  - Модель начинает со временем работать хуже
- На самом деле это могут быть ошибки разработчиков модели
  - Например, утечки разметки (dataleak)

Выбросы vs Крайние случаи

- Выбросы (Outliers)
  - Необычные входные данные
  - Обычно могут быть проигнорированы
  - Человек с необычно редким или частым пульсом
- Крайние случаи (Edge cases)
  - Необычные результаты
  - Обычно не могут быть проигнорированы
  - Человек с обычными показателями, но тяжело больной

Как тебе такое, Илон Маск?

[CVPR'20 Workshop on Scalability in Autonomous Driving] Keynote - Andrej Karpathy

Вырожденная обратная связь

- Модель: рекомендует ролик про котят
- Человек: щелкает по ролику про котят
- Модель: ага, ролик про котят хороший!
- Модель: рекомендует ролик про котят
- …

Проблемы вырожденной обратной связи

- Изначально ролики с котятами ранжировались чуть-чуть выше роликов с собачками (возможно, случайное отличие)
- Так как они выше, пользователь выбирает их чаще
- Так как он выбирает их чаще, их ранг растет
- Спустя какое-то время рекомендации становятся однородными (он смотрит только ролики про котят!)

- Проблема возникает только в реальной работе.
  Трудно выявить во время обучения

Как выявлять?

- Average Rec Popularity (ARP)
  - Средний % трафика у рекомендованного товара
- Average Percentage of Long Tail Items (APLT)
  - % редких рекомендаций (получающих менее 1% трафика)
- Насколько разнятся рекомендации для разных пользователей?
- % товаров, которые никогда не попадают в рекомендации
- Снижение CTR на первых страницах рекомендаций
- Рост CTR на вторых-третьих страницах

Борьба с вырожденной обратной связью

- Рандомизация
  - Разбавлять выдачу случайными рекомендациями
  - Собирать обратную связь
- Расширение рекомендаций
  - Добавлять не-ML рекомендации, например
    новые поступления, рекомендации экспертов, топ категорий
- Позиционные признаки:
  - Учитывать разницу CTR на разных позициях выдачи

Позиционные признаки

- Вариант — признак позиции
- Добавить в обучающие данные признак «был на 1 позиции»
- Выставлять признак в False во время предсказания

- Вариант — вес сэмпла при обучении
- Чем выше позиция — тем меньше вес обучающего примера
- Например для первых 3 позиций 0,6 0,75 0,9, для остальных 1

Сдвиг распределения

- Наша модель аппроксимирует зависимость F: X → Y
- Covariate shift
  - Распределение X изменилось, F не изменилась
- Label Shift
  - Распределение Y изменилось, F не изменилась
- Concept Drift
  - Изменилась форма зависимости F

Covariate Shift

- Наша модель аппроксимирует зависимость F: X → Y
- Распределение X изменилось, F не изменилась
- Часто Covariate Shift влечет за собой Label Shift
- Взвесьте обучающие данные по степени близости к рабочим
- Rethinking Importance Weighting for Deep Learning under Distribution Shift →

Label Shift

- Наша модель аппроксимирует зависимость F: X → Y
- Label Shift
  - Распределение Y изменилось, F не изменилась
- Иногда — просто следствие Covariate Shift
- Проверьте баланс классов
- Возможно, взвесьте классы

Concept Drift

- Наша модель аппроксимирует зависимость F: X → Y
- Concept Drift
  - Изменилась форма зависимости F
- Ситуация на рынке изменилась, например
- Может быть цикличным/сезонным изменением
- Может быть временным эффектом какого-то события

Как меняются данные

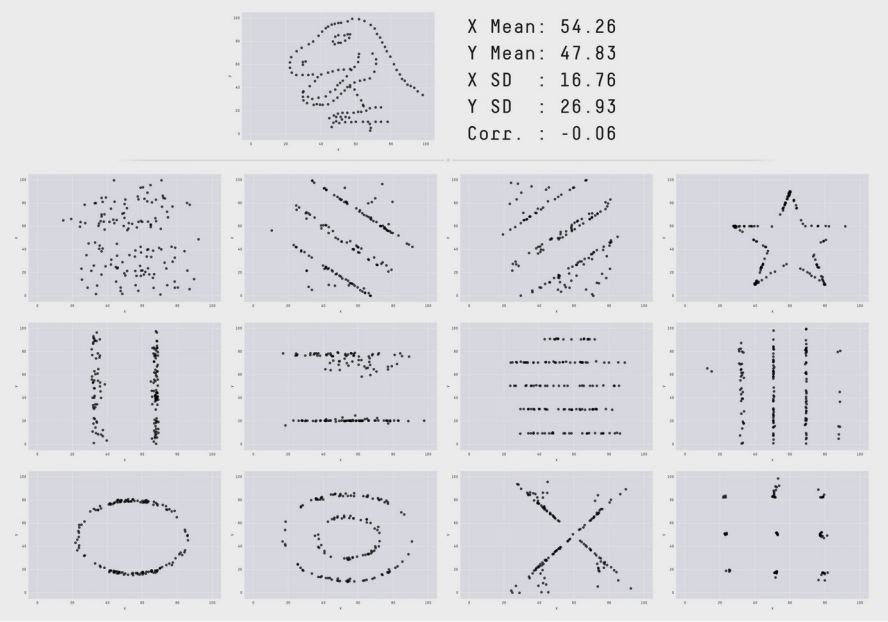
- Признаки меняются
  - Добавляются
  - Удаляются
  - Меняется схема данных
- Разметка меняется
  - Появляются новые классы
  - Изчезают классы
  - Меняется смысл меток. Например, BMI 27.8 -> 25

Временное окно важно

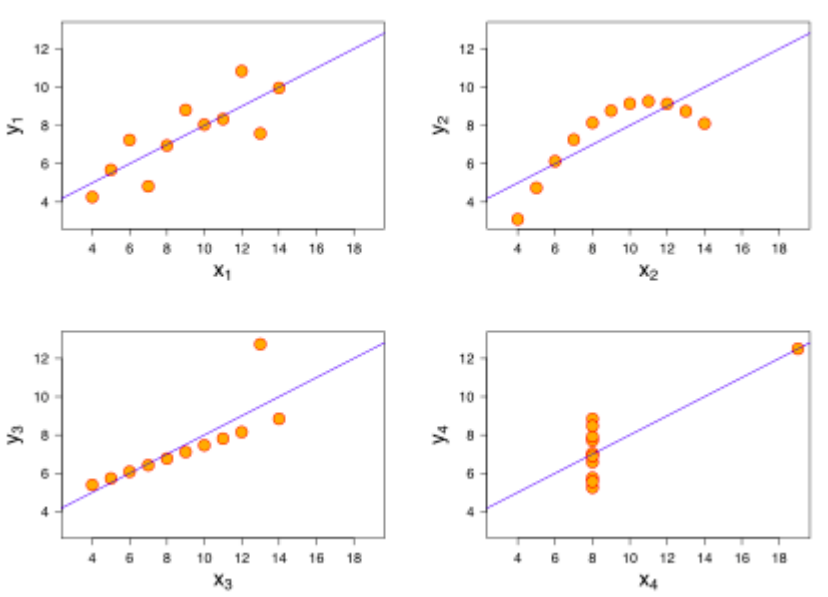
https://news.alphastreet.com/why-you-should-probably-sit-through-the-next-market-crash-2018-prediction/

Как сравнить распределения

Описательные статистики- среднее, дисперсия, квантили



Same Stats, Different Graphs



Anscombe's quartet

Как сравнить распределения

- Тест Колмогорова-Смирнова
- Тест Хи-квадрат
- Population Stability Index PSI строго → и на пальцах →
- См. блог EvidentlyAI «Which test is the best?» →
- How to Compare Two or More Distributions → и ее перевод →
- alibi-detect →
- evidently.ai →
- Подход ВТБ →

Сдвиги разные

- Резкие сдвиги распределений заметить легче, чем постепенные
- Сдвиг только у новых пользователей можно заметить когортным анализом
- Сдвиг у тех же самых пользователей можно заметить, сравнивая их поведение с историей

Что делать со сдвигом данных

- Если новых данных много — переобучите модель
- Если новых данных мало — отберите из старых данных похожие на новые
- Или взвесьте старые данные степенью похожести на новые
- Если модель переобучить нельзя (долго, дорого)
  - Скорректируйте пороги
  - Обучите вторую модель исправлять ошибки первой

Мониторинг vs Наблюдаемость

- Мониторинг (monitoring)
  - Процесс сбора, анализа и хранения метрик, которые могут помочь нам определить, что что-то пошло не так
- Наблюдаемость (observability)
  - Свойство системы, которое позволяет нам исследовать ее работу
  - Наблюдаемость обеспечивается на этапе создания системы
  - Существующая система может быть доработана
- Мониторинг основан на наблюдаемости

Еще раз про SLA

- SLI — что мы можем измерить
- SLO — какой уровень метрик нас устраивает
- SLA — какой уровень ошибок неприемлим
- SLA > SLO

Операционные метрики

- Задержка: не более 200 мс
- % запросов с кодом ответа 2XX: не менее 99% за 30 минут
- Доступность: 99.9% (9 часов в год, между прочим)
- Обычно их просто измерить
- Их часто включают в SLA

ML-метрики — что мониторить?

- Accuracy и проч
  - Зависит от задержки в обратной связи
  - Собирайте так много обратной связи, как можете
- Распределение предсказаний
  - Обычно выходная размерность мала. Легко считать статистику, проверять распределения, делать стат.тесты
  - Обычно изменение распределения предсказаний означает изменения во входных данных

ML-метрики — что мониторить?

- Accuracy и проч
- Распределение предсказаний
- Распределения признаков
  - Great Expectations →
  - Pydantic →
  - TensorFlow Data Validation →

Проблемы мониторинга

- Много данных, большая нагрузка на систему
- Мониторинг может удвоить нагрузку на ваши сервера
- И на ваш кошелек
- Alert Fatigue — когда люди начинают игнорировать сообщения
- Схема данных меняется — нужно переделывать валидаторы

Из чего состоит мониторинг

- Логи. Собираем все, до чего дотянемся
  - «If it moves, we track it. Sometimes we'll draw a graph of something that isn't moving yet, just in case it decides to make a run for it» →
- Дашборды. Делаем мониторинг доступным в нужный момент
- Алерты
  - Alert Policy: когда отправляем сообщение
  - Notification Channel: куда и кому отправляем
  - Description: что включаем в сообщение
- Алерт должен быть руководством к действию
- Если алерт не требует действий, он не нужен

Дополнительные материалы

- Adversarial Validation Approach to Concept Drift Problem in User Targeting Automation Systems at Uber →
- Degenerate Feedback Loops in Recommender Systems →
- Addressing Delayed Feedback for Continuous Training with Neural Networks in CTR prediction →
- Beyond NDCG: behavioral testing of recommender systems with RecList →

Все будет в телеграм-канале