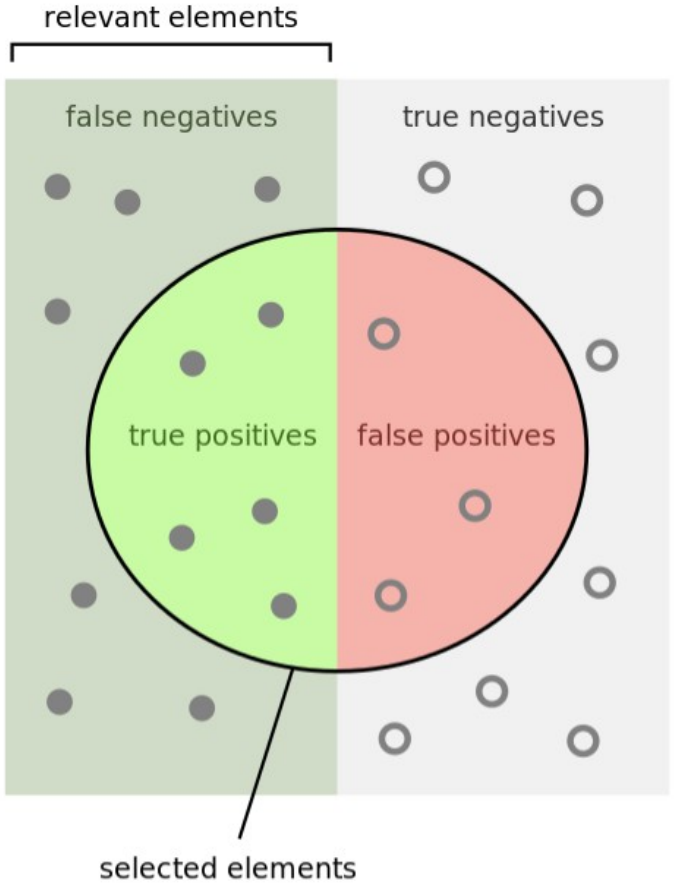
Деревья

Рыжиков Артём, НИУ ВШЭ,
17.11.2018

Offtop. Метрики качества



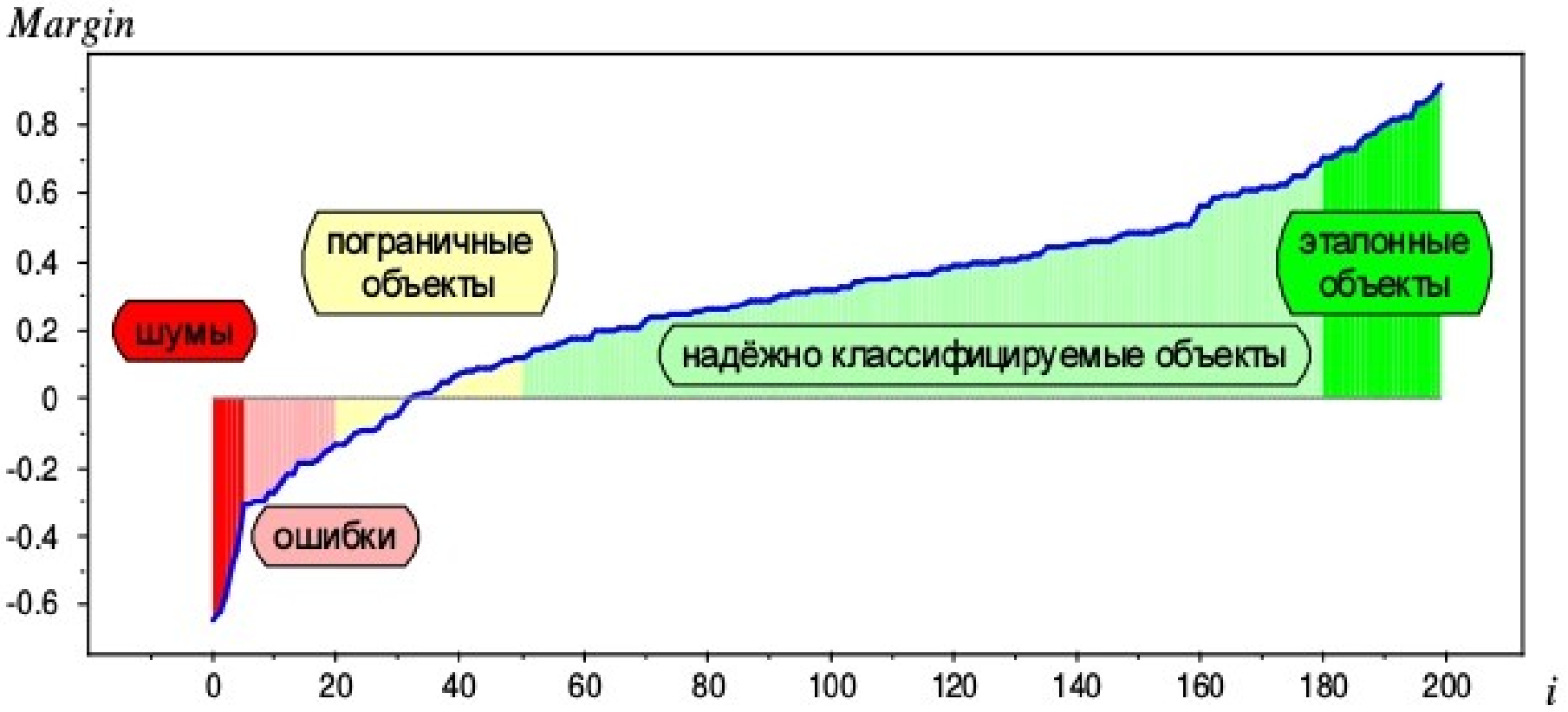
How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

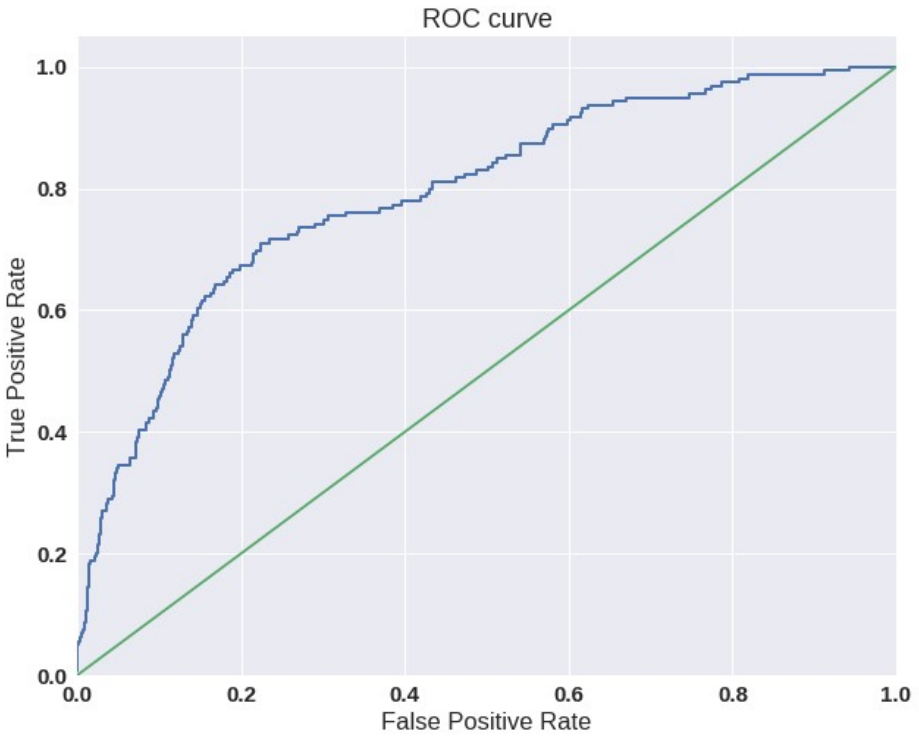
How many relevant items are selected?

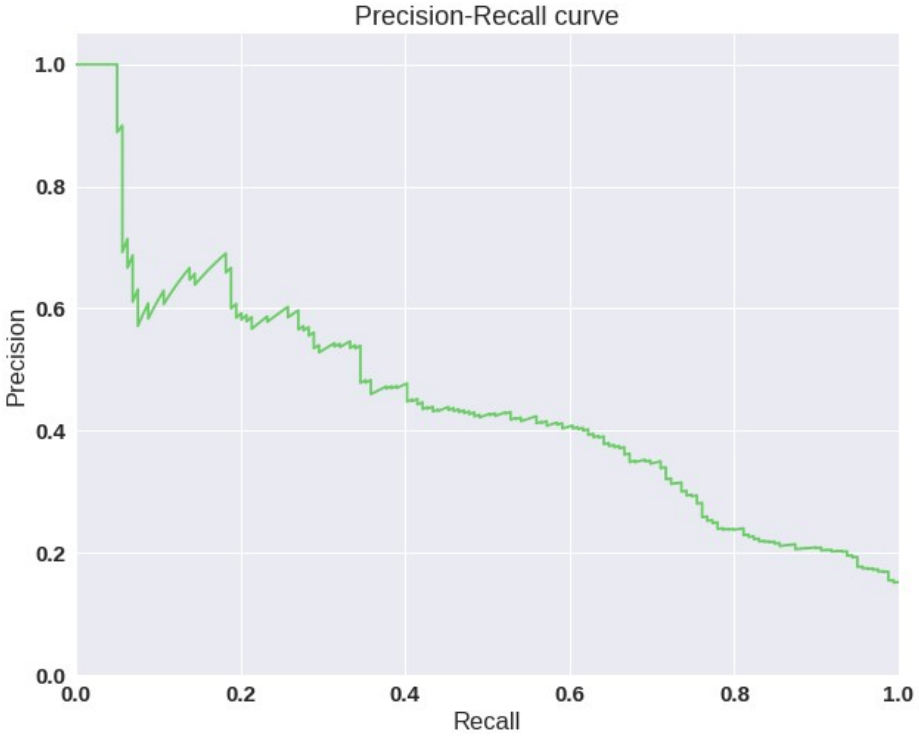
$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

Threshold


Margin
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
20
40
60
80
100
120
140
160
180
200
i
шумы
пограничные объекты
надёжно классифицируемые объекты
эталонные объекты
ошибки

Метрики

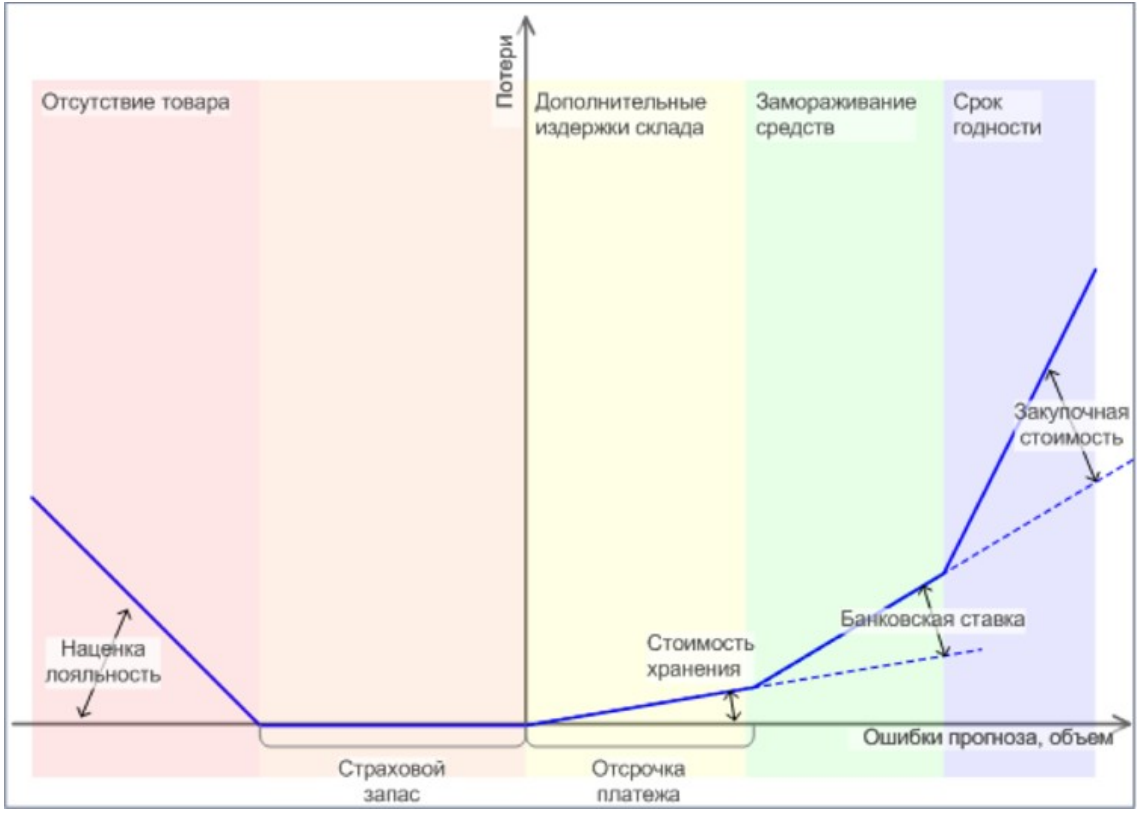

ROC curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall curve
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Ссылки

- Habr
- Подробнее про различие между ROC AUC и PR AUC

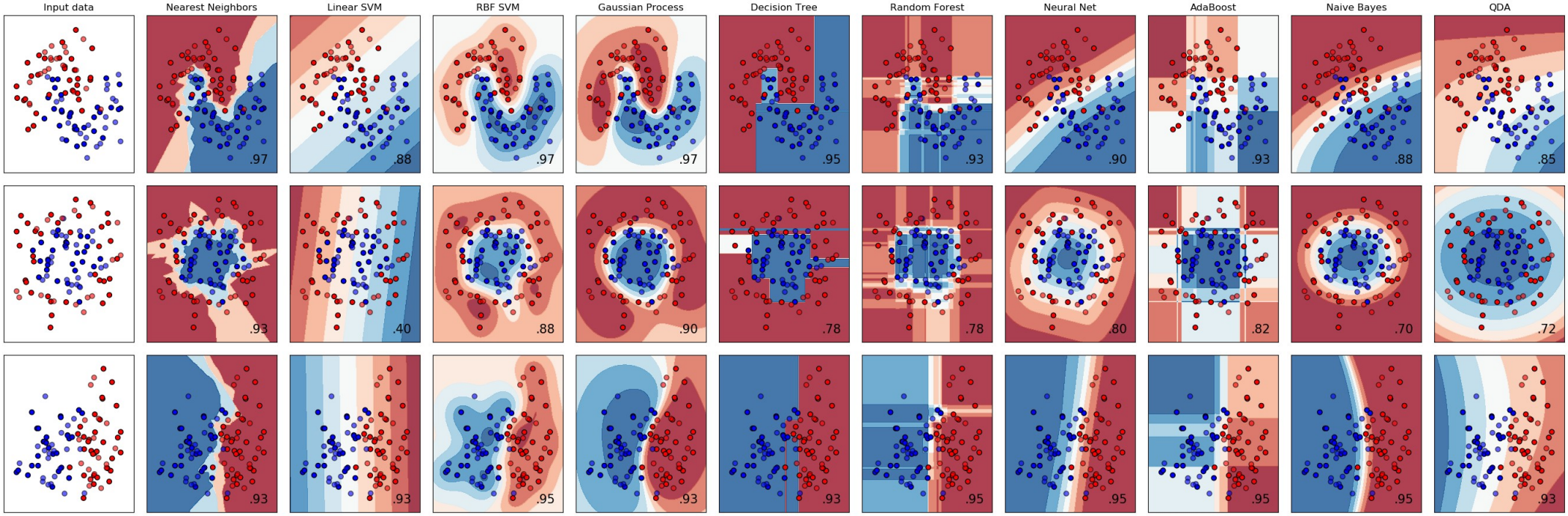
Метрики в задаче регрессии



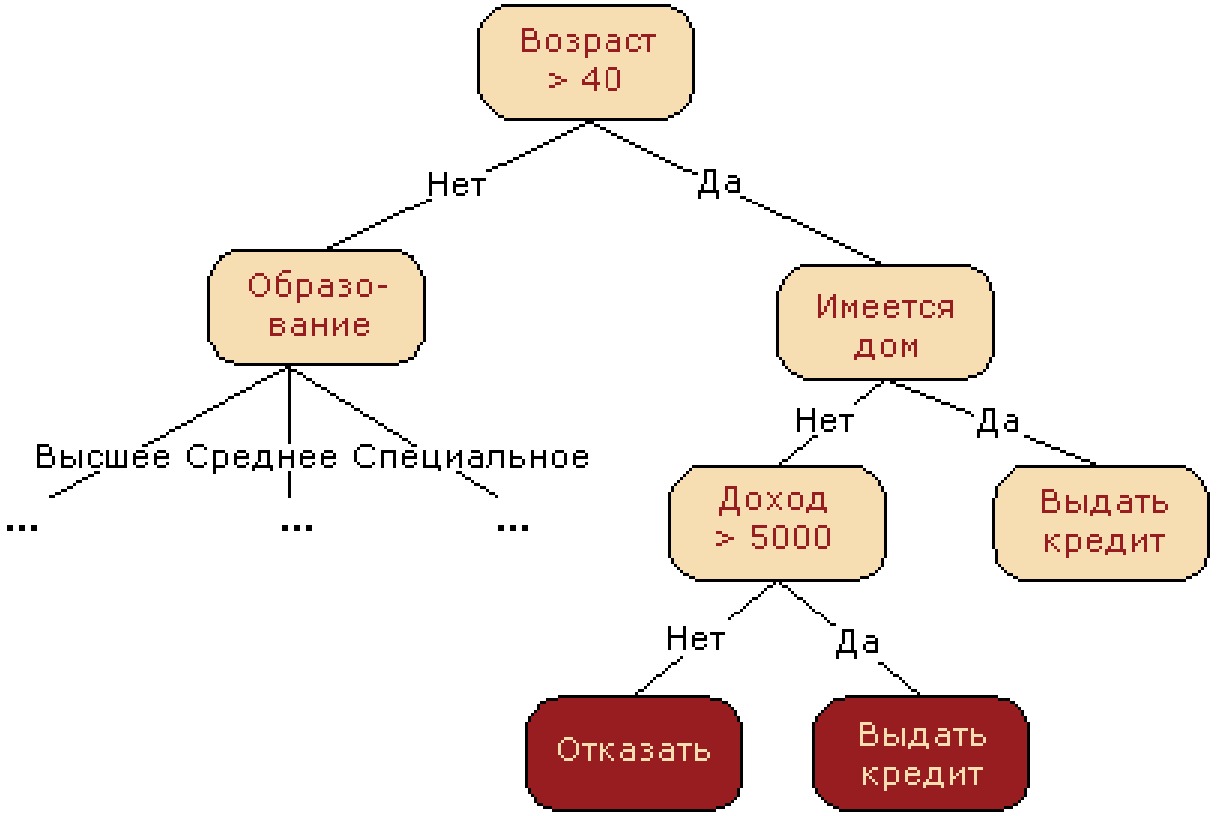
Ссылки

- Лекции Евгения Соколова

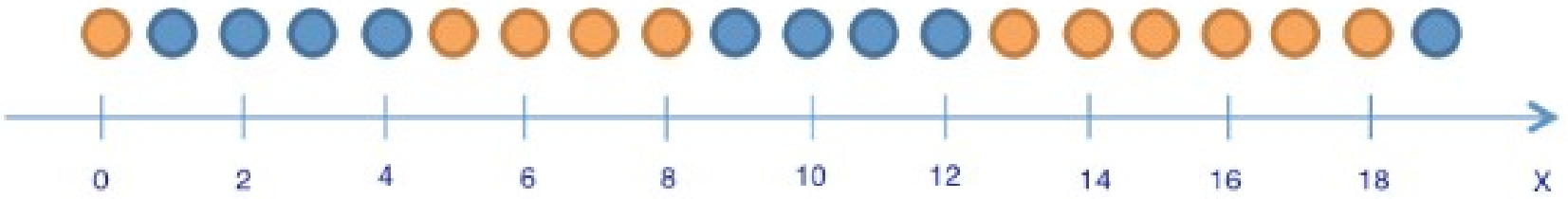
Логические методы классификации



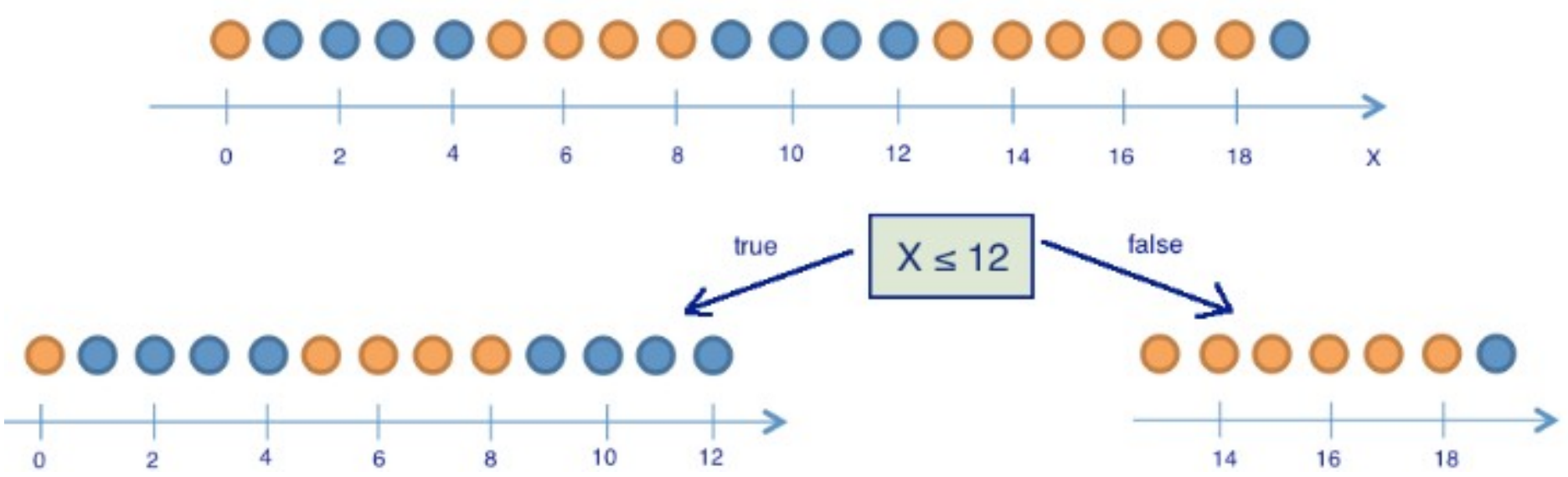
Дерево принятия решений



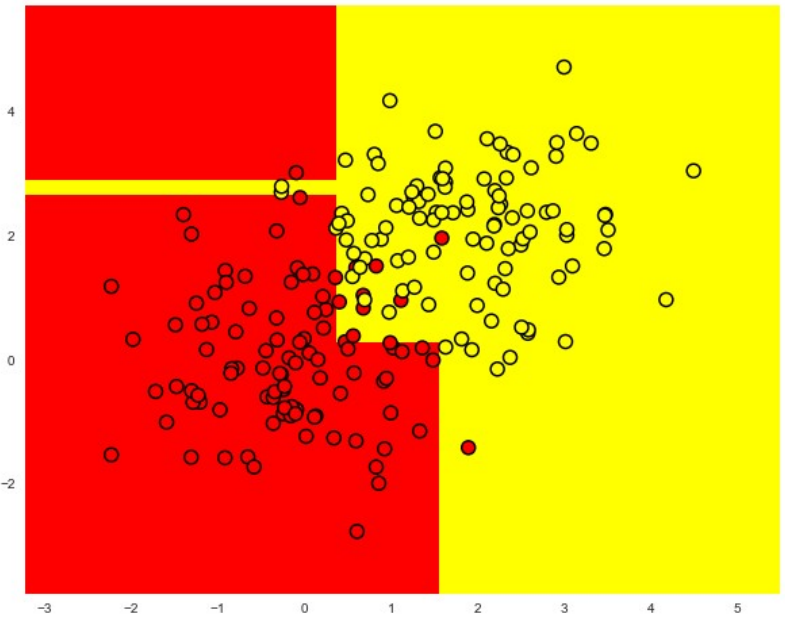
Информационный прирост

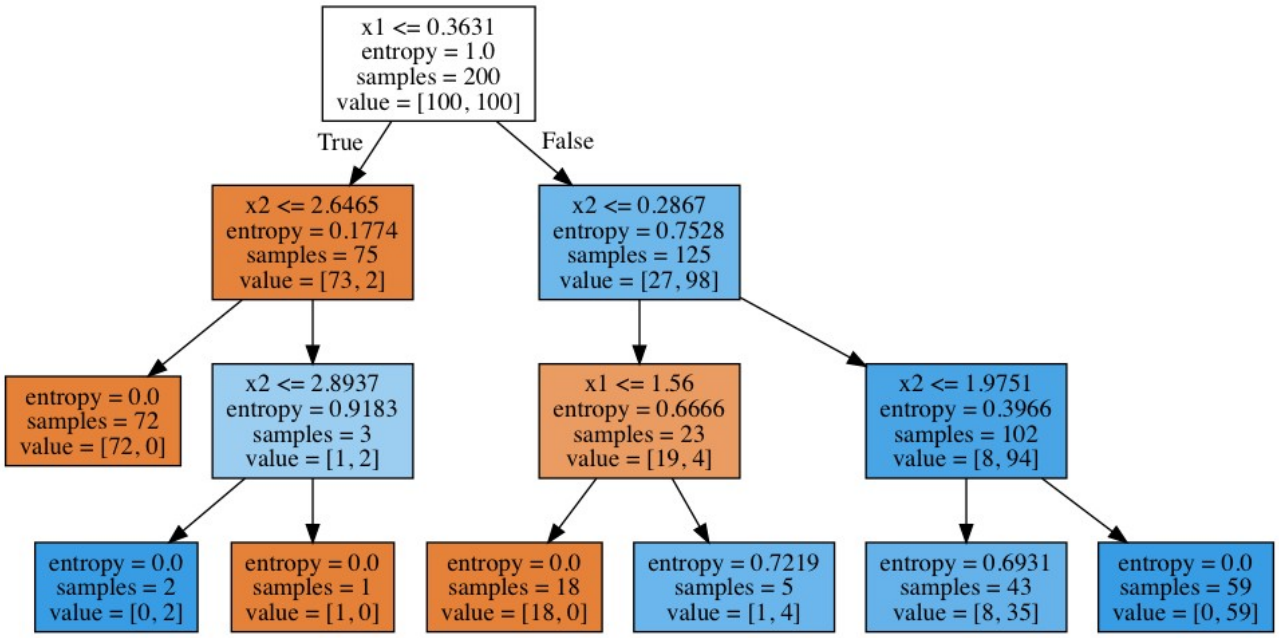


Информационный прирост

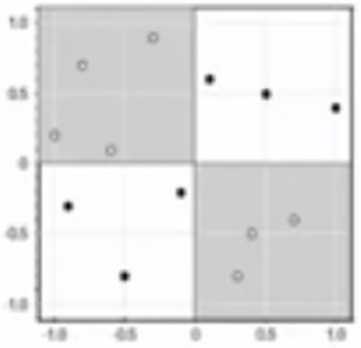


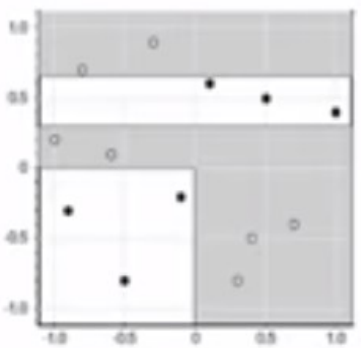
Пример дерева



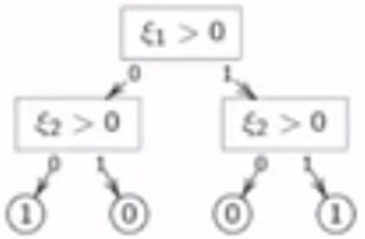


Переобучаемость ID3

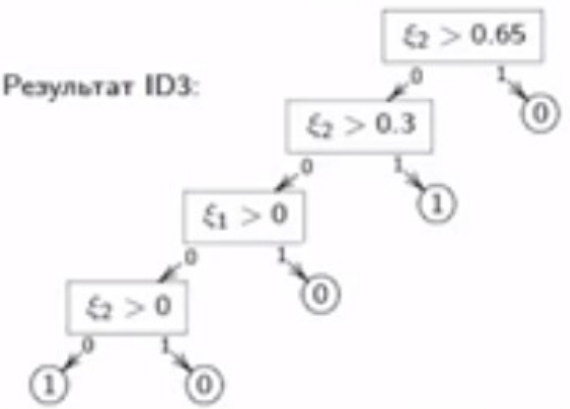




Оптимальное дерево для задачи XOR:



Результат ID3:



К.В.Воронцов, Видеолекции

C4.5, CART

Редукция дерева («стрижка», pruning: C4.5, CART)

X^k — независимая контрольная выборка, $k \approx 0.5\ell$.

1: **для всех** $v \in V_{\text{внутр}}$
2: $S_v :=$ подмножество объектов X^k, дошедших до v;
3: **если** $S_v = \varnothing$ **то**
4: **вернуть** новый лист v, $c_v :=$ Мажоритарный класс(U);
5: число ошибок при классификации S_v четырьмя способами:
$r(v)$ — поддеревом, растущим из вершины v;
$r_L(v)$ — поддеревом левой дочерней вершины L_v;
$r_R(v)$ — поддеревом правой дочерней вершины R_v;
$r_c(v)$ — к классу $c \in Y$.
6: в зависимости от того, какое из них минимально:
сохранить поддерево v;
заменить поддерево v поддеревом L_v;
заменить поддерево v поддеревом R_v;
заменить поддерево v листом, $c_v := \arg\min_{c \in Y} r_c(v)$.

Ссылки

- Видеолекции
- Habr