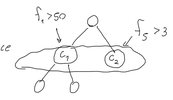
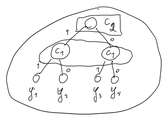
LSML #8

Random Forest → каждое дерево ①
→ построение сплита в дереве

Catboost

1) CPU/GPU training & inference
2) Мало гиперпараметров
3) Новые loss функции
4) Oblivious trees





$\boxed{11}$ $\boxed{10}$ = 2

	$c_1 = 1$	$c_1 = 0$
$c_2 = 1$	$\boxed{y_3}$	$\boxed{y_2}$
$c_2 = 0$	y_1	y_0

· можем за 1 пар быстро найти нужный y_i

$[y_0 ; y_1 ; y_2 ; y_3]$

5) Catboost + Spark

RecSys

User \ Items	1	2	3	4
1		4	5	2
2	5	①		①
3		5		2
4	1		5	4
5		①		②
6	①	5	0	1

✓User-based CF
✓Item-based CF

$$corr(u,v) = \frac{\sqrt{\sum_{i \in I}} \left[(r_{ui} - r_u)(r_{vi} - r_v) \right]}{\sqrt{\left(\sum_{i \in I} (r_{ui} - r_u)^2\right) \cdot \left(\sum_{i \in I} (r_{vi} - r_v)^2\right)}}$$

$$\hat{r}_{ui} = r_u + \frac{\sum_{v \in K} corr(u,v) \cdot (r_{vi} - r_v)}{\sum_{v \in K} |corr(u,v)|}$$

n - users
m - items

$n \cdot m \cdot n$ $\boxed{n^2 \cdot m}$

· Обычно, матрица рейтингов очень спарсовая (~99.99% нули)

ratings [u_id, i_id, rating]
↓
u_id : key
(i_id, rating) : value
↓
groupbykey()
↓
flatMap() [(i_id$_1$, r_1) . — (i_id$_k$, r_k)]

∀ i_1 r_1 in values:
  ∀ i_2 r_2 in values[i_1:]:
    yield (i_1, i_2), $(r_1 - m_1) \cdot (r_2 - m_2)$
    key, value

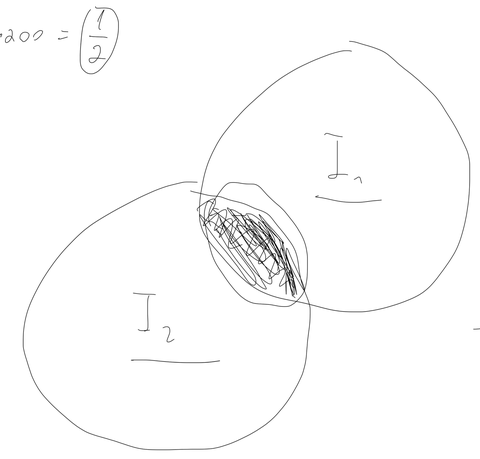
groupbykey()
↓
sum()

✓20k - взаим.
2k - users
200 - item
взаимодейств разномерны

1 user → 10 1 item → 100

$\frac{10}{200} = \left(\frac{1}{20}\right)^2$ $\left(\frac{1}{400}\right) \cdot 200 = \frac{1}{2}$

$\frac{100}{2000} = \left(\frac{1}{20}\right)^2$ $\frac{1}{4\cancel{00}} \cdot 20\cancel{00} = \boxed{5}$



check
1f - i0f
item

$$\frac{|I_1 \& I_2|}{|I_1 | I|}$$

Netflix Prize

A

$A^T A$

Funk SVD

$X \in \mathbb{R}^{\boxed{n} \times k}$

$Y \in \mathbb{R}^{\boxed{m} \times k}$

$$\min_{X,Y} \sum_{(u,i) \in R} (r_{ui} - x_u^T y_i)^2 + \lambda \cdot (\|x_u\|_2^2 + \|y_i\|_2^2)$$

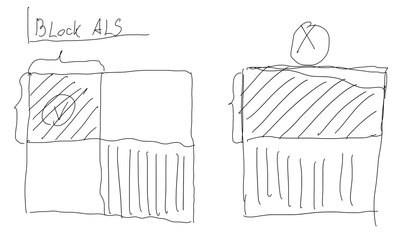
ALS

$$\min_{X,Y} \sum_{(u,i) \in R} (r_{ui} - x_u^T y_i)^2$$

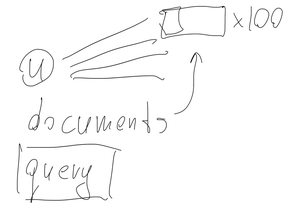
$\boxed{Y}$ fixed

$r_{ui} = x_u^T y_i$
$r_{vi} = x_v^T y_i$

$Y = X^{(u)}$



$m = \boxed{10^6} \to 100 \to \boxed{10}$

DSSM

HNSW