Машинное Обучение МФТИ

Семинар: Определение Возраста

Концепция

1. Основная идея этой серии семинаров – провести студентов от цели принести пользу компании (заработать денег) и наличиях каких-то данных к: постановке задаче, сбору данных, построению решения с использованием алгоритмов машинного обучения грамотному измерению качества и выкатке решен ия в продакшен.

2. Такой семинар предворяет контест, на реальных данных., где студенты смогут попробовать обсужденные на семинаре подходы к решению задачи.

3. Семинар обязательно должен проходить в формате диалога, студенты должны думать что все придумали сами. А задача семинариста их за ручку провести через это =)

4. Семинары будет состоять из 6 основных блоков

   (a) Вы датасаинтесты и у вас есть данные, как заработать денег?

   (b) Как свести это к задаче машинного обучения?

   (c) Откуда взять разметку?

   (d) Думаем как решить задачу (в командах), обсуждаем, находим баги (самая жирная часть)

   (e) Как измерить качество?

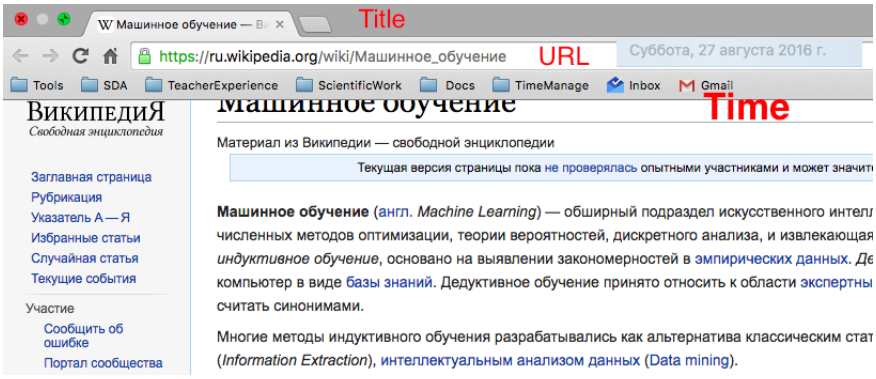
   (f) Как катить в прродакшен? (Второе дно)

План семинара

1. Вы работаете датасаинтастами в крупной интернет компании, к примеру Гугл, и у вас есть сервис аналитики.

2. Сервис аналитики, это такая штука которая ставит вам куку(en.wikipedia.org/wiki/HTTP_cookie) в браузер, а дальше видит, на какие странички ходил человек с такой кукой.

3. Что значит видит? – Есть страничка и на есть урл и заголовок(тайтл) и человек заходит на страничку в какой-то момент времени, соответственно одна транзакция, это строка `user_id, url, title, time`



4. **Вопрос + командная работа** Как заработать денег используя эти данные?

   Возможные варианта ответа:

   (a) Показывать персоанализированную рекламу

   (b) Рекомендовать хороший контент, чтобы пользователь проводил больше времени на сайте и показывать рекламу в это время

   (c) ...

5. **Убеждение** Дальше всех нужно убедить в таком сценарии, что приходят к тебе рекламу покупать, и говорят мы хотим продавать вне-дорожники от 100к$ ценой, кто их купит – ответ мужик старше 40 лет с большим доходом. Ну и во! Нам нужно уметь знать про людей пол возраста и доход. Давайте начнем с пола.

6. **Вопрос + командная работа (ответ придумывают студенты)** Как решить эту задачу методами машинного обучения?

   (a) Без учителя и набрать правил руками – не выйдет, очень плохое решение

   (b) Надо собрать обучающих данных, как???!!

   (c) А у большой компании есть почта в которой все указывают пол возраст, что там еще?

7. Супер у на есть признаки и метки, давайте к машинному обучению

   (a) Какую задачу решать классификация регрессия, можно-ли и так и так?

   (b) Какие фичи? Какие методы использовать?

   (c) Почему все что вы придумали будет работать плохо

      i. Линейные модели, Голые деревья – бесполезно, зато очень быстро учатся

      ii. Скорее всего у вас мешок урлов – это очень здоровая матрица

      iii. Бустинг и Форест на голых данных (а какой размерности у нас матрица?) будет учиться очень долго, очень много признаков – переобучитя, низкая обобщающая способность,

      iv. Сжимать признаки, хмм хороший вариант (а какой размерности у нас матрица?) чем вы ее сожмете? – Очень большая, ничем интеллектуальным типа svd и t-sne

      v. На самом деле можно, но надо учиться на подбвыборках

   (d) попробовали Бустинг и Форест – не работает, линейные модели плохо, сжимать svd не можем. Что делать?

   (e) Давайте придумаем свои способы пожать пространство

      i. Хешинг трик

ii. Отсортировать от признаки по частоте заходов от младших к старшим и разбить на интервалы
iii. Что там еще?
iv. Какие минусы у этих подходов и как их устранить –не учитываем порядок фичей и прочее RNN, word2vec

(f) как адаптировать к много-классовой классификации – возрастных груп штук 5 точно хватить

(g) сбалансирован ли датасет?

(h) Ну отлично, что-то сделали, как качество мерить, в итоге:
i. как построить процес кросс-валидации? – использовать для теста аудитория сайта где рекламу будем показывать, разные модели на разных сайтах затащат
ii. регрессия не интерпретируема давайте на классификации мериться
iii. Как мерить многоклассовую – ну да есть варианты, но давайте измерим лучше качество бинарнах задач?
iv. f1 – не объяснить боссу, точность полнота приятнее
v. давайте мерить полноту на заданном уровне точности

(i) Как катить в продакшен? Что считать в онлайне а что хранить в быстрой базе данных? Сколько у нас вообще данные? как быстро должны отвечать.

(j) Применять модель один раз за сутки, результаты в быструю базу данных, мап-редюс?