

# Машинное обучение, ФКН ВШЭ

## Семинар №19

### 1 Факторизационные машины

Пусть обучающая выборка представлена матрицей  $X \in \mathbb{R}^{\ell \times d}$ , где  $\ell$  – кол-во объектов, а  $d$  – кол-во признаков. В случае решения задачи построения рекомендательной системы объектами, как правило, являются пары пользователь-товар, для которых известны оценки  $r_{ui}$  степени заинтересованности пользователя в товаре. При этом признаки могут разбиваться на группы, описывающие различные составляющие информации о паре (см рис. 1).

Feature vector $\mathbf{x}$																			Target $y$				
$\mathbf{x}_1$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...	5	$y_1$	
$\mathbf{x}_2$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...	3	$y_2$	
$\mathbf{x}_3$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...	1	$y_3$	
$\mathbf{x}_4$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...	4	$y_4$	
$\mathbf{x}_5$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...	5	$y_5$	
$\mathbf{x}_6$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...	1	$y_6$	
$\mathbf{x}_7$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...	5	$y_7$	
	A	B	C	...	TI	NH	SW	ST	...	TI	NH	SW	ST	...	Time	TI	NH	SW	ST	...			
	User				Movie					...	Other Movies rated					Time	Last Movie rated					...	

Рис. 1. Представление данных

Факторизационные машины учитывают признаков вплоть до некоторой степени  $r$ , однако чаще всего данные модели используют  $r = 2$ . В этом случае модель будет выглядеть следующим образом:

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j + \frac{1}{2} \sum_{j=1}^d \sum_{k=1, k \neq j}^d \langle v_j, v_k \rangle x_j x_k, \quad (1.1)$$

где  $w_0, w_1, \dots, w_d \in \mathbb{R}$ ,  $v_1, \dots, v_d \in \mathbb{R}^r$  – параметры модели,  $x = (x_1, \dots, x_d)$  – признаковое описание объекта. Первое слагаемое отвечает константному предсказанию, второе слагаемое является линейной моделью, третье слагаемое содержит все попарные взаимодействия  $x_j x_k$ . В отличие от полиномиальной модели, где каждому попарному взаимодействию  $x_j, x_k$  ставится индивидуальный вес  $w_{jk}$ , здесь вес взаимодействия моделируется скалярным произведением двух векторов  $\langle v_j, v_k \rangle$ . Иначе

можно сказать, что матрица всех весов при слагаемых второго порядка  $W$  представляема как  $W = VV^T$ , где  $V = [v_1, \dots, v_d]^T$ .

Из разложения  $W = VV^T$  можно сделать выводы о том, в каких ситуациях применяются факторизационные машины — в частности, в случаях, когда количество признаков  $d$  очень велико. При  $d \gg \ell$  обучение полной матрицы весов попарных взаимодействий быстро приводит к переобучению, поэтому в подобных ситуациях параметр  $r$  обычно выбирают много меньше  $d$ .

## 2 Методы обучения

Пусть имеется обучающая выборка  $X = \{(x_i, y_i)\}_{i=1}^\ell$ . Параметрами модели являются  $\Theta = (w_0, w_1, \dots, w_d, \mathbf{v}_1, \dots, \mathbf{v}_d) = (w_0, w_1, \dots, w_d, v_{11}, \dots, v_{dr})$ . Задача обучения факторизационной машины формулируется стандартным образом:

$$Q(a, X) = \sum_{i=1}^{\ell} L(a(x_i; \Theta), y_i) + \sum_{\theta \in \Theta} \lambda_{\theta} \theta^2 \rightarrow \min_{\Theta}, \quad (2.1)$$

где  $\lambda_{\theta}$  — коэффициент регуляризации параметра  $\theta$ ,  $L(z, y)$  — функция потерь, которая в случае решения задачи регрессии может задаваться как MSE:

$$L(z, y) = (z - y)^2,$$

а в случае бинарной классификации (для  $y \in \{0, 1\}$ ):

$$L(z, y) = -(1 - y) \log(1 - z) - y \log z.$$

Второе слагаемое является регуляризатором и может индивидуально выбираться для каждого  $\theta$ . На практике параметры разбивают на некоторые группы (как на рис. 1), для каждой из которых выбирают индивидуальный параметр  $\lambda_{\theta}$ .

Существует несколько способов обучения факторизационных машин:

- SGD (stochastic gradient descent) — оптимизация параметров происходит методом стохастического градиентного спуска по объектам обучающей выборки.
- ALS (Alternating Least-Squares) — оптимизация параметров происходит попеременно по каждому параметру при фиксированных значениях остальных параметров.
- MCMC (Markov Chain Monte Carlo) — параметры сэмпляются из апостериорного распределения.

### §2.1 SGD

Обучение факторизационных машин при помощи SGD осуществляется стандартным образом — случайным образом выбирается один из объектов обучающей выборки, после чего обновляются все параметры модели. Ниже приведён подробный алгоритм обучения.

**Вход:** Обучающая выборка  $X = \{(x_i, y_i)\}_{i=1}^\ell$ , параметры регуляризации  $\{\lambda_{\theta}\}_{\theta \in \Theta}$ , шаг градиентного спуска  $\eta$ , дисперсия инициализации  $\sigma$ .

**Выход:** Параметры модели  $\Theta^* = (w_0^*, \dots, w_d^*, \mathbf{v}_1^*, \dots, \mathbf{v}_d^*)$ .

$w_j := 0, j = \overline{0, d}, v_{js} \sim \mathcal{N}(0, \sigma), j = \overline{1, d}, s = \overline{1, r};$

**повторять**

**для**  $(x_i, y_i) \in S$

$$w_0 := w_0 - \eta \left( \frac{\partial}{\partial w_0} L(a(x_i; \Theta), y_i) + 2\lambda_0 w_0 \right)$$

**для**  $j \in \{1, \dots, d\} \wedge x_{ij} \neq 0$

$$w_j := w_j - \eta \left( \frac{\partial}{\partial w_j} L(a(x_i; \Theta), y_i) + 2\lambda_{w_j} w_j \right)$$

**для**  $s \in \{1, \dots, r\}$

$$v_{js} := v_{js} - \eta \left( \frac{\partial}{\partial v_{js}} L(a(x_i; \Theta), y_i) + 2\lambda_{v_{js}} v_{js} \right)$$

**пока** веса не стабилизируются

## §2.2 ALS

Alternating Least-Squares или Coordinate Descent (координатный спуск) — метод обучения, при котором каждый параметр настраивается независимо, т.е. все параметры модели настраиваются поочередно при фиксированных значениях остальных параметров. При этом для каждого параметра можно вывести аналитическую формулу для его оптимального значения на каждом шаге данного метода обучения.

**Задача 2.1.** Выразите оптимальное значение некоторого параметра  $\theta \in \Theta$  при фиксированных значениях остальных параметров  $\Theta \setminus \{\theta\}$  при оптимизации функционала  $Q(a, X)$  для задачи регрессии с функционалом  $MSE$ .

**Решение.** Заметим, что для любого параметра  $\theta \in \{w_0, w_1, \dots, w_d, v_{11}, \dots, v_{dr}\}$  модель  $a(x)$  линейна по нему и может быть линейризована следующим образом:

$$\begin{aligned} a(x) &= g_\theta(x) + \theta h_\theta(x), \\ h_{w_0}(x) &= 1, \\ h_{w_j}(x) &= x_j, j = \overline{1, d}, \\ h_{v_{js}}(x) &= x_j \sum_{k \neq j} v_{ks} x_k, j = \overline{1, d}, s = \overline{1, r}, \end{aligned}$$

где функции  $g_\theta(x)$ ,  $h_\theta(x)$  не зависят от значения параметра  $\theta$ .

Запишем оптимальное значение  $\theta^*$  параметра  $\theta$  для решаемой оптимизационной задачи:

$$\begin{aligned} \theta^* &= \arg \min_{\theta} \left( \sum_{i=1}^{\ell} (a(x_i; \theta) - y_i)^2 + \sum_{\alpha \in \Theta} \lambda_{\alpha} \alpha^2 \right) \\ &= \arg \min_{\theta} \left( \sum_{i=1}^{\ell} (g_\theta(x_i) + \theta h_\theta(x_i) - y_i)^2 + \sum_{\alpha \in \Theta} \lambda_{\alpha} \alpha^2 \right) \\ &= \frac{\sum_{i=1}^{\ell} (y_i - g_\theta(x_i)) h_\theta(x_i)}{\sum_{i=1}^{\ell} h_\theta^2(x_i) + \lambda_\theta}. \end{aligned}$$

■

Как правило,  $g_\theta(x)$  не выражают в явном виде, а используют представление  $g_\theta(x) = a(x; \Theta) - \theta h_\theta(x)$ , из которого можно вывести формулу обновления параметра  $\theta$  при использовании метода ALS:

$$\begin{aligned}\theta^* &= \frac{\sum_{i=1}^{\ell} (y_i - g_\theta(x_i)) h_\theta(x_i)}{\sum_{i=1}^{\ell} h_\theta^2(x_i) + \lambda_\theta} = \frac{\sum_{i=1}^{\ell} (y_i - a(x_i; \Theta) + a(x_i; \Theta) - g_\theta(x_i)) h_\theta(x_i)}{\sum_{i=1}^{\ell} h_\theta^2(x_i) + \lambda_\theta} = \\ &= \frac{\theta \sum_{i=1}^{\ell} h_\theta^2(x_i) + \sum_{i=1}^{\ell} h_\theta(x_i) (y_i - a(x_i; \Theta))}{\sum_{i=1}^{\ell} h_\theta^2(x_i) + \lambda_\theta}.\end{aligned}$$

Это выражение использует текущие значения параметров модели и её прогнозы на объектах обучающей выборки и используется для итеративного обновления каждого параметра  $\theta$  при фиксированных значениях остальных параметров при использовании метода ALS для обучения модели.

## 3 Обобщение факторизационных моделей

### §3.1 Matrix Factorization

Пусть решается задача построения рекомендательной системы, имеются множество пользователей  $U$  и множество товаров  $I$ . Воспользуемся факторизационными машинами для решения этой задачи. Естественный путь описания пары  $(u, i) \in U \times I$  — это бинарный вектор, состоящий из нулей и содержащий ровно две единицы:

$$(u, i) = (\underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|U|}, \underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|I|}),$$

где первая единица соответствует индексу пользователя и стоит на  $u$ -ом месте, а вторая единица соответствует индексу товара и стоит на  $(|U| + i)$ -ом месте. Тогда факторизационная машина, обученная на данных, представленных таким образом, будет выглядеть следующим образом:

$$a(x) = w_0 + w_u + w_{|U|+i} + \langle v_u, v_{|U|+i} \rangle. \quad (3.1)$$

Как видим, данная модель совпадает с моделью в подходе, использующем матричное разложение (со сдвигами  $w_u, w_i$ ). Вектора  $v_u, v_{|U|+i} \in \mathbb{R}^r$  — латентные вектора пользователей и товаров соответственно.

### §3.2 Pairwise Interaction Tensor Factorization

После того, как мы выпустили первую версию рекомендательной системы, разработчики добавили в интерфейс системы проставление тегов. Теперь, помимо множества пользователей  $U$  и множества товаров  $I$  у нас есть множество тегов  $T$ . Естественный путь описания тройки  $(u, i, t) \in U \times I \times T$  — бинарный вектор, состоящий из нулей и содержащий ровно три единицы:

$$(u, i, t) = (\underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|U|}, \underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|I|}, \underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|T|}),$$

где первая единица соответствует индексу пользователя и стоит на  $u$ -ом месте, вторая единица соответствует индексу товара и стоит на  $(|U| + i)$ -ом месте, а третья единица соответствует индексу тега и стоит на  $(|U| + |I| + t)$ -ом месте. Тогда факторизационная машина, обученная на данных, представленных таким образом, будет выглядеть следующим образом:

$$a(x) = w_0 + w_u + w_{|U|+i} + w_{|U|+|I|+t} + \langle v_u, v_{|U|+i} \rangle + \langle v_u, v_{|U|+|I|+t} \rangle + \langle v_{|U|+i}, v_{|U|+|I|+t} \rangle. \quad (3.2)$$

Как видим, данная модель совпадает с моделью Pairwise Interaction Tensor Factorization. Разумеется, пример с тегом является очень частным, и вместо тега мы можем использовать любой категориальный признак.

### §3.3 SVD++

Продолжаем развитие нашей рекомендательной системы. Пусть на этот раз мы столкнулись с категориальным признаком, который принимает ненулевые значения на некотором **подмножестве** категорий. Таким образом, объект обучающей выборки представлен двумя категориальными признаками: пользователь  $u$ , товар  $i$ , и множественно-категориальным признаком  $l$ , который может принимать  $m$  ненулевых значений. На Рис. 1 первые три группы признаков соответствуют такому представлению. Как и раньше, опишем такое представление числовым вектором:

$$(u, i, l = \{l_1, \dots, l_m\}) = (\underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|U|}, \underbrace{0, \dots, 0, 1, 0, \dots, 0}_{|I|}, \underbrace{0, \dots, 1/m, \dots, 1/m, \dots, 0}_{|L|}),$$

Как и в предыдущих пунктах, ненулевые значения полученного вектора являются индикаторами категорий признаков (подмножества категорий в случае множественно-категориального признака). Модель для факторизационных машин выглядит следующим образом:

$$\begin{aligned} a(x) = & w_0 + w_u + w_i + \langle v_u, v_i \rangle + \frac{1}{m} \sum_{f=1}^k \langle v_i, v_{l_j} \rangle \\ & + \frac{1}{m} \sum_{j=1}^m w_{l_j} + \frac{1}{m} \sum_{j=1}^m \langle v_u, v_{l_j} \rangle + \frac{1}{m^2} \sum_{j=1}^m \sum_{j'=1}^m \langle v_{l_j}, v_{l_{j'}} \rangle \end{aligned}$$

Первая часть (первая строка уравнения) соответствует модели под названием SVD++. Вторая строчка модели содержит дополнительные взаимодействия между признаками.

### §3.4 Заключение

Мы рассмотрели три модели, и на этих примерах выяснили, что факторизационные машины являются некоторым обобщением факторизационных моделей. Фак-

торизационные машины предоставляют удобный способ учёта взаимодействий высоких порядков, в случае, когда пространство признаков велико по сравнению с размером обучающей выборки.