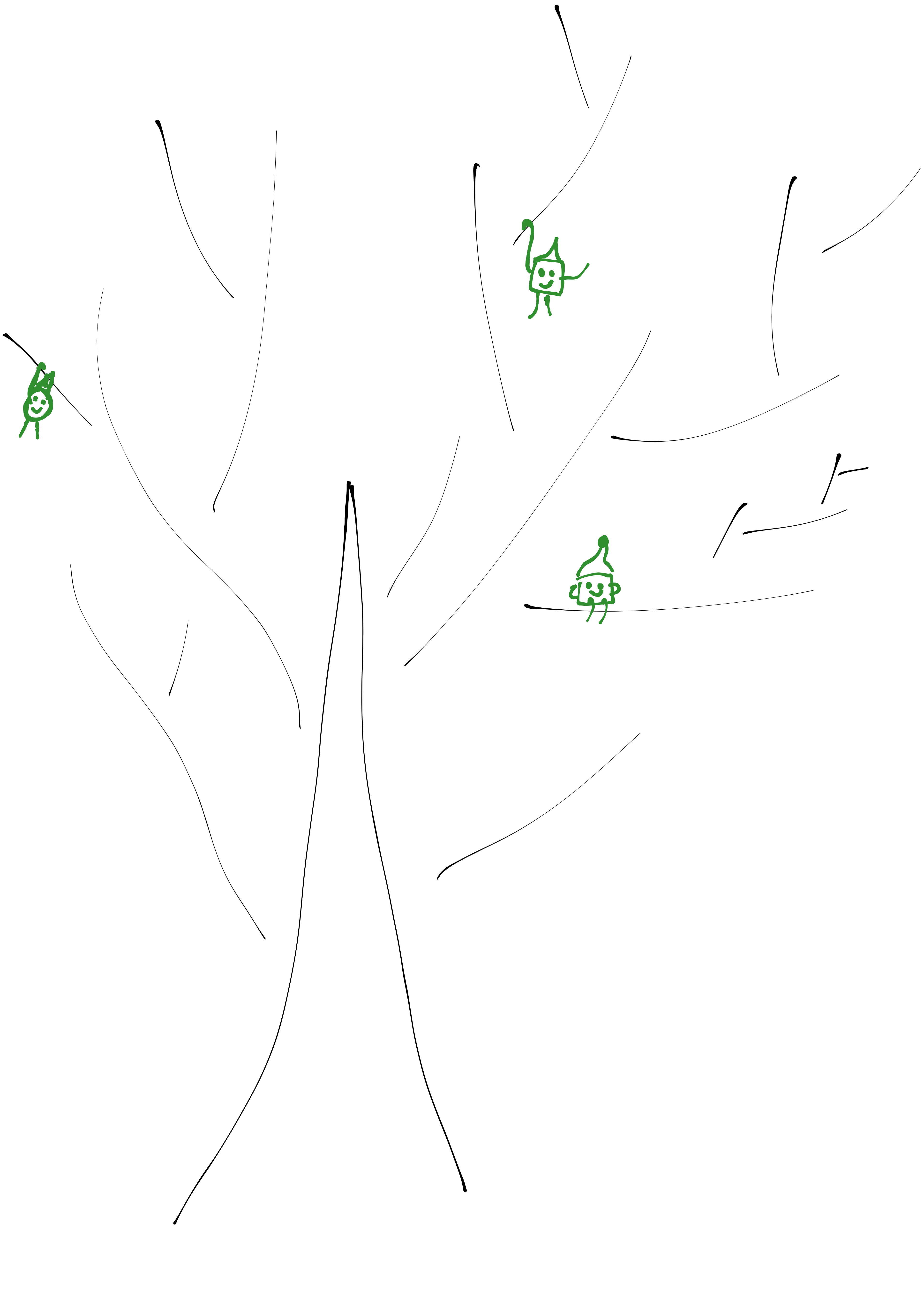
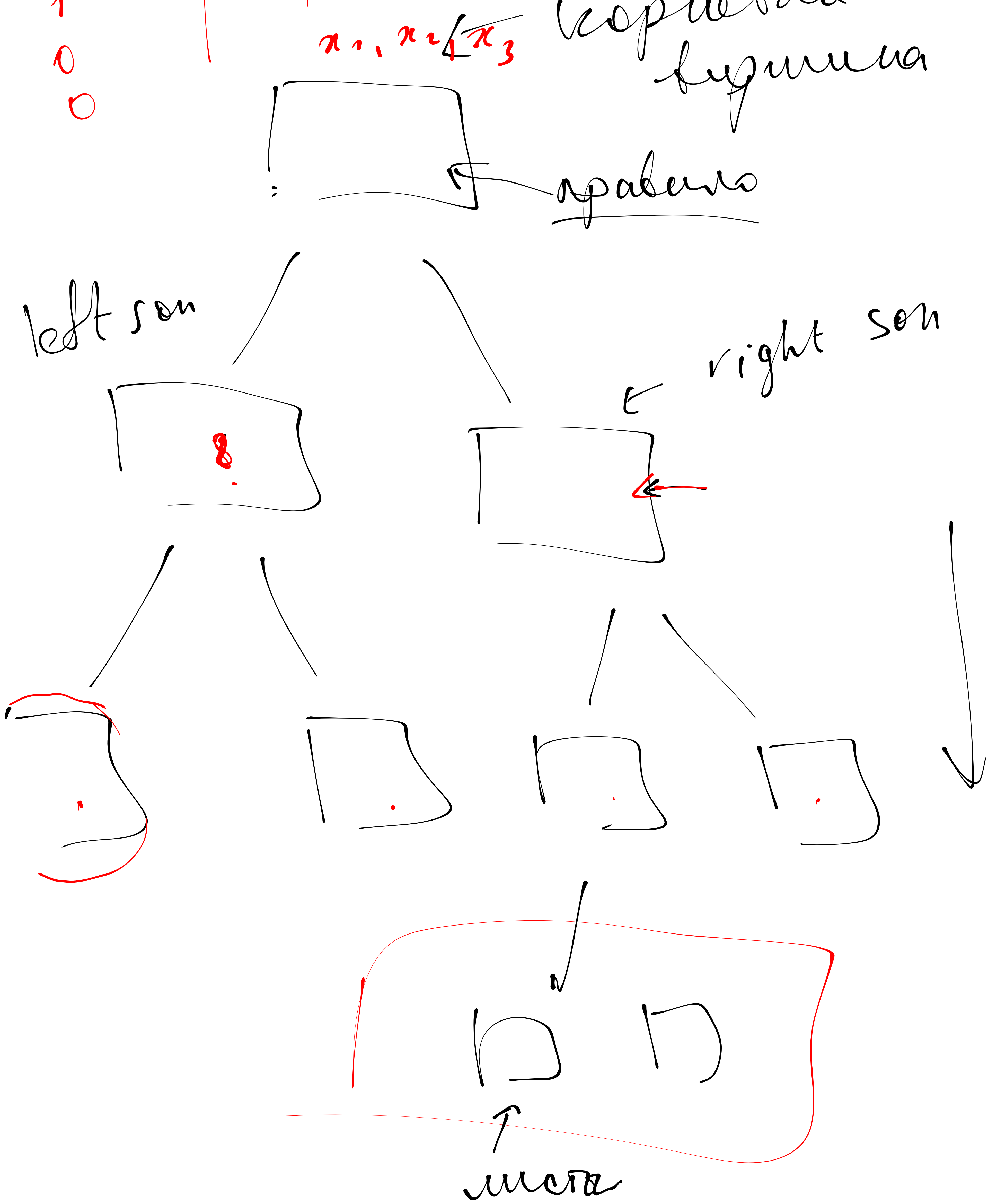


y	x_1	x_2	x_3
0			
1			
0			
0			

Дерево



Гиперпараметры:

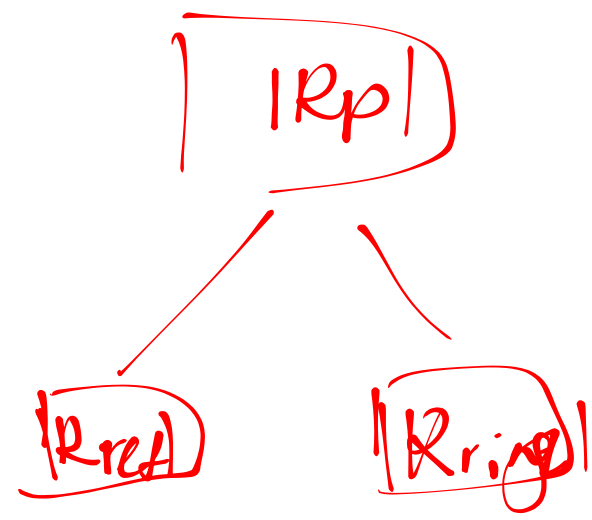
max_depth - макс. глубина

выбор функц.

min_samples_leaf

max_features -

min_samples_split



information gain

$$IG = I(R_p) - \frac{|R_{left}|}{|R_p|} I(R_{left}) - \frac{|R_{right}|}{|R_p|} I(R_{right}) \to \max$$

① Ошибка классиф.

$$\boxed{I = 1 - \max_k p_k}$$

② $$I = -\sum_{i=1}^{c} p_i \log p_i$$

$\hat{p}_1 \ldots \hat{p}_k$

вершина

$$L = \frac{1}{|R_m|} \prod_{j=1}^{K} c_j^{\sum_{i=1}^{|R_m|}[y_i = k]} \to \max_{c}$$

R_m

Пример.

$$L = p_1^{100} \cdot p_2^{50} \cdot \ldots p_k^{500}$$

$$\sum_{j=1}^{K} c_j = 1$$

потери

① $$\ell = -\frac{1}{|R_m|} \sum_{i=1}^{|R_m|} \sum_{j=1}^{K} [y_i = j] \overset{\log}{c_j} \to \min$$

$$\sum_{j=1}^{k} c_j = 1$$

$$\mathcal{L} = -\frac{1}{|R_m|} \sum \sum [y_i = j] \overset{\log}{c_j} + \lambda\left(\sum c_j - 1\right) \to \min_{c_1, \ldots, c_k}$$

$$L'_{c_k} = \underbrace{-\frac{1}{|R_m|} \sum_{i=1}^{|R_m|} [y_i = k]}_{p_k} \frac{}{c_k} + \lambda = 0$$

$$-\frac{p_k}{c_k} + \lambda = 0 \qquad c_k = \frac{p_k}{\lambda}$$

$$\sum_{i=1}^{k} c_i = 1$$

$$\sum_{i=1}^{k} p_i = 1.$$

$$\sum c_i = \frac{\overbrace{\sum p_i}^{1}}{\hat{\lambda}} = 1 \rightarrow \hat{\lambda} = 1$$

$$\Longrightarrow p^* = (p_1 \ \ldots \ p_k)$$

$$p_1 = \frac{\sum_{i=1}^{|R_m|} [y_i = 1]}{|R_m|}$$

$$\boxed{H(R_m) = -\sum_{i=1}^{K} p_i \log p_i}$$

x	0	1	5
	0.2	0.3	0.5

$$H(R_m) = -(0.2 \log 0.2 + 0.3 \log 0.3 + 0.5 \log 0.5)$$

для какого-то k

$p_k = 1$. $\Rightarrow$

$p_1 = 0 \ldots \quad p_{k-1} = 0$

$$H = -(1 \log_2 1 + 0 \log_2 0 \ldots) =$$

$= 0$

Покажем, что энтропия ограничена сверху

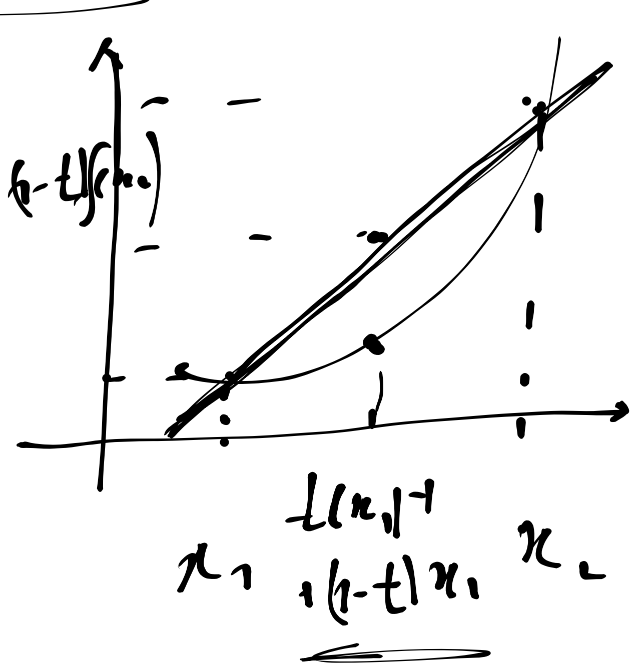
$$\boxed{\sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i} = -\sum p_i \log_2 p_i}$$

$\boxed{\text{Нер-во Йенсена}}$

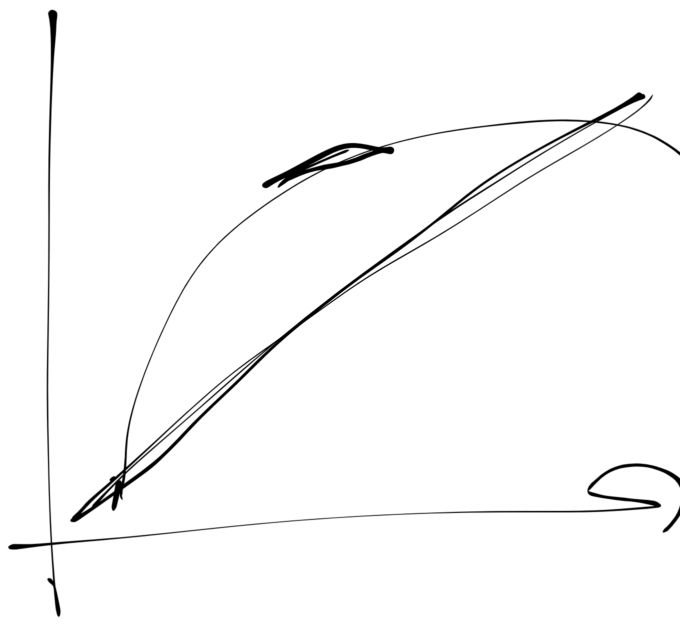
f — выпуклая

$$f\left(\sum a_i x_i\right) \le \sum a_i f(x_i)$$

$\sum a_i = 1$



f – вогнутая = выпуклая вверх

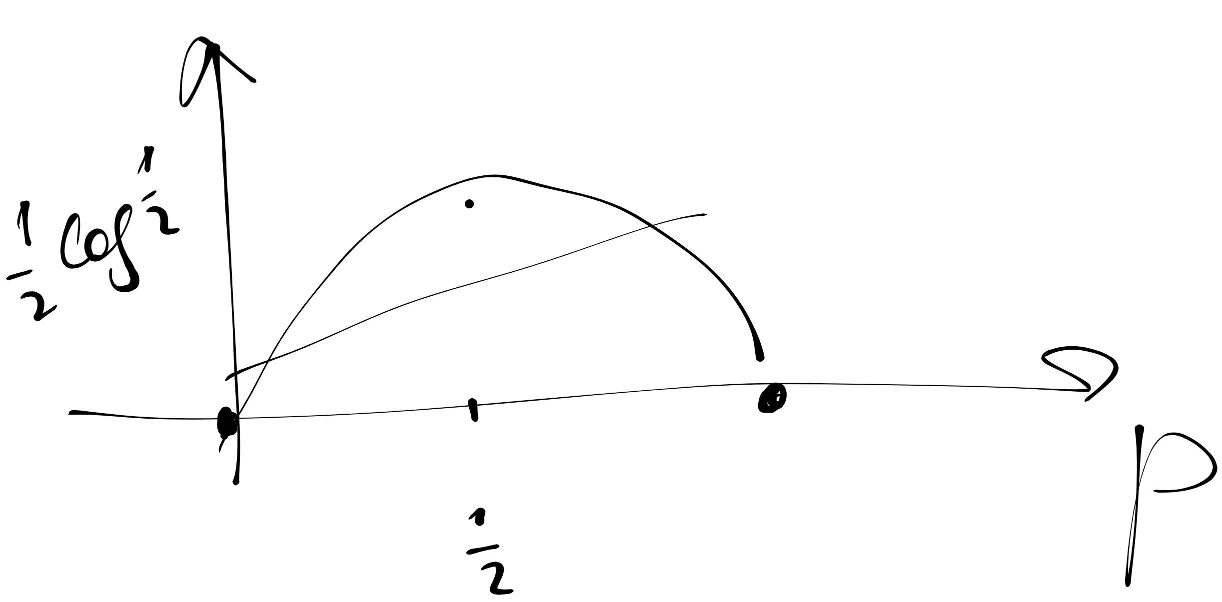


$$f\left(\sum a_i x_i\right) \ge \sum a_i f(x_i)$$

$\boxed{0 \log_2 0 = 0}$

$(p \log p)$ $0 \le p \le 1$

мур?



$p \log p \dots$

$$\log p + \frac{p}{p} = 0$$

$$\log p = -1.$$

$$\log p_{p=\frac{1}{2}}$$

$$(p \log p)'_p = 0$$

$$\hat{p} = \frac{1}{2}$$

$$\sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i} \le \log_2 \left(\sum_{i=1}^{K} p_i \cdot \frac{1}{p_i}\right) = \boxed{\log_2 K}$$

1	2	...	K
$\frac{1}{K}$	$\frac{1}{K}$		$\frac{1}{K}$

$$\sum_{i=1}^{K} \underbrace{\frac{1}{K}}_{=p_i} \log_2 \underbrace{K}_{=\frac{1}{p_i}} = \boxed{\log_2 K.}$$

№1. x_1

Марки.

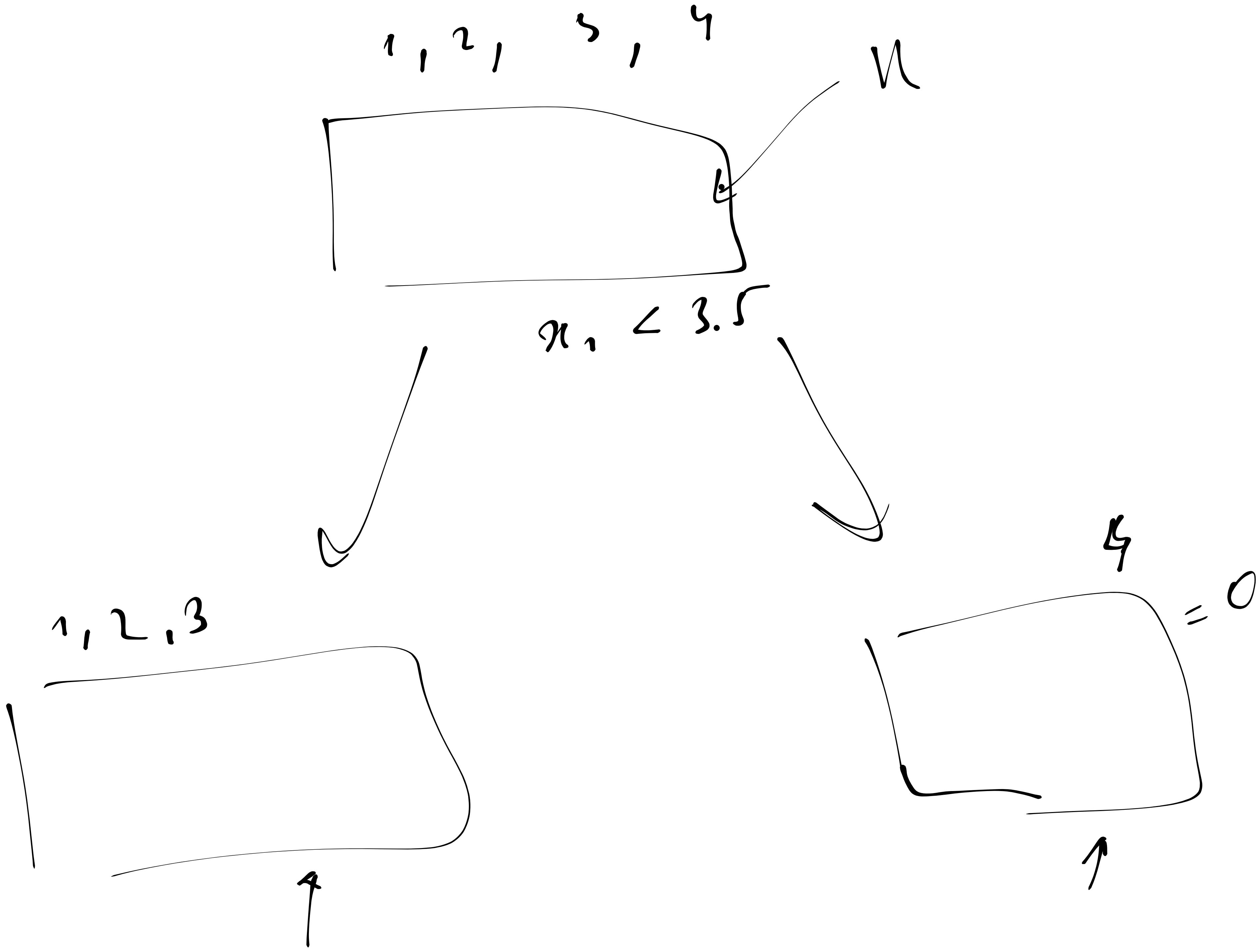
30 – про природу 1

20 – про погоду 2

3 – про ж. 3

4 – про м. 4

$$H(R_m) = \left(\frac{30}{57} \log \frac{57}{30} + \frac{20}{57} \log \frac{57}{20} \ldots \right)$$



$$IG = H(R_p) - (\_)\, H(R_l) - (\_)\, H(R_r)$$