Подготовка к кексу

В течение одного из следущих семенаров вам предстоит решить кекс. То есть рассказать о том как бы вы с помощью методов машинного обучения попробовали бы решить жизненную проблему. Откуда бы вы брали данные, какие бы модели строили и тп. Эта небольшая pdf-ка призвана вам помочь немного привести своё кексовое мышление в порядок и вспомнить о каких жизненных ситуациях мы говорили на семинарах. Несмотря на то, что по сравнению с предстоящим кексом, они были довольно мелки, в них были полезные мысли.

Итак, на семинарах мы с вами прошлись по основным задачам машинного обучения и посмотрели на то, как они естественным образом возникают в маркетинге. Давайте тезисно вспомним о том, что это были за задачи в идеологическом плане.

Кластеризация

Кекс про торговлю подарками

Британский интернет-магазин подарков собрал данные по своим транзакциям за 2010—2011 годы. В основном магазин работает с оптовиками. Он хочет сегментировать своих клиентов по их характеристикам и более точечно проводить различные рекламные кампании. Нужно ему помочь!

Наши рассуждения:

- Можно попробовать сегментировать клиентов стандартными приёмами, основанными на нашей экспертности. Например, по географическому положению, объёмам сделок, сделать RFM-сегментацию и тп.
- Можно попробовать задействовать всю мощь машинного обучения! Например, в нашем датасете содержится огромное количество описаний товаров. Давайте обработаем эти тексты и на их основе проведём кластеризацию. Скорее всего, каждый оптовый покупатель концентрируется на каких-то определённых разновидностях подарков. Кластеризация по описаниям поможет нам выявить разновидности и настроить нашу рекламную кампанию, отталкиваясь от полученных по товарам кластеров.
- Подпробное описание задачи и код с картинками и решением
- **Ещё немного необязательной информации**. На паре мы выяснили, что кластеры, построенные по описаниям пересекаются. Это не удивительно, мы часто используем одни и те же слова для различных целей... К счастью есть более сложные модели, которые помогают учитывать наложение кластеров друг-на друга. Например про них и многое другое можно почитать вот тут. Это мой небольшой рисёрч. Он получился довольно длинным, очень неформальным и слегка доставляющим.

Кекс про рекламу

Международное круизное агентство ”Carnival Cruise Line”решило себя разрекламировать с помощью баннеров и обратилось для этого к вам. Чтобы протестировать, велика ли от таких баннеров польза, их будет размещено всего 20 штук по всему миру. Вам надо выбрать 20 таких локаций для размещения, чтобы польза была большой, и агентство продолжило с Вами сотрудничать.

Агентство крупное, и у него есть офисы по всему миру. Вблизи этих офисов оно и хочет разместить баннеры - легче договариваться и проверять результат. Также эти места должны хорошо просматриваться.

Наши рассуждения:

- Банеры. Нужно, чтобы их чаще смотрели. В точках, где они стоят нужны большие скопления людей. Агенство круизное, значит нам нужны туристы.
- Как найти большое скопление людей? По геолокации! Нужна база чекинов.
- Тогда мы сможем кластеризовать чекины, найти самые популярные места в окрестнсоти каждого офиса и поставить там банеры. Задача будет решена, на банеры будут смотреть, а нам дадут денег.
- Ещё можно очистить данные от чекинов людей, которые точно не являются туристами. Например, можно посмотреть какие люди чекинились в районе в разные дни. Маловероятно, что турист будет чекиниться в одном и том же месте и в понедельник и в пятницу. Скорее всего, это местный житель. Эту идею, кстати говоря, придумала Ира :) Можно придумать и другие улучшения!
- Подпробное описание задачи и код с картинками и решением

Классификация

Кекс про приложение

Мы хотим выпустить своё классное приложение под IOS, вот только беда в том, что мы не знаем как стать успешными. Срочно нужно что-нибудь придумать и разобраться с этой проблемой!

Наши рассуждения:

- На самом деле нам предстоит решать две задачи: на первых этапах нужно привлечь пользователей, а на дальнейших этапах нужно их не потерять.
- Можно было бы собрать много-много примеров успешных приложений и провальных и внимательно изучить каждый конкретный случай. Тогда бы мы могли понять какие именно вещи стоит повторить, а какие повторять не стоит. Это хорошая идея и практика, однако мы тут занимаемся машинным обучением и хотим большего!

- Можно попробовать обучить модель понимать какое приложение хорошее, а какое плохое. Тогда бы мы могли посмотреть на какие факторы ориентируется модель и при создании своего приложения обратить внимние именно на них.
- Собираем данные с Appstore. Делим все прилодения на молодые (получил мало оценок) и старые (получили много оценок). Обучаем две логистические регрессии. Сравниваем какие факторы вносят положительный вклад на первых этапах развития приложения и на последующих.
- Подпробное описание задачи и код с картинками и решением Обратите внимание, что мы довольно грубо оценили нашу модель. В реальности с интерпретацией коэффициентов и их величин нужно быть очень аккуратным. Для того, чтобы выработать эту аккуратность, в дополнение к машинке нужно заботать матстат и эконометрику.
- Машинное обучение ни в коем случае не отменяет экспертный подход, в котором мы собираем успешные и неудачные кейсы. Эти подходы в данной задаче должны сочетать друг-друга.

Регрессия

Кекс про продажи

Walmart продаёт продукты в разных регионах США. Каждый магазин содержит несколько отделов. Магазин очень хочет спрогнозировать по каждому отделу для каждого магазина объём продаж.

Зачем? Если мы привезли в магазин слишком мало товара, потребителем его не хватит. Мало того, что они не принесут нам денег, так ещё и станут к нам менее лояльными: ”Не поедем в этот магазин. Там вечно ничего нет.”

Если мы привезли в магазин слишком много товара, то возникают лишние расходы, связанные с хранением товаров, а также лишние расходы, связанные с просрочкой товаров. Хотелось бы уметь избегать всех этих лишних расходов и привозить в каждый магазин ровно столько товара, сколько у нас купят.

Ясное дело, что для разных типов товаров мы будем нести разные расходы на хранение, более того разные товары портятся с разной скоростью. В идеале было бы круто предсказывать продажи для каждой отдельной группы товаров.

Например, для овощей у нас одна модель, для телевизоров вторая, а функции потерь зависят от специфика каждого товара. На практике, скорее всего, так и делают. Мы только учимся и такое разнообразие задач нас угробит. Мы без детализации посмотрели на аггрегированную статистику Walmart. Мы оценивали регрессию, которая должна была бы предсказать продажи.

При оценке модели мы поговорили о работе с выбросами, в данном случае это праздники. Мы боролись с праздниками. В них образуются акции по уценке товаров, из-за этого возникают всплески в продажах, то есть аномалии, то есть выбросы. Мы специфицировали нашу модель так,

чтобы учесть праздники. Выбросы довольно опасная штука.

Кстати говоря, помнить про квантильную ошибку? Когда мы в прошлом семестре присваивали недопрогнозу больший вес, чем перепрогнозу? Тут можно попробовать использовать её. Потеря клиентов явно опаснее протухания продукта на складе. Подпробное описание задачи и код с картинками и решением.

Другие полезные мелочи

Ещё кексы

В Яндексе периодически проходят всякие крутые встречи, на которых люди делятся друг с другом опытом работы с данными. На одной из таких встреч обсуждали маркетинговые задачи. Рассказ о парочке маркетинговых кексов можно найти, например, вот в этой 30-минутной лекции: https://events.yandex.ru/lib/talks/6063/.

Возможно, вы почерпнёте из неё какие-то классные идеи, а потом используете их при решении своего кекса.

О том как ритейл собирает данные о нас

В кексе важно будет руководствоваться реалистичными предпосылками и источниками данных. Чтобы примерно представлять себе откуда ритейл берёт данные, здесь есть небольшой рассказ о них. В какой-то степени он дублирует лекцию выше.

Для сбора данных и аналитики сайты используют специальные сервисы. Например, Яндекс.Метрику и Google Analytics. Эти сервисы позволяют анализировать то, что люди делают на сайте, с каких страниц они приходят, откуда они приходят (из поиска, с конкретного рекламного баннера и тп), какими демографическими характеристиками они обладают (пол, возраст, география и тп).

Более того, для каждого пришедшего пользователя существуют очень разношёрстные данные о визитах: в какой последовательности он смотрел страницы, куда кликал мышью, как ей двигал и т.п. Сервисы фактически показывает полную информацию о том, что происходит на сайте, позволяют выгрузить сырые данные и на их основе обучить какие-то модели.

Предположим, что ЛОЛита два года назад зашла на сайт интернет-магазина сделать парочку покупок. Вчера она сделал это повторно. Как понять, что эти два захода принадлежат одному и тому же человеку? Обычно для этого используют систему из разных id.

Если ЛОЛита заходил на сайт, используя свой личный кабинет, мы поймём, что это один и тот же человек по его внутреннему id. Другой способ идентифицировать человека — использовать его аккаунт в google или яндексе. Если человек зашёл на сайт, был залогинен в своей почте, и на сайте стояла метрика, мы сможем его отследить.

Более слабым идентификатором является device-id устройства, с которого работает человек. Ясное дело, что сначала человек может зайти с телефона, потом с компьютера и, если он не залогинен, то система будет думать, что это два разных человека.

Самым слабым идетификатором является id, построенный на основе куки человека. Если вы почистите в браузере куки, то этот id перезатрётся, и система будет думать, что вы новый человек. Используя такую вложенную систему из адишников, мы можем понимать где выполнял действия один и тот же человек.

В оффлайн-ритейле дело обстоит немного сложнее. Когда у магазина есть два чека в базе данных, он никак не может понять принадлежат они одному и тому же человеку или нет. Чтобы как-то исправить эту ситуацию и научиться агрегировать покупки, магазины придумывают всякие ухищрения, позволяющие им накопить данные. Например, систему бонусных карт, по id которых можно понять, что чеки принадлежат одному и тому же человеку.