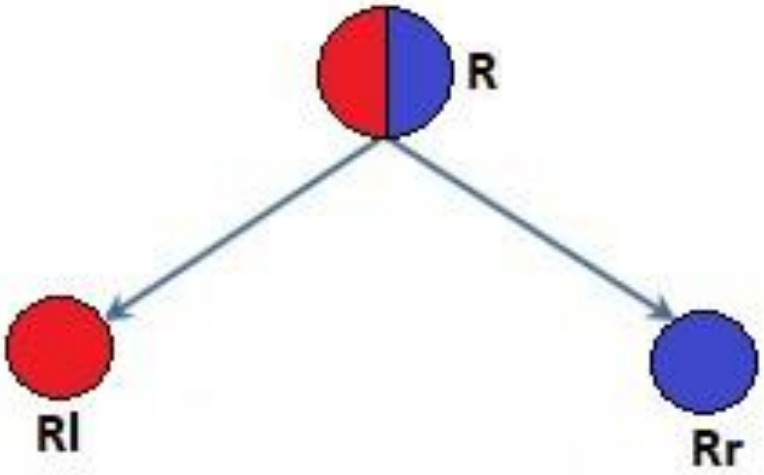
КРИТЕРИИ ИНФОРМАТИВНОСТИ

В каждой вершине оптимизируем функционал $Q(X, j, t)$.

- Пусть R – множество объектов, попадающих в вершину на данном шаге, а R_l и R_r - объекты, попадающие в левую и правую ветки после разбиения.

Цель: хотим, чтобы после разбиения объектов на две группы внутри каждой группы как можно больше объектов было одного класса.



КРИТЕРИИ ИНФОРМАТИВНОСТИ

- Пусть R – множество объектов, попадающих в вершину на данном шаге, а R_l и R_r - объекты, попадающие в левую и правую ветки после разбиения.

Цель: хотим, чтобы после разбиения объектов на две группы внутри каждой группы как можно больше объектов было одного класса.

- Функция $H(R)$ - критерий информативности - оценивает меру однородности целевых переменных внутри группы R.

- Чем меньше разнообразие целевой переменной внутри группы, тем меньше значение $H(R)$. То есть хотим

$$\boldsymbol{H(R_l) \to min, H(R_r) \to min}$$

КРИТЕРИИ ИНФОРМАТИВНОСТИ

- Пусть R – множество объектов, попадающих в вершину на данном шаге, а R_l и R_r - объекты, попадающие в левую и правую ветки после разбиения.

Цель: хотим, чтобы после разбиения объектов на две группы внутри каждой группы как можно больше объектов было одного класса.

- Чем меньше разнообразие целевой переменной внутри группы, тем меньше значение $H(R)$. То есть

$$\boldsymbol{H(R_l) \to min, H(R_r) \to min}$$

- Определим функционал Q по формуле:

$$\boldsymbol{Q(R, j, t) = H(R) - \frac{|R_l|}{|R|} H(R_l) - \frac{|R_r|}{|R|} H(R_r) \to \max_{j,t}}$$

H(R) В ЗАДАЧЕ РЕГРЕССИИ

$$\boldsymbol{H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - \bar{y})^2},$$

где $\bar{y} = \frac{1}{|R|} \sum_{(x_j, y_j) \in R} y_j$

- Значит, информативность в листе – это дисперсия целевой переменной (для объектов, попавших в этот лист). Чем меньше дисперсия, тем меньше разброс целевой переменной объектов, попавших в лист.

H(R) В ЗАДАЧЕ КЛАССИФИКАЦИИ

Будем в каждой вершине в качестве ответа выдавать не класс, а распределение вероятностей классов:
$c = (c_1, \dots, c_K), \sum_i c_i = 1.$

- $\boldsymbol{H(R) = \sum_{k=1}^{K} p_k(1 - p_k)}$ (критерий Джини).
- $\boldsymbol{H(R) = -\sum_{k=1}^{K} p_k \log p_k}$ (энтропийный критерий)