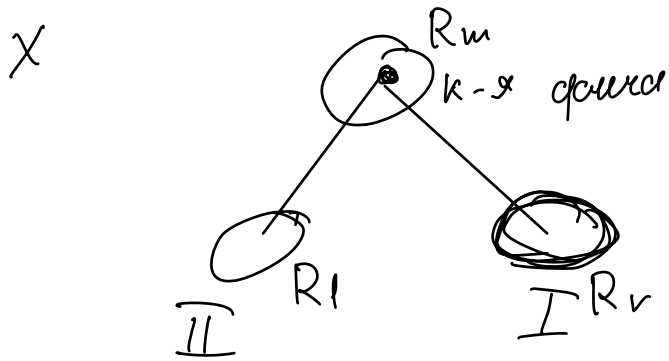
Выбор предикатов



R_m - мн-во объектов

$$Q(R_m) = H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_r)$$

$H(R)$ - критерий информативности.

$H(R) \to \min$

$Q(R_m) \to \max$

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x,y) \in R} L(y, c)$$

Регрессия

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x,y) \in R} (y - c)^2$$

$$H(R) = \min_{c \in Y} \frac{1}{|R|} \sum_{(x,y) \in R} (y - \bar{y})^2, \quad \bar{y} = \frac{1}{|R|} \sum y$$

Классификация

R, $k \in \{1 \dots K\}$

p_k – доля объектов класса k в R

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

k^* – наиболее мощный класс

$k^* = \arg\max_k p_k$

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

$c = p_{k^*}$

$\Leftarrow \quad 1 - p_{k^*}$

Критерий Джини

$c = (c_1, \dots, c_K)$, $\sum_{k=1}^{K} c_k = 1$

Критерий Бриера

$$H(R) = \min_{\sum_k c_k = 1} \frac{1}{|R|} \sum_{(x,y) \in R} \sum_{k=1}^{K} \left(c_k - [y_i = k] \right)^2$$

$$c^* = (p_1, \ldots, p_K)$$

$$H(R) = \sum_{k=1}^{K} p_k (1 - p_k) = \sum_{k=1}^{K} (p_k - p_k^2) =$$

$$= \underbrace{\sum_{k=1}^{K} p_k}_{1} - \sum_{k=1}^{K} p_k^2$$

Задача 1. $H(R) = \sum_{k \neq k'} p_k p_{k'} = \sum_{k=1}^{K} \sum_{k' \neq k}^{K} p_k p_{k'} =$

$$= \sum_{k=1}^{K} p_k \left(\sum_{k' \neq k}^{K} p_{k'} \right) = \sum_{k=1}^{K} p_k (1 - p_k)$$

Задача 2. Есть вершина т. объекты R $a(x)$, которая выбирает класс случайно. но класс k выбирается с вер-ю p_k E(частота ошибок) равно индексу Джини.

$$E \frac{1}{|R|} \sum_{x,y} [y \neq a(x)] = \frac{1}{|R|} \sum_{x,y} E[y \neq a(x)] =$$

$$= \frac{1}{|R|} \sum_{x,y} (1 - p_y) =$$

$$= \frac{1}{|R|} \sum_{x,y} \sum_{k=1}^{K} [y_i = k](1 - p_k) =$$

$$= \sum_{k=1}^{K} \underbrace{\frac{\sum_{x,y} [y_i = k]}{|R|}} (1 - p_k) = \sum_{k=1}^{K} p_k (1 - p_k)$$

$$Q(R_m) = H(R_m) - \frac{|R_l|}{|R_m|} H(R_l) - \frac{|R_r|}{|R_m|} H(R_r) \to \max$$

$$= -\frac{1}{|R_m|} \left(|R_l| H(R_l) + |R_r| H(R_r) \right)$$

p_{mk} – доля k в вершине m

$$= -\frac{1}{|R_m|} \left(|R_l| - |R_l| \cdot \left(\sum_{k=1}^{K} p_{lk}^2 \right) + |R_r| - |R_r| \left(\sum_{k=1}^{K} p_{rk}^2 \right) \right) =$$

$$|R_l| \left(\sum_{k=1}^{K} p_{lk}^2 \right) + |R_r| \left(\sum_{k=1}^{K} p_{rk}^2 \right)$$

R_m

Число пар (x_i, x_j) т. ч. оба объекта попадут в одно и то же поддерево и $y_i = y_j$

$p_{lk} |R_l|$ $\left(\sum_{k=1}^{K} p_{lk}^2 \right) |R_l|^2 + \left(\sum_{k=1}^{K} p_{rk}^2 \right) |R_r|^2$

Энтропийный критерий

$$\prod_{k=1}^{K} p_k^{[y_i = k]}$$

$$H(R) = \min_{\sum c_k = 1} \left(-\frac{1}{|R|} \sum_{x,y} \sum_{k=1}^{K} [y_i = k] \log c_k \right) \to \min_{c_k}$$

$$L(c, \lambda) = -\frac{1}{|R|} \sum_{x,y} \sum_{k=1}^{K} [y_i = k] \log c_k + \lambda \sum_{k=1}^{K} c_k \to \min_{c_k}$$

$$\frac{\partial L}{\partial c_k} = -\frac{1}{|R|} \sum_{x,y} [y_i = k] \frac{1}{c_k} + \lambda \qquad \frac{p_k}{c_k} + \lambda = 0$$

$$c_k = p_k / \lambda$$

$$1 = \frac{1}{\lambda} \sum_{k=1}^{K} p_k = \frac{1}{\lambda} \Rightarrow \lambda = 1 \qquad c_k = p_k$$

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

$$\log_2 K$$