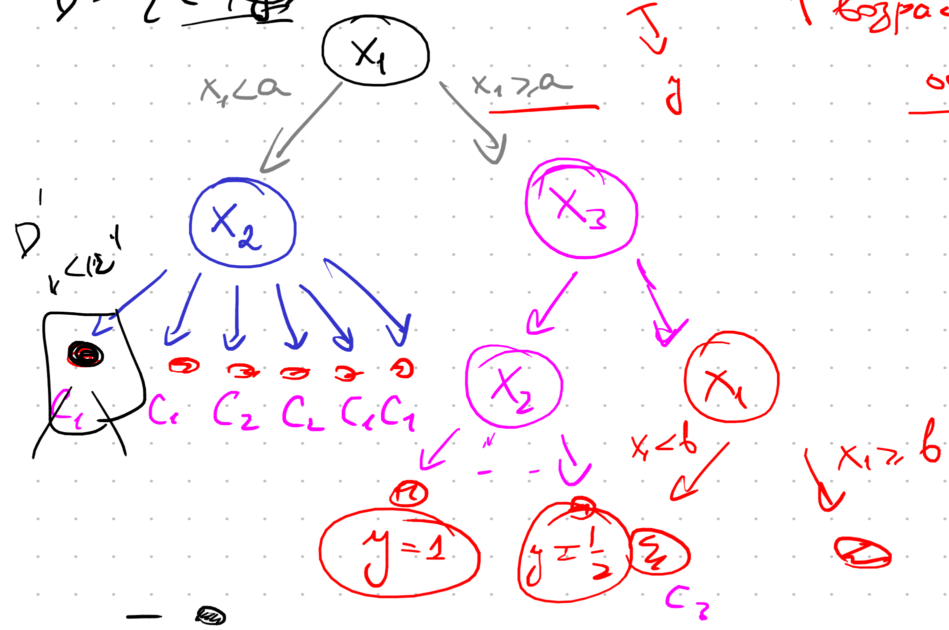
① Decision trees

$D = \{(\bar{x}, y)\}$

$\bar{x} = (x_1, x_2, \dots, x_d)$, $x_1 \in \mathbb{R}$

возраст $\in$ { "<12", "12-16", "17-20", "21-30", "31-45", ">45" }

ordinal



— рекурсивно:

— хотим ли мы расщепляться? если нет, stop

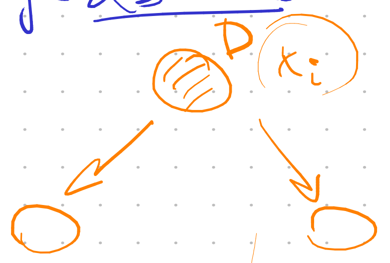
— выбрать признак i x_i

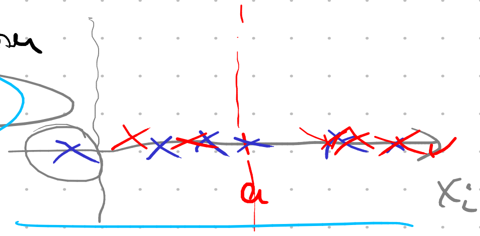
— выбрать, как:

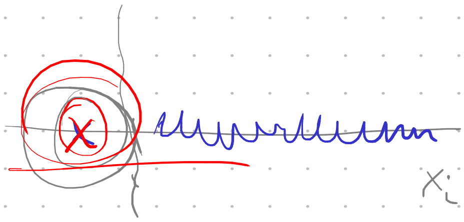
— если x_i — катег. то по значениям

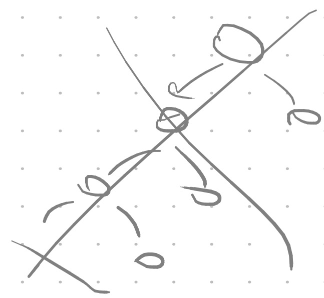
— если $x_i \in \mathbb{R}$, то выбрать a

$(\bar{x}, y \in \{C_1, \dots, C_K\})$







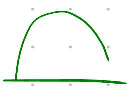


$D_1 = D|_{x_i < a}$ $D_2 = D|_{x_i \geq a}$

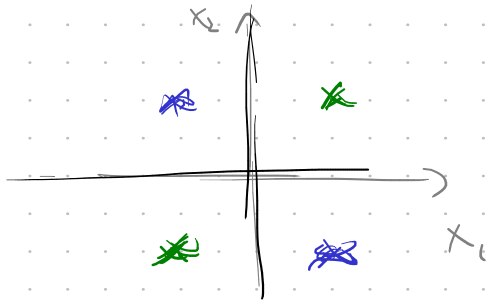
$$\boxed{H(D_1) = -\sum_k p_k \log p_k}, \text{ где } p_k = \frac{\#\{n \mid y_n \in C_k\}}{|D_1|}$$

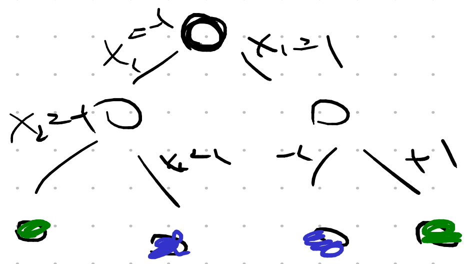
$H(D_2)$

$$Gain(D, x_i) = -\sum_{s=1}^{S} \frac{|D_s|}{|D|} H(D_s) + \cancel{H(D)}$$



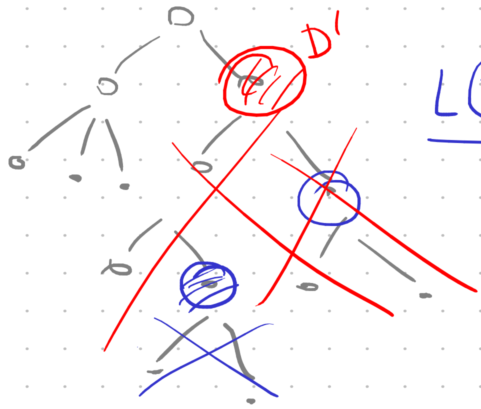
$$Gini(D) = \sum_k p_k (1 - p_k)$$





pruning

Random forest



$$\underline{L(T)} + \underline{\lambda \cdot |T|}$$

② Model combination

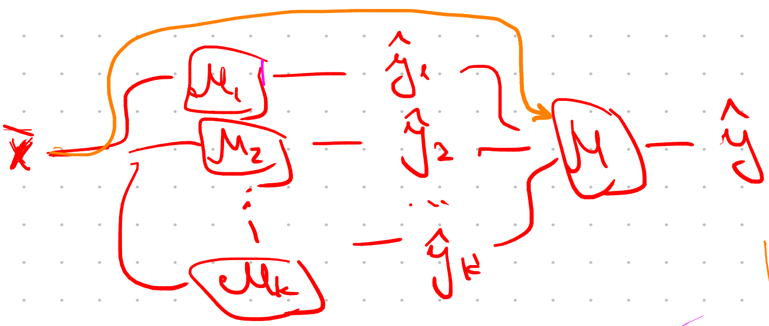
- model selection — BIC, TIC, AIC

$$M_1, M_2, \dots, M_K \longrightarrow M^* = M_{k^*}$$

$$p(\bar{\theta} \mid D, M_k) = \frac{\text{---}''\text{---}}{p(D \mid M_k)}$$

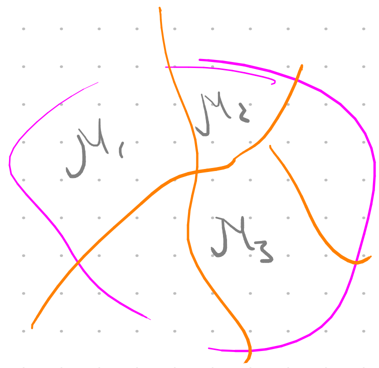
$$k^* = \arg\max\, p(M_k \mid D) = \arg\max\, \overline{p(D \mid M_k)}\, p(\cancel{M_k})$$

- blending



- mixture of experts (MoE)

$$\bar{x} \;—\; \underset{\text{router}}{\boxed{M}} \;—\; k \;—\; \boxed{M_k} \;—\; \hat{y}$$



- committee

$$M(\bar{x}) = \underline{\frac{1}{K} \sum_{k=1}^{K} M_k(\bar{x})}$$

$$y_k(\bar{x}) = \underline{h(\bar{x}) + \varepsilon_k(\bar{x})}$$

$$\text{Error} = E_{\bar{x}}\left[(y_k - h)^2\right] = E_{\bar{x}}\left[\varepsilon_k^2(\bar{x})\right]$$

$$\text{Error}_{avg} = \frac{1}{K} \sum_k E_{\bar{x}}\left[\varepsilon_k^2\right]$$

$$\text{Error}_{comm} = E_{\bar{x}}\left[\left(\frac{1}{K}\sum y_k - h\right)^2\right] = E_{x}\left[\left(\left(\frac{1}{K}\right)\sum \varepsilon_k(\bar{x})\right)^2\right] =$$

$$= \frac{1}{k^2}\left(\sum_k \mathbb{E}_x\left[\varepsilon_t^2(\bar x)\right] + \sum_{k\neq l} \cancel{\mathbb{E}_x\left[\varepsilon_t(\bar x)\varepsilon_l(\bar x)\right]}_{\to 0}\right) = \frac{1}{k^2}\sum_t \mathbb{E}_x\left[\varepsilon_t^2\right]$$

$$\frac{E_{avg}}{k} \le Err_{comm} \le E_{avg}$$

- bagging — bootstrap aggregation

$$X_1, X_2, \dots, X_N \sim X$$

$X_1 - X_n \sim \dots$

$X_1 - X_n \sim \dots$

$X_1 \dots X_N \sim \dots$

$$\frac{\text{Sample}}{\text{Population}} = \frac{\text{Subsample}}{\text{sample}}$$

$$D \to \begin{cases} D_1 \sim D \to M_1 \\ D_2 \sim D \to M_2 \\ \vdots \\ D_m \sim D \to M_m \end{cases} \qquad y = \frac{1}{m}\sum_{i=1}^{m} M_i(\bar x)$$

3) **Boosting**

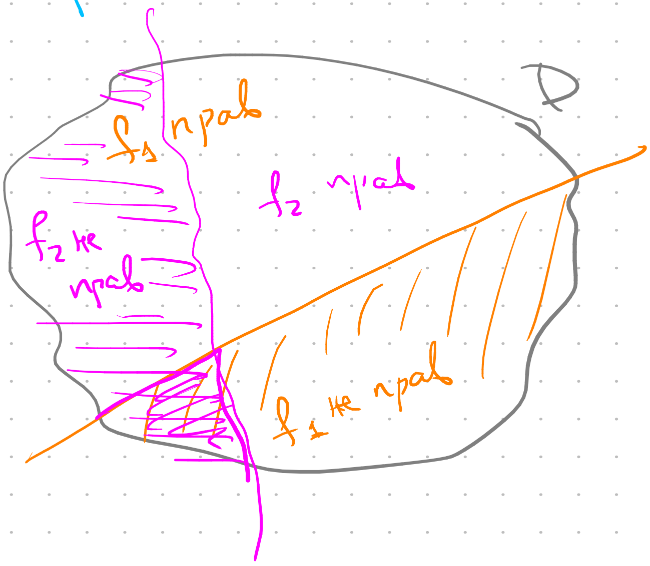
$$F(\bar x) = \alpha_1 f_1(\bar x) + \alpha_2 f_2(\bar x) + \dots + \alpha_T f_T(\bar x)$$

$$F_k(\bar x) = \underbrace{F_{k-1}(\bar x)}_{\text{фиксир.}} + \alpha_k f_k(\bar x)$$

1997
Freund, Schapire
AdaBoost

— мы учим обуч. слабые классиф. f_t на D, минимизируя

$$\varepsilon_t = \sum_{i=1}^{N} w_i^{(t)}\left[f_t(\bar x_i) \neq y_i\right]$$



AdaBoost:
- init $w_i^{(1)} = 1/N$
- for $t = 1 \dots T$:
  - обучаем f_t (с весами $w_i^{(t)}$), получаем $\varepsilon_t < \frac{1}{2}$
  - $\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$
  - $w_i^{(t+1)} := \frac{1}{Z_t} w_i^{(t)} \cdot e^{-\alpha_t \cdot y_i f_t(\bar x_i)}$, где $Z_t = \sum_i w_i^{(t)} e^{-\alpha_t y_i f_t(\bar x_i)}$

$y_i \in \{\pm 1\}$

$f_t(\bar x_i) \in \{\pm 1\}$

$$-\quad \hat{y} = \text{sign}\Big(\underbrace{\sum_t \alpha_t f_t(\bar{x})}_{= F(\bar{x})}\Big)$$

$$\text{error} \quad \text{sign}(F(\bar{x}_i)) \neq y_i \qquad e^{-y_i F(\bar{x}_i)} \geq 1$$

$$\underline{\varepsilon_t = \frac{1}{2} - \gamma_t, \quad \gamma_t > 0}$$

$$w_i^{(t+1)} = \frac{1}{N}\,\frac{e^{-y_i f_{t+1}(\bar{x}_i)}}{Z_1 \ldots Z_t}$$

$$w_i^{(T+1)} = w_i^{(1)} \cdot \frac{1}{Z_1} e^{-\alpha_1 y_i f_1(\bar{x}_i)} \cdot \frac{1}{Z_2} e^{-\alpha_2 y_i f_2(\bar{x}_i)} \cdot \ldots \cdot \frac{1}{Z_T} e^{-\ldots}$$

$$= \frac{1}{N} \cdot \frac{e^{-y_i(\alpha_1 f_1(\bar{x}_i) + \ldots + \alpha_T f_T(\bar{x}_i))}}{Z_1 Z_2 \ldots Z_T} = \frac{1}{N}\,\frac{e^{-y_i F(\bar{x}_i)}}{Z_1 \ldots Z_T}$$

$$\underline{\text{Error}} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{[\text{sign}(F(\bar{x}_i)) \neq y_i]}_{\leq e^{-y_i F(\bar{x}_i)}} \leq \frac{1}{N}\sum_{i=1}^{N} e^{-y_i F(\bar{x}_i)} =$$

$$= \sum_{i=1}^{N} w_i^{(T+1)} Z_1 Z_2 \ldots Z_T = \underline{Z_1 Z_2 \ldots Z_T}$$

$$Z_t = \sum_i w_i^{(t)} e^{-\alpha_t y_i f_t(\bar{x}_i)} = \sum_{\underline{i: y_i = f_t(\bar{x}_i)}} w_i^{(t)} e^{-\alpha_t} + \sum_{i: y_i \neq f_t(\bar{x}_i)} w_i^{(t)} e^{+\alpha_t}$$

$$= e^{-\alpha_t} \cdot (1 - \varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t \neq e^{-\alpha_t}\left(\frac{1}{2} + \gamma_t\right) + e^{\alpha_t}\left(\frac{1}{2} - \gamma_t\right)$$

$$= \left[\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}\right] = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \cdot (1 - \varepsilon_t) + \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \cdot \varepsilon_t =$$

$$= 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = 2\sqrt{\frac{1}{4} - \gamma_t^2} = \sqrt{1 - 4\gamma_t^2}$$

$$\text{Error} \leq \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2} \leq e^{-2\sum_{t=1}^{T}\gamma_t^2} \qquad 1 + x \leq e^x$$

(Friedman, 2000) :

$$L = \sum_{i=1}^{N} e^{-y_i F(\bar{x}_i)} \qquad F(\bar{x}_i) = \frac{1}{2}(\alpha_1 f_1 + \ldots + \alpha_T f_T)$$

F_{t-1} = фикс. $F_t(\bar{x}) = F_{t-1}(\bar{x}) + \frac{1}{2}\alpha_t f_t(\bar{x})$

$$L = \sum_{i=1}^{N} e^{-y_i (F_{t-1}(\bar{x}_i) + \frac{1}{2}\alpha_t f_t(\bar{x}_i))} =$$

$$= \sum_i e^{-y_i F_{t-1}(\bar{x}_i)} \cdot e^{-\frac{1}{2} y_i \alpha_t f_t(\bar{x}_i)} = \sum_{i=1}^{N} N \cdot w_i^{(t)} \cdot \underbrace{\prod z_s}_{\text{Const}} \cdot e^{-\frac{1}{2} y_i \alpha_t f_t(\bar{x}_i)}$$

$$e^{-y_i F_{t-1}(\bar{x}_i)} = N w_i^{(t)} \cdot \prod_{s=1}^{t-1} z_s$$

$$= \text{Const} \cdot \sum_i w_i^{(t)} e^{-\frac{1}{2} y_i \alpha_t f_t(\bar{x}_i)} =$$

± 1

$$= \text{Const} \cdot \left(\sum_{i:\text{прав.}} w_i^{(t)} e^{-\frac{1}{2}\alpha_t} + \sum_{i:\text{неправ.}} w_i^{(t)} e^{+\alpha_t/2} \right) =$$

$$\pm \sum_{i:\text{неправ}} w_i^{(t)} e^{-\alpha_t/2}$$

$$= \text{Const} \left(\underbrace{\sum_{i=1}^{N} w_i^{(t)} e^{-\alpha_t/2}}_{\text{Const}} + \underbrace{\left(e^{\frac{\alpha_t}{2}} - e^{-\frac{\alpha_t}{2}}\right)}_{\text{Const}} \cdot \sum_{i:\text{неправ.}} w_i^{(t)} \right)$$

$f_t \to \min$

④ Gradient boosting

$f_1, f_2, \ldots, f_T$

$$\hat{y}_i = \sum_{t=1}^{T} f_t(\bar{x}_i) = (f_1 + \ldots + f_{T-1}) + f_T(\bar{x}_i)$$

$$L = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \longrightarrow \min$$

$$L^{(T)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(T-1)} + f_T(\bar{x}_i)) + \Omega(f_T) \xrightarrow{f_T} \min$$

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$$

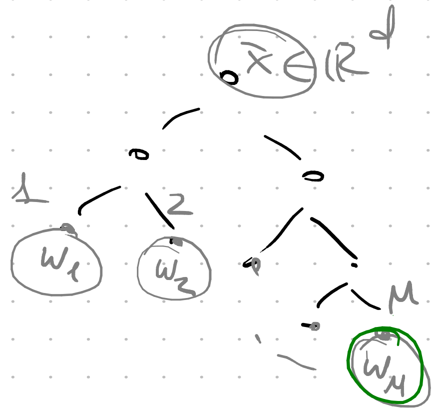
$$L^{(T)} = \sum_{i=1}^{N} \left(\left(y_i - \hat{y}_i^{(T-1)}\right) - f_t(\bar{x}_i) \right)^2 + \Omega(f_t) = \qquad \xrightarrow{f_t} \min$$

$$= \text{Const} + \sum_{i=1}^{N} \left(f_t^2(\bar{x}_i) - 2 f_t(\bar{x}_i)\left(y_i - \hat{y}_i^{(T-1)}\right) \right) + \Omega(f_T)$$

$$l(y_i, \hat{y}_i) = l\left(y_i, \hat{y}_i^{(T-1)} + f_T(\bar{x}_i)\right) \approx \overbrace{l\left(y_i, \hat{y}_i^{(T-1)}\right)}^{\text{Const}} + \underbrace{\boxed{\frac{\partial l}{\partial \hat{y}}\Big|_{\hat{y}_i^{(T-1)}}}}_{=g_i} f_T(\bar{x}_i) + \frac{1}{2} \underbrace{\boxed{\frac{\partial^2 l}{\partial \hat{y}^2}\Big|_{\hat{y}_i^{(T-1)}}}}_{=h_i} f_T^2(\bar{x}_i)$$

$$L^{(T)} \approx \sum_{i=1}^{N} \left(\frac{h_i}{2} f_T^2(\bar{x}_i) + g_i f_T(\bar{x}_i) \right) + \Omega(f_T) \xrightarrow{f_T} \min$$

5 **XGBoost** $f_t(\bar{x})$ – decision tree



$\bar{w} \in \mathbb{R}^M$, $q: \mathbb{R}^d \to \{1, \dots, M\}$

$$f_T(\bar{x}) = w_{q(\bar{x})}$$

$$\Omega(f_T) = \gamma \cdot M + \frac{\lambda}{2} \sum_{j=1}^{M} w_j^2$$

$$L^{(T)} \approx \sum_{i=1}^{N} \left(g_i \, w_{q(\bar{x}_i)} + \frac{h_i}{2} w^2_{q(\bar{x}_i)} \right) + \gamma M + \frac{\lambda}{2} \sum_{j=1}^{M} w_j^2 =$$

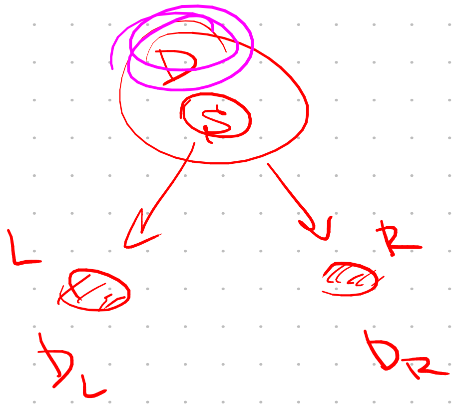
$$= \sum_{j=1}^{M} \left(w_j \underbrace{\boxed{\sum_{i: q(\bar{x}_i) = j} g_i}}_{= G_j} + \frac{1}{2} w_j^2 \underbrace{\boxed{\sum_{i: q(\bar{x}_i) = j} h_i}}_{= H_j} \right) + \gamma M + \frac{\lambda}{2} \sum_{j=1}^{M} w_j^2 =$$

$$= \sum_{j=1}^{M} \left(G_j \cdot w_j + \frac{H_j}{2} \cdot w_j^2 \right) + \gamma M + \_\_\_ =$$

$$= \sum_{j=1}^{M} \underbrace{\left(w_j \overset{q}{G_j} + \frac{1}{2} w_j^2 (\overset{q}{H_j} + \lambda) \right)}_{\xrightarrow{w_j} \min} + \gamma M \xrightarrow{q,\ \bar{w}} \min$$

$$\underbrace{w_j^* = -\frac{G_j}{H_j + \lambda}}$$

$$= \sum_{j=1}^{M} \left(-\frac{G_j^2}{H_j+\lambda} + \frac{G_j^2}{2(H_j+\lambda)} \right) + \gamma M = \boxed{-\frac{1}{2} \sum_{j=1}^{M} \frac{G_j^2}{H_j+\lambda} + \gamma M}$$



$$L_{before} = -\frac{1}{2} \sum_j \frac{G_j^2}{H_j+\lambda} + \gamma M$$

$$L_{after} = -\frac{1}{2} \sum_{j \neq s} \frac{G_j^2}{H_j+\lambda} - \frac{1}{2} \frac{G_L^2}{H_L+\lambda} - \frac{1}{2} \frac{G_R^2}{H_R+\lambda} + \gamma (M+1)$$

$$Gain_s = L_{before} - L_{after} = \frac{1}{2} \left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{G_s^2}{H_s+\lambda} \right) - \gamma$$

XGBoost: for $t = 1 \ldots T$:

- while $\exists s: Gain_s > 0$
  - grow tree f_t
- $F_t = f_1 + \ldots f_t$
- compute $\hat{y}_i^{(t)}, h_i^{(t)}, g_i^{(t)}$