Решающие деревья

Кантонистова Е.О.

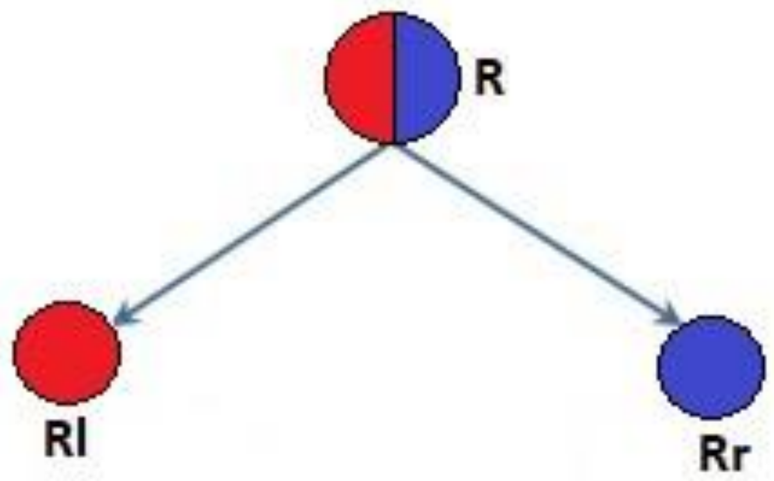
ВШЭ, 2019

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:

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

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

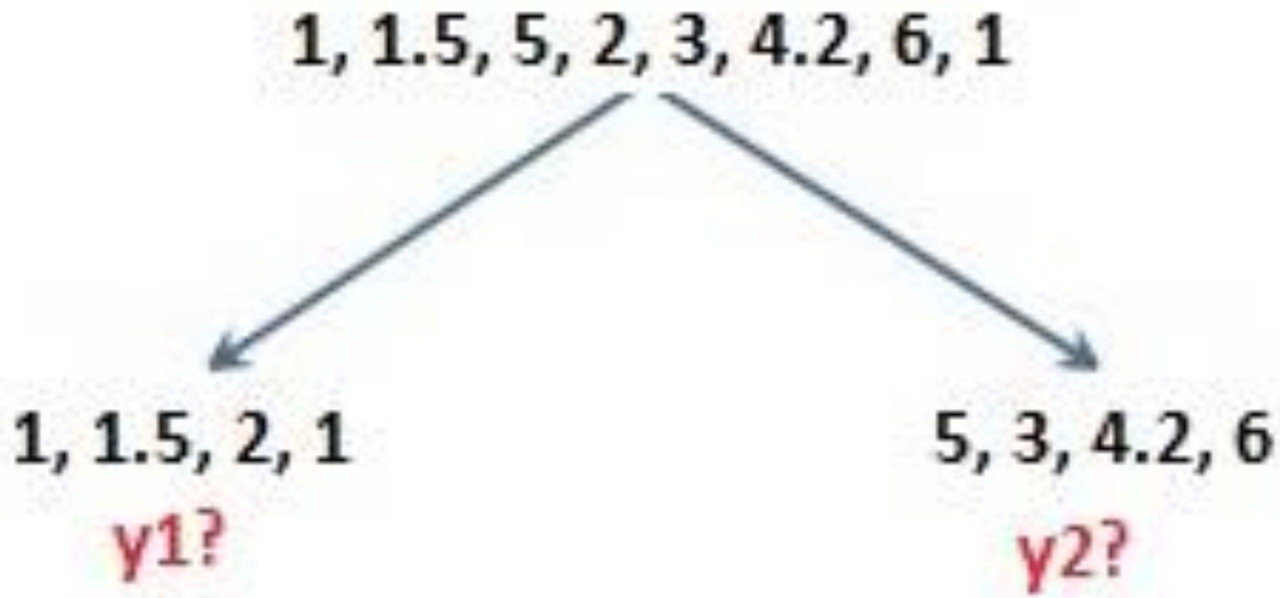
$$\boldsymbol{H(R_l) \to min, H(R_r) \to min}$$

- Определим функционал Q по формуле:

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

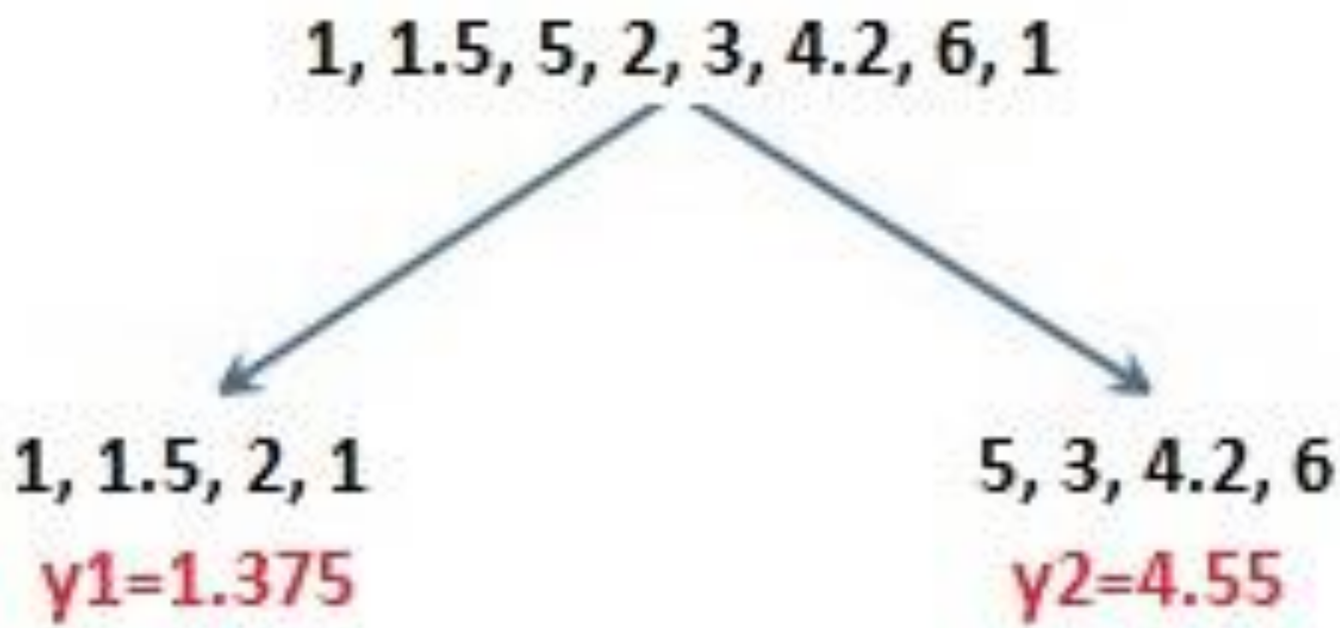
ПРИМЕР: РЕШАЮЩЕЕ ДЕРЕВО В ЗАДАЧЕ РЕГРЕССИИ

Предположим, что в лист дерева попало несколько объектов. В каждом листе дерево предсказывает константу. Какую константу выгоднее всего выдать в качестве ответа?



ПРИМЕР: РЕШАЮЩЕЕ ДЕРЕВО В ЗАДАЧЕ РЕГРЕССИИ

Если в качестве функционала ошибки в листе использовать среднеквадратичную ошибку, то в качестве ответа надо выдавать среднее значение целевых переменных всех объектов, попавших в лист.



Вид критерия информативности

- В каждом листе дерево выдает константу c (вещественное число – в регрессии, класс или вероятность класса – в классификации).
- Чем лучше объекты в листе предсказываются этой константой, тем меньше средняя ошибка на объектах:

$$\boldsymbol{H(R) = \min_{c \in \mathbb{R}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c)},$$

где $L(y, c)$ – некоторая функция потерь.

H(R) В ЗАДАЧАХ РЕГРЕССИИ

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

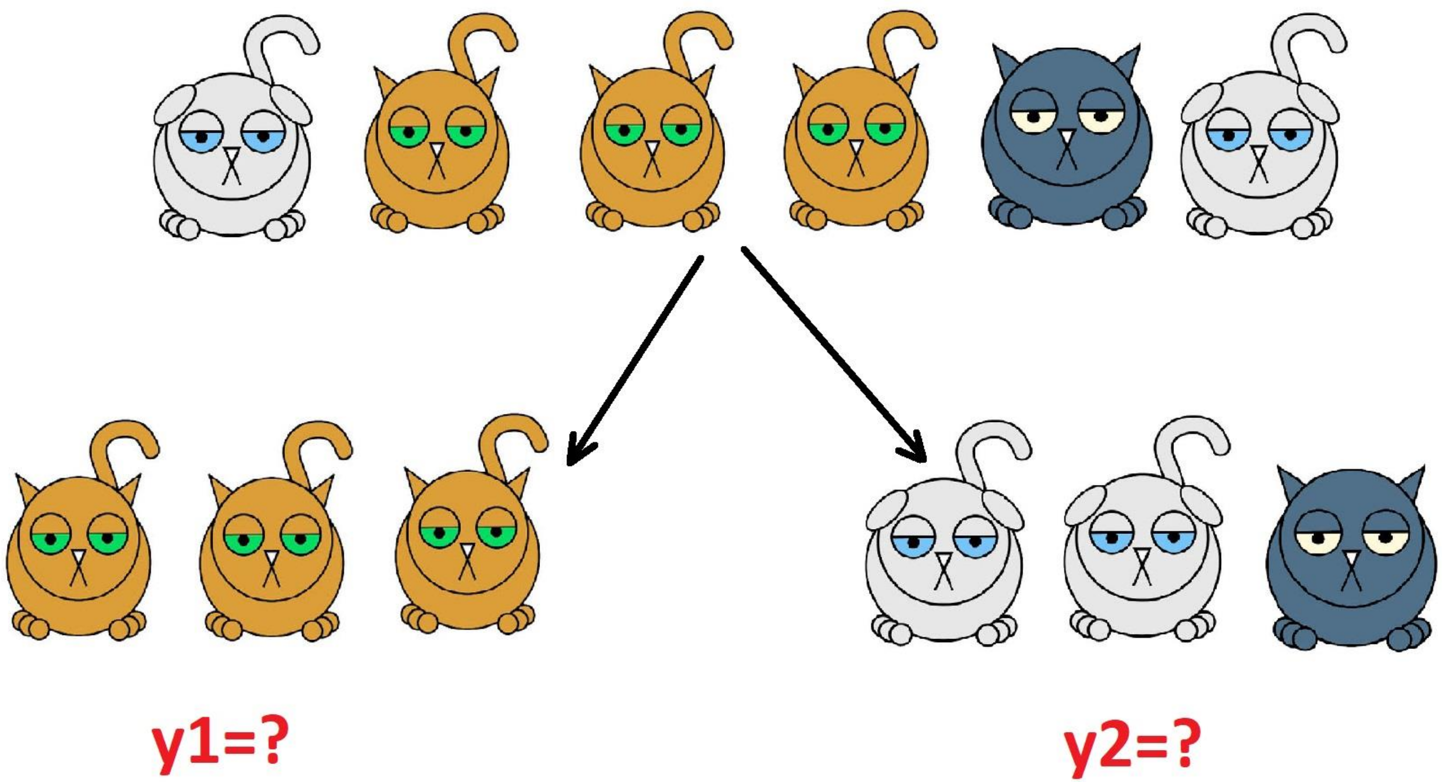
- Минимум будет достигаться, если c – это среднее значение целевой переменной, то есть

$$H(R) = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_j, y_j) \in R} y_j \right)^2$$

- 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.

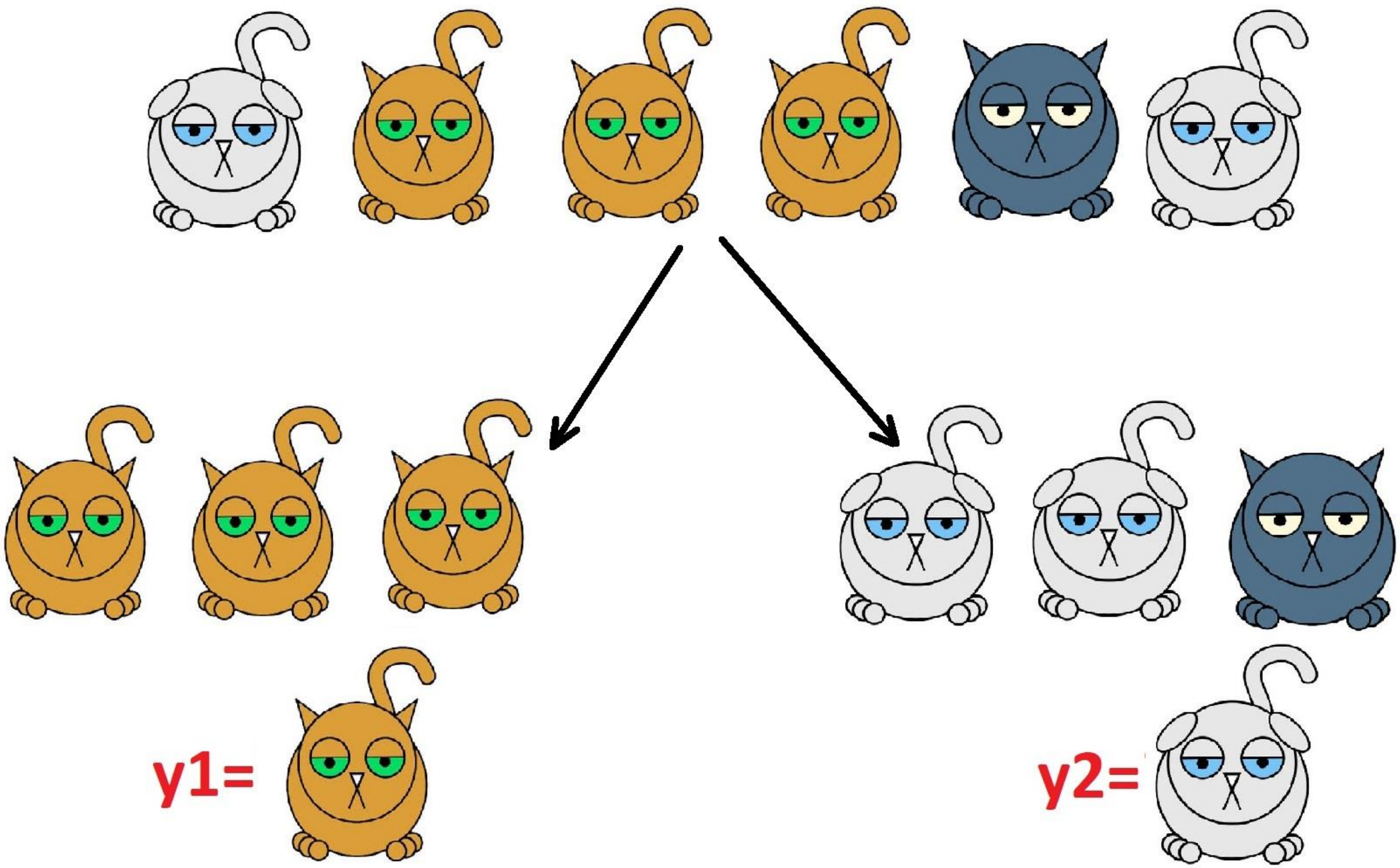
ПРИМЕР: РЕШАЮЩЕЕ ДЕРЕВО В ЗАДАЧЕ КЛАССИФИКАЦИИ

Предположим, что в лист дерева попало несколько объектов. В каждом листе дерево предсказывает класс объекта. Какой класс выгоднее всего выдать в качестве ответа?



Пример: решающее дерево в задаче классификации

Разумнее всего в качестве ответа в листе выдавать самый представительный класс.



H(R) в задачах классификации

Решаем задачу классификации с K классами: $1, 2, \ldots, K$.

- Пусть p_k доля объектов класса k, попавших в вершину:

$$p_k = \frac{1}{|R|} \sum_{(x_i, y_i) \in R} [y_i = k]$$

- Пусть k_* - самый представительный класс в данной вершине:

$$k_* = \underset{k}{argmax}\, p_k$$

H(R) В ЗАДАЧАХ КЛАССИФИКАЦИИ

Решаем задачу классификации с K классами: $1,2,\dots,K$.

- Пусть p_k доля объектов класса k, попавших в вершину:

$$p_k = \frac{1}{|R|} \sum_{(x_i,y_i)\in R} [y_i = k]$$

- Пусть k_* - самый представительный класс в данной вершине:

$$k_* = \underset{k}{argmax}\, p_k$$

Ошибка классификации:

$$H(R) = \min_{c\in Y} \frac{1}{|R|} \sum_{(x_i,y_i)\in R} \boldsymbol{[y_i \neq c]}$$

H(R) В ЗАДАЧАХ КЛАССИФИКАЦИИ

Решаем задачу классификации с K классами: $1,2,\dots,K$.

- Пусть p_k доля объектов класса k, попавших в вершину:

$$p_k = \frac{1}{|R|} \sum_{(x_i,y_i)\in R} [y_i = k]$$

- Пусть k_* - самый представительный класс в данной вершине:

$$k_* = \underset{k}{argmax}\, p_k$$

Ошибка классификации:

$$H(R) = \min_{c\in Y} \frac{1}{|R|} \sum_{(x_i,y_i)\in R} [y_i \neq c]$$

$$\boldsymbol{H(R) = \frac{1}{|R|} \sum_{(x_i,y_i)\in R} [y_i \neq k_*] = 1 - p_{k_*}}$$

H(R) В ЗАДАЧАХ КЛАССИФИКАЦИИ

Критерий Джини

- Будем в каждой вершине в качестве ответа выдавать не класс, а распределение вероятностей классов: $c = (c_1, \dots, c_K), \sum_i c_i = 1.$
- Качество распределения можно измерить с помощью критерия Бриера:

$$H(R) = \min_c \frac{1}{|R|} \sum_{(x_i,y_i)\in R} \sum_{k=1}^{K} (c_k - [y_i = k])^2$$

Утверждение.

1) Минимальное значение функционала $H(R)$ достигается на векторе, состоящем из долей классов: $c_* = (p_1, \dots, p_K)$

2) На векторе c_* функционал (*) переписывается в виде

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

H(R) в задачах классификации

Энтропийный критерий

Запишем логарифм правдоподобия:

$$H(R) = \min_c \left(-\frac{1}{|R|} \sum_{(x_i, y_i) \in R} \sum_{k=1}^{K} [y_i = k] \log c_k \right) \quad (*)$$

На векторе $c_* = (p_1, \dots, p_K)$ функционал (*) записывается в виде

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

(энтропия распределения)

H(R) В ЗАДАЧАХ КЛАССИФИКАЦИИ

Энтропийный критерий

Запишем логарифм правдоподобия:

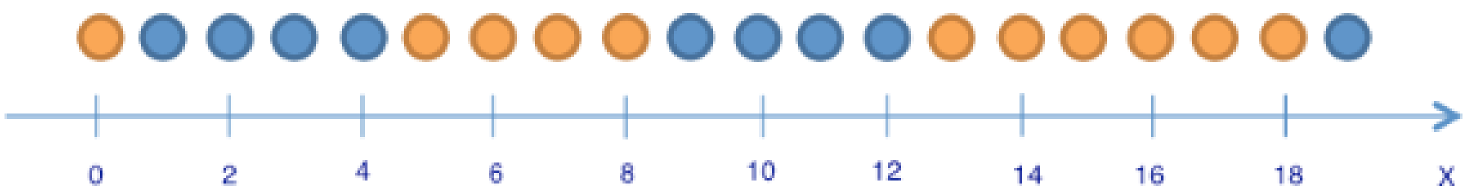
$$H(R) = \min_{c}\left(-\frac{1}{|R|}\sum_{(x_i,y_i)\in R}\sum_{k=1}^{K}[y_i = k]\log c_k\right)\ (*)$$

На векторе $c_* = (p_1, \dots, p_K)$ функционал (*) записывается в виде

$$\boldsymbol{H(R) = -\sum_{k=1}^{K} p_k \log p_k}$$ (энтропия)

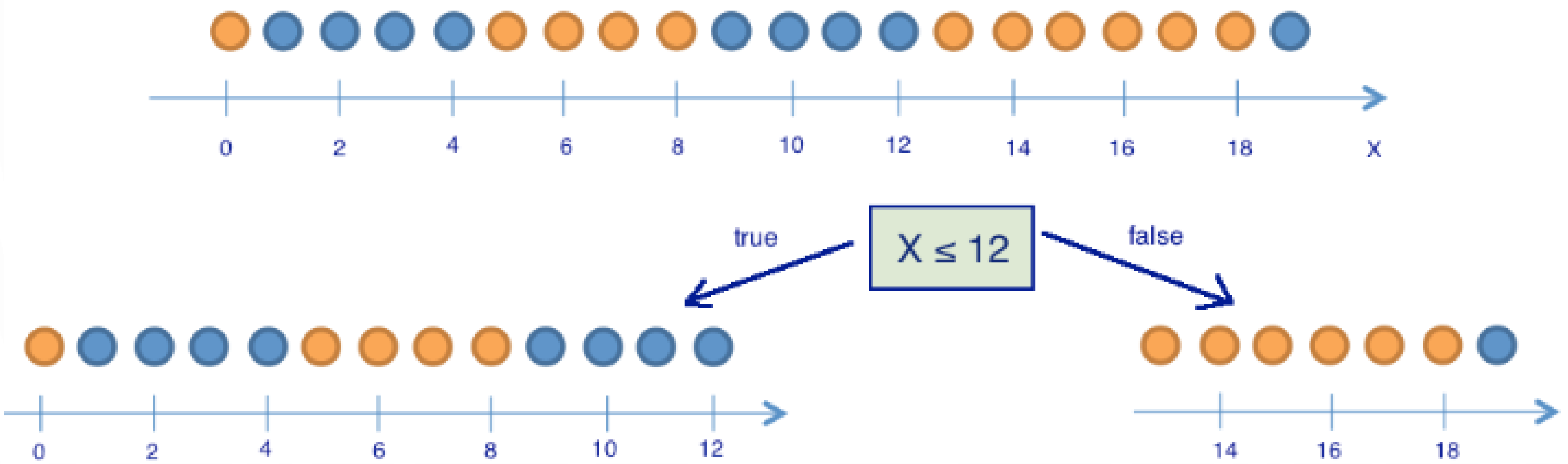
- **Энтропия $H(R) \geq 0$ (минимум на распределении $p_i = 1, p_j = 0, j \neq i$)**
- **$max\ H(R)$ достигается на равномерном распределении $p_1 = \dots = p_K = \frac{1}{K}$.**

ПРИМЕР ИСПОЛЬЗОВАНИЯ ЭНТРОПИЙНОГО КРИТЕРИЯ



- $p_1 = \frac{9}{20}, p_2 = \frac{11}{20} \Rightarrow$ энтропия $H_0 = -\frac{9}{20}\log\frac{9}{20} - \frac{11}{20}\log\frac{11}{20} \approx 1$

ПРИМЕР ИСПОЛЬЗОВАНИЯ ЭНТРОПИЙНОГО КРИТЕРИЯ



- В левой части $H_l = -\frac{5}{13}\log\frac{5}{13} - \frac{8}{13}\log\frac{8}{13} \approx 0.96$
- В правой части $H_r = -\frac{1}{7}\log\frac{1}{7} - \frac{6}{7}\log\frac{6}{7} \approx 0.6$

То есть $Q = H_0 - \frac{|R_l|}{R}H_l - \frac{|R_r|}{|R|}H_r = 1 - \frac{13}{20}\cdot 0.96 - \frac{7}{20}\cdot 0.6 \approx 0.16$