Автоматизация научных исследований в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2018

Постановка задачи прогнозирования дефолта по банковским картам

В крупном банке требуется спрогнозировать остатки на счетах клиентов и вероятность их ухода в дефолт по банковским картам на начало каждого месяца на год вперед. У каждого клиента есть анкета и известны на каждый день остатки на счету, MCC-коды покупок.
Требуется

1) поставить задачу формально,
2) указать технические детали собираемых данных,
3) дополнить и уточнить постановку согласно этим деталям.

Обозначения для формальной постановки задачи

Обозначим индексы через
$k \in \mathcal{K}$ — множество клиентов банка,
$\tau \in \mathcal{T}$ — множество дней мониторинга,
$t \in \mathcal{T}$ — множество дней на начало месяца в запросе на прогноз.
Очевидно поэлементное вложение

$$t \hookrightarrow \tau.$$

Объектом прогнозирования является пара (k, t) — пара клиент (индекс клиента) и момент времени, в который выполняется запрос на прогноз. Обозначим индекс объекта

$$(k, t) \mapsto i \ni \mathcal{I}.$$

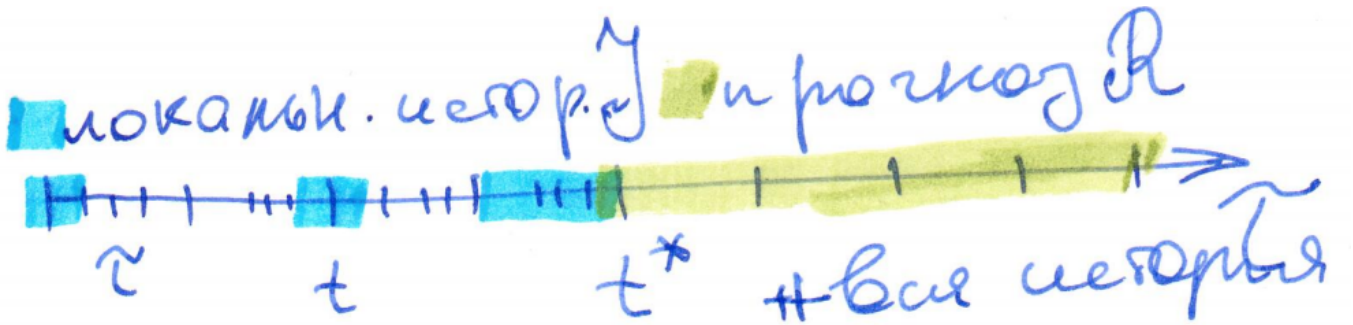
Построение множества индексов описаний

Индексы вектора описания объекта обозначим $\mathcal{J} \ni j$ и поставим им в соответствие локальную предысторию для фиксированного момента времени $t = t^*$ и любого клиента $k \in \mathcal{K}$:

$$h : \mathcal{T}|_{\tau < t^*} \to \mathcal{J}.$$

Индексы вектора ответа обозначим $\mathcal{R}$:

$$r : \mathcal{T}|_{t^* \leqslant t < t^*+12} \xrightarrow{\text{id}} \mathcal{R}.$$



Пример h: для построения локальной предыстории используются дни последнего, относительно момента t^* месяца, нескольких дней последних месяцев из истории в окрестности t^*, дней текущего для t^* месяца последних лет. Назовем такой выбор предыстории априорным (до появления выборки) выбором, или построением, признакового пространства.

Построение выборки по историческим данным

Обозначим временные ряды остатков по дням $s^{\text{res}}_{k\tau}$, типов потребления $s^{\text{MCC}}_{k\tau}$, ухода в дефолт ν^{def}_{kt}, остатков на счетах на начало месяца ν^{res}_{kt}, анкетных данных клиента $\mathbf{u}_k \in \mathbb{R} \times \cdots \times \mathbb{R}$. Построим выборку, в которой объекты с индексами $\mathcal{I} \ni i = i(k, t)$ для запросов на прогнозирование в моменты времени $t = t^*$ заданы как

$$\mathbf{x}_i = [\mathbf{x}_{k,\mathbf{h}(t)}, \mathbf{u}_k]^{\mathsf{T}},$$

и ответы

$$\mathbf{y}_i = \boldsymbol{\nu}_{k,\mathbf{r}(t)},$$

Зададим модель

$$\mathbf{f}(\mathbf{w}) : \mathbf{x}_i \mapsto \mathbf{y}_i$$

и ее функцию ошибки

$$S(\mathbf{w}) = \sum_{i \in \mathcal{I}} \|\mathbf{f}_i - \mathbf{y}_i\|.$$

Вопросы для уточнения постановки задачи

1. Как явно задать функции h, r и обратные им, чтобы было удобно делать анализ выбранных апостериори признаков?
2. Точно ли заданы первые две формулы на слайде 4?
3. Как изменится постановка задачи, в случае, когда клиенты имеют недостаточную длину истории?
4. Как задать отображение вектора $\mathbf{u}$ из произвольных шкал $\mathbb{L} \times \cdots \times \mathbb{L}$ в пространство действительных чисел $\mathbb{R} \times \cdots \times \mathbb{R}$ меньшей размерности?
5. Как породить новые признаки описания клиента с помощью его анкеты в метрическом (евклидовом пространстве) оптимальной размерности?
6. Как привести исходные данные с пропусками и выбросами в тот вид, который требуется для построения адекватной модели?