2.4 Criteria of branching quality.

It is necessary to have a criterion, which will allow to compare all various variants of node branching and to choose the best of them.

The frequency of mistakes (PR problem) or a relative variance (RA problem) can be considered as such criterion.

Let the node be divided on *L* new nodes.

Let the number of observations appropriate to *l-th* new node be N_l,

$Data_l$ will be a set of these observations,

$\hat{Y}(l)$ will be the decision, attributed to *l-th* node and

N_l^{ω} will be a number of observations of ω–th class, which corresponds to *l-th* node (PR problem).

The general number of observations in the initial node will be *N*. Formulas for the calculation of the criteria are similar to which were used at defining the decision tree quality (because the variant of branching also represents a tree):

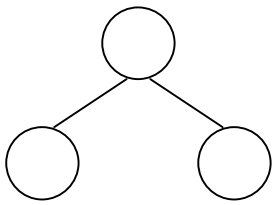
$$N_{err} = \sum_{l=1}^{L} \sum_{\substack{\omega=1 \\ \omega \neq \hat{Y}(S)}}^{K} N_l^{\omega} \text{ (PR problem);}$$

$$d_{om} = \frac{d_{oc}}{d_0} ,$$

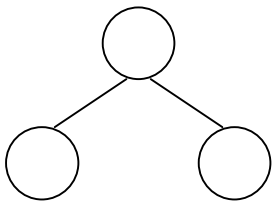
where

$$d_{oc} = \frac{1}{N} \sum_{l=1}^{L} \sum_{i \in Data_l} (\hat{Y}(l) - y^i)^2 , \; d_0 = \frac{1}{N} \sum_{i=1}^{N} (y^i - \bar{y})^2 , \; \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y^i .$$

For the PR problem more precisely methods of definition exist, as the objects of different classes are divided (in literature the term *"impurity" is used*, which can be interpreted as a degree of "pollution" of observations by objects from another classes). For example, we consider two variants of division (figure 11).



$N_1^1=0, \quad N_2^1=20,$
$N_1^2=20, \quad N_2^2=10,$
$N_{err}=10$



$N_1^1=5, \quad N_2^1=25,$
$N_1^2=15, \quad N_2^2=5,$
$N_{err}=10$

Fig. 11

The numbers of mistakes for these variants coincide, however it is clear, that the first variant is more preferable, since one of the new nodes does not need more branching because all objects in it are correctly referred to the same class.

To take into account similar cases, for the definition of division quality, it is possible to use entropy criterion or Gini's criterion.

The entropy criterion of splitting is defined by the formula:

$$H(L)=\sum_{l=1}^{L}\frac{N_l}{N}\sum_{\omega=1}^{K}-\frac{N_l^{\omega}}{N_l}log\frac{N_l^{\omega}}{N_l}=\frac{1}{N}\left(\sum_{l=1}^{L}N_l\,log\,N_l-\sum_{l=1}^{L}\sum_{\omega=1}^{K}N_l^{\omega}\,log\,N_l^{\omega}\right)$$

The lesser the entropy value is, the more information contains in the variant of division. Let the entropy for the initial node denoted as

$$H(0)=\sum_{\omega=1}^{K}\frac{N^{\omega}}{N}log\frac{N^{\omega}}{N},$$

where N^{ω} means the number of observations of ω-th class.

It is possible to use for given branching the value *gain=H(L)-H(0)* as a measure of "usefulness" or "gain".

Note the next properties of entropy criterion:

1) If the number of classes is constant and frequencies of various classes converge to each other, then value *H* is increased.
2) If various classes are equiprobable and the number of classes is increasing, then *H* is increasing logarithmically (i.e. proportionally to $\log_2 K$).

Some researches recommend that it is better to use *L* as the basis of *log*.

Gini's criterion for splitting is calculated by the following formula:

$$G(L)=\sum_{l=1}^{L}\frac{N_l}{N}\left(1-\sum_{\omega=1}^{K}\left(\frac{N_l^{\omega}}{N_l}\right)^2\right)=1-\frac{1}{N}\sum_{l=1}^{L}\sum_{\omega=1}^{K}\frac{\left(N_l^{\omega}\right)^2}{N_l}$$

The smallest value of this parameter corresponds to the best division of objects.

For the definition of quality, one can also use a parameter of branching *"gain"*, which is defined as a difference between the value of the given criterion for the initial node and the value of the variant of its division.