RecSys Challenge 2015: ensemble learning with categorical features

Кудрявцев Георгий

30 сентября 2015 г.

Постановка задачи

Дана последовательность кликов пользователя, совершенных во время сессии в интернет-магазине. Требуется определить:

- Собирается ли пользователь покупать предметы во время сессии.
- Если да,то какие предметы он купит.

Набор данных

Тренировочная выборка.

- yoochoose-clicks.dat – $9.2 * 10^6$ сессий
  - (Session ID, Timestamp, Item ID, Category)
- yoochoose-buys.dat
  - (Session ID, Timestamp, Item ID, Price, Quantity)

Тестовая выборка.

- yoochoose-test.dat – $2.3 * 10^6$ сессий
  - (Session ID, Timestamp, Item ID, Category)

Только 5.5% пользователей купили хотя бы один предмет.
Количество различных Item ID – 54.287
Количество различных Category ID – 347

Функционал качества

- SI – сессии, представленные в файле решения
- S – все сессии в тестовом файле.
- s – сессия в тестовом файле.
- Sb – сессии в тестовом файле, которые завершились покупкой.
- A_s – предсказанные купленные предметы.
- B_s – реальные купленные предметы.

$$Score(SI) = \sum_{\forall s \in SI} \begin{cases} \frac{|Sb|}{|S|} + \frac{A_s \cap B_s}{A_s \cup B_s} & \text{если } s \in Sb \\ -\frac{|Sb|}{|S|} & \text{иначе} \end{cases}$$

Функционал качества

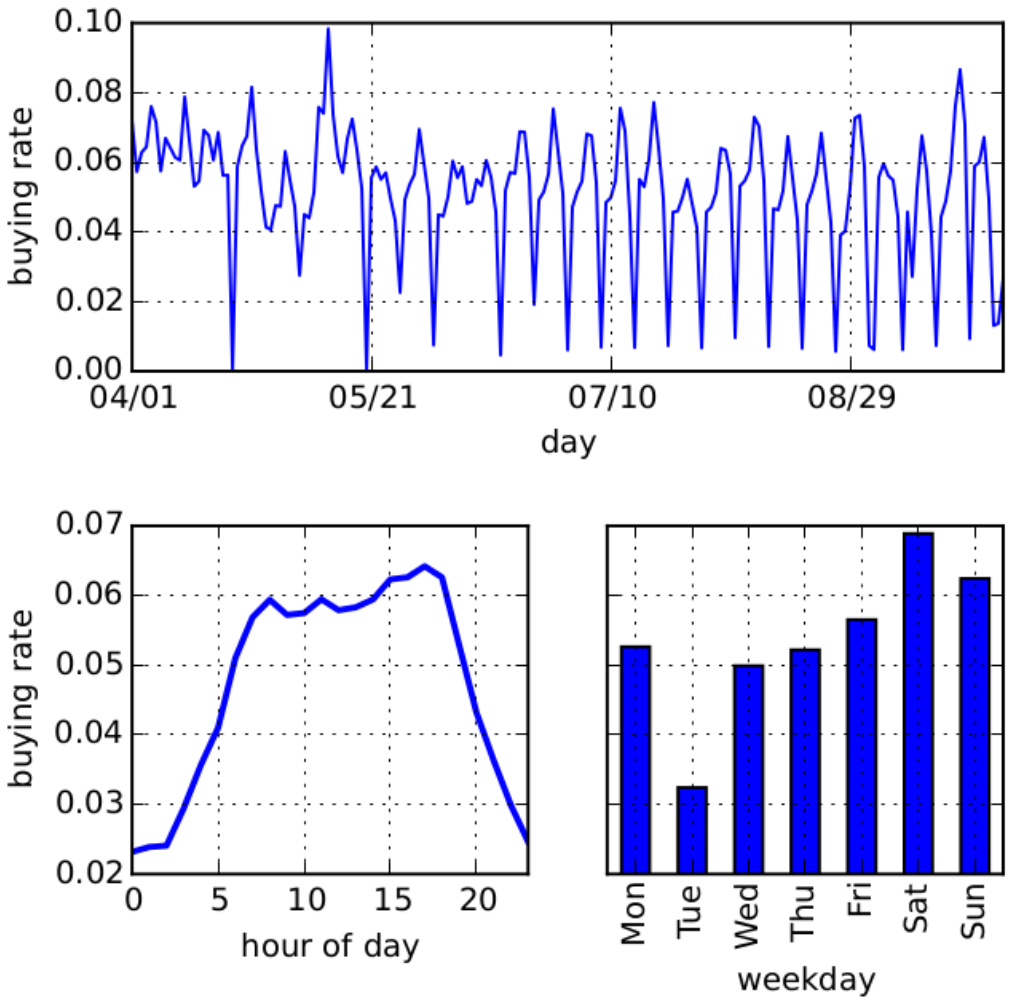
Функционал качества можно переписать следующим образом.
Пусть y(s) – множество предметов, купленных во время сессии s,
h(s) – множество предсказанных предметов, купленных во время сессии

$$Q(h, S_{test}) = \sum_{s \in S_{test}: h(s) > 0} \left(\frac{|S^b_{test}|}{|S_{test}|} (-1)^{isEmpty(y(s))} + J(y(s), h(s)) \right)$$

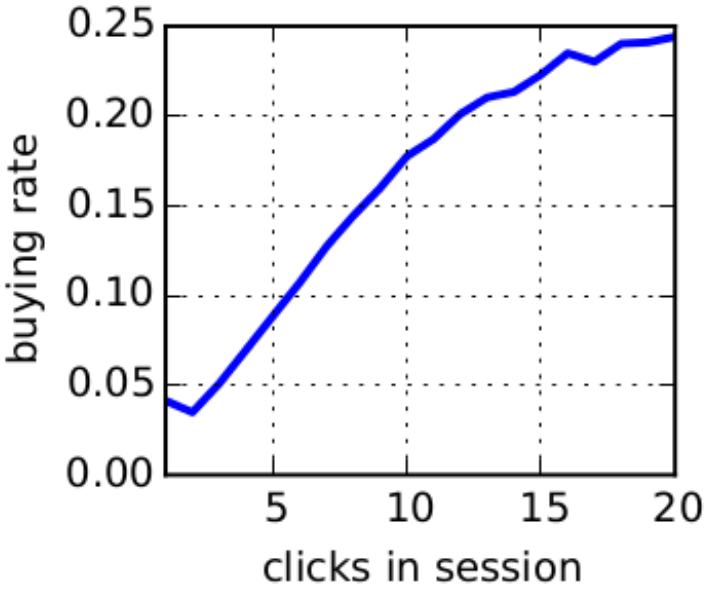
перепишем формулу.

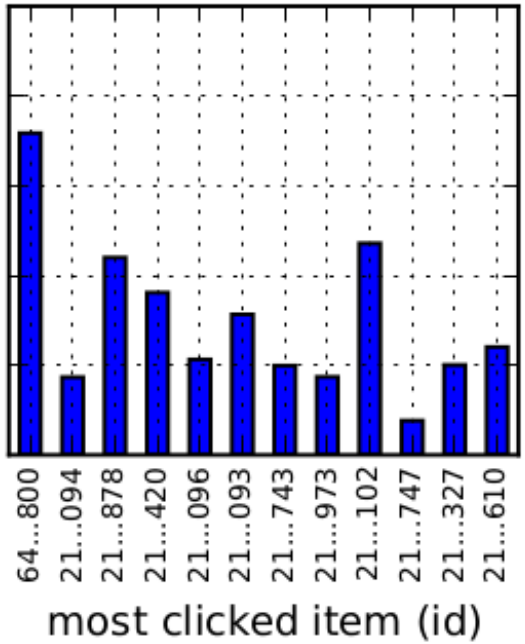
$$Q(h, S_{test}) = \frac{|S^b_{test}|}{|S_{test}|} (TP - FP) + \sum_{s \in S_{test}} J(y(s), h(s))$$

Интересные наблюдения



Интересные наблюдения





Структура итоговой модели

Была использована двух-ступенчатая классификация:

- купил ли пользователь что-то за сессию
- если купил, то какие именно предметы

Признаки сессии

- Начало/конец сессии(месяц, день и т.п)(2×7 Вещест.)
- Начало/конец сессии(месяц, день и т.п) (2×7 Категор.)
- Продолжительность сессии в секундах (1 Вещест.)
- Число кликов, категорий, предметов и категорий предметов (4 Вещест.)
- Топ 10 предметов и топ 5 категорий по числ (15 Категор.)
- ID первого/последнего предмета, кликнутого 1-6 раз (12 Категор.)
- Вектор из 100 числа кликов и общей продолжительности для 100 предметов и 50 категорий, которые были самыми популярными.(150×2 Вещест.)

Признаки пары сессия-предмет

- ID предмета (1 Категор.)
- Общее и относительное число кликов в сессии (2 Вещест.)
- Время первого/последнего клика на предмет(месяц, день и т.п.) (2×7 Вещест.)
- Время первого/последнего клика на предмет(месяц, день и т.п.) (2×7 Категор.)
- Время между первым и последним кликом предмета (15 Категор.)
- Суммарная длительность кликов предмета и категорий предмета в дан. сессию(12 Категор.)
- Число уникальных категорий для предмета в данной дан. сессию)(2 Вещест.)

Итоговая модель

Пусть $h_p(s)$ - детектор покупки, $h_i(s,j)$ - детектор j предмета.
Функция потерь – logloss

$$logloss(h, S_{test}) = \sum_{s \in S_{test}} y(s)log(h(s)) + (1 - y(s))log(1 - h(s))$$

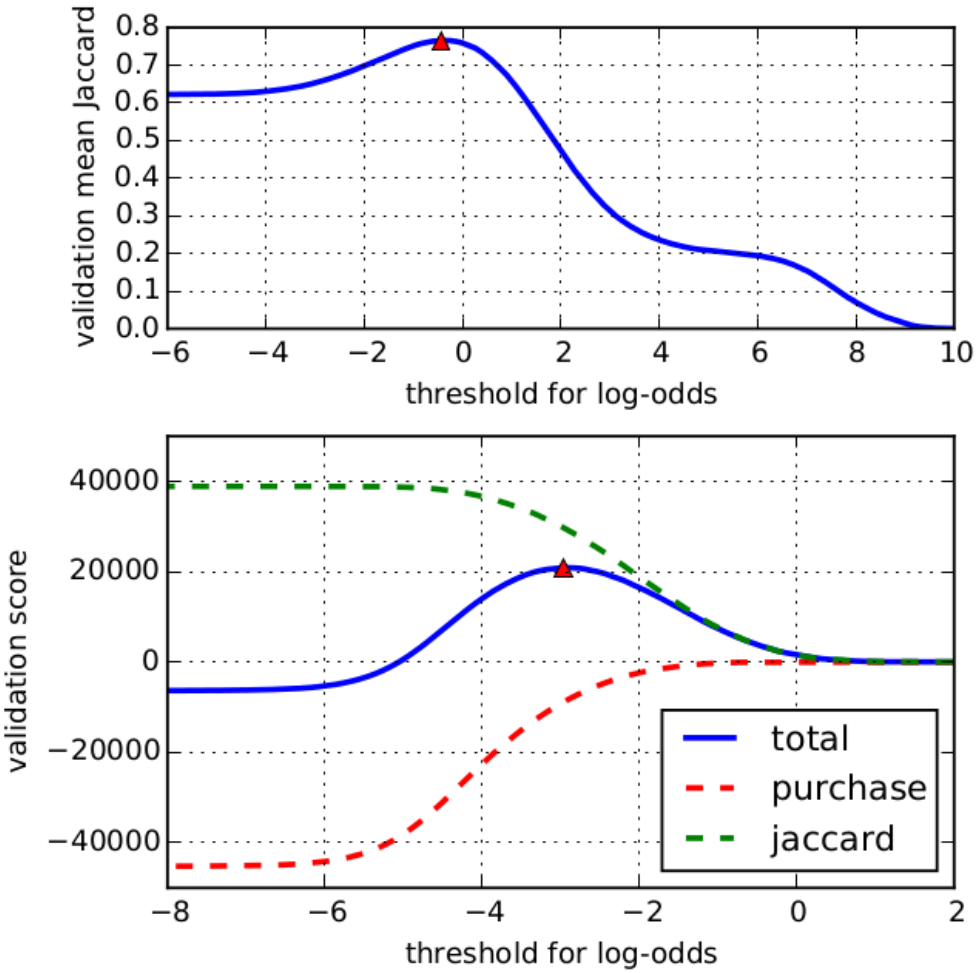
Итоговая модель:

$$h(s) = \begin{cases} \emptyset & \text{если } h_p(s) < \alpha_p \\ \{j | h_i(s,j) > \alpha_j\} & \text{если } h_p(s) > \alpha_p \end{cases}$$

Итоговая модель

Для обучения детектора покупки и предмета был использован MatrixNet.
Обучение – 12 часов на 150 машинах.
Предсказание – 10 минут на 1 машине(около 4000 сессий в секунду).

Оптимизация по порогу



Другие метрики

- AUC детектора покупки – 0.85
- Precision детектора покупки – 16%
- Recall детектора покупки – 77%
- среднее значение меры Жаккара детектора предмета – 0.765