1. What is a decision tree?

Let us consider the following example of a recognition problem. During a doctor‘s examination of some patients the following characteristics are determined:

X_1 - temperature, X_2 - coughing, X_3 - a reddening throat,

$Y=\{W_1,W_2,W_3,W_4,W_5\}$= {a cold, quinsy, the influenza, a pneumonia, is healthy} - a set from the possible diagnoses, demanding more profound inspection.

It is required to find a model, where Y depends on X. The example (figure 1) illustrates such a model, which can be seen as a decision tree.

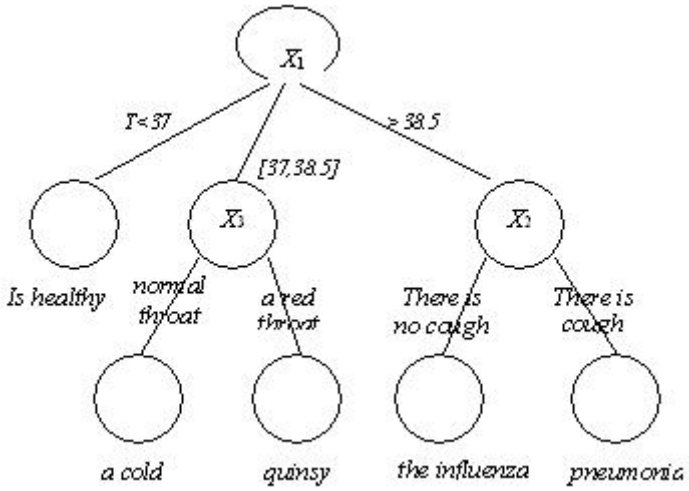

Fig. 1

The ordinary tree consists of one root, branches, nodes (places where branches are divided) and leaves. In the same way the decision tree consists of nodes which stand for circles, the branches stand for segments connecting the nodes. A decision tree is usually drawn from left to right or beginning from the root downwards, so it is easier to draw it. The first node is a root. The end of the chain " root - branch - node-... - node " is called "leaf". From each internal node (i.e. not a leaf) may grow out two or more branches. Each node corresponds with a certain characteristic and the branches correspond with a range of values. These ranges of values must give a partition of the set of values of the given characteristic.

When precisely two branches grow out from an internal node (the tree of such type is called a dichotomic tree), each of these branches can give a true of false statement concerning the given characteristic as is shown on figure 2.

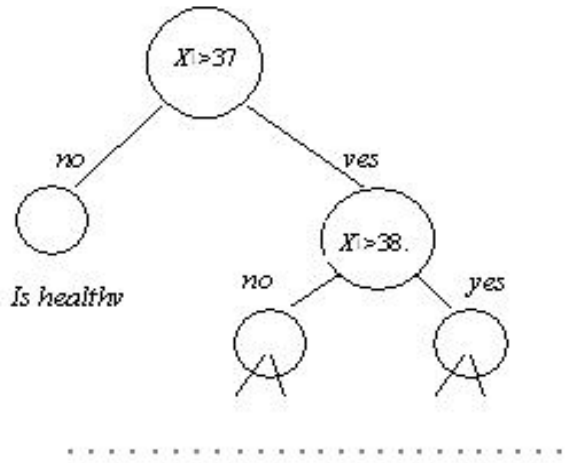

Fig. 2

The value Y is ascribed for each terminal node of a tree (named "leaf"). In case of pattern recognition problem the given value is a certain class, and in a regression analysis case the given value represents a real number. Decision trees for a cluster analysis problem will be considered separately in §4.

For any observation of x, using a decision tree, we can find the predicted value Y. For this purpose we start with a root of a tree, we consider the characteristic, corresponding to a root and we define, to which branch the observed value of the given characteristic corresponds. Then we consider the node in which the given branch comes. We repeat the same operations for this node etc., until we reach a leaf. The value YS ascribed to S-th leaf will be the forecast for x. Thus, the decision tree gives the model T of dependence Y from X: Y=T(X).

Decision trees, which are considered in a regression analysis problem, are called regression trees.

In the given manual we consider the simplest kind of decision trees, described above. There are, however, more complex kinds of trees, in which each internal node corresponds to more complex statements, not one but several characteristics are given. For example, these statements can be defined by a linear combination of quantitative characteristics (for example, expression 10x1+5x2-1>0) i.e. corresponding to various subregions of multivariate space which are split by a hyper plane).

From this point of view, the hyper planes of the considered decision trees are perpendicular to the numerical axes.

The decision tree should be consistent, which means that on the way from the root to a leaf, there should be no mutually excluding variants, for example «$X_1 > 37$» и «$X_1 < 30$».

It is possible to allocate the following features of the decision trees.

Decision trees allow to process both quantitative and qualitative characteristics simultaneously.

A set of logic statements about values of characteristics corresponds to decision trees. Each statement is obtained by passing the way from root to leaf. So, for example, for the tree represented on figure 1 the following list of statements corresponds to:

1. If $X_1 < 37$, *Y="is health"*.

2. If $X_1 \in [37,38.5]$ and *X3="there is no reddening of throat"*, then *Y="to catch cold";*

3. If $X_1 \in [37,38.5]$ and *X3="there is reddening of throat"*, then *Y="angina";*

4. If $X_1 > 38.5$ and *X2="there is no cough"*, then *Y="influenza";*

5. If $X_1 > 38.5$ and *X2="there is cough"*, then *Y="pneumonia";*

Thus, the decision tree represents a logic model of regularities of the researched phenomenon.

The lack of decision trees is the fact that in a case where all characteristics are quantitative, the decision trees may represent sufficiently rough approximation of the optimum solution. For example, the regression tree, which is drawn on figure 3, is piecewise a constant approximation of the regression function. On the other hand, it is possible to compensate this lack by increasing the number of leaves, i.e. by decreasing the length of appropriate "segments" or "steps".

Let us consider a decision tree with M leaves. This decision tree corresponds to the decomposition of the characteristic space into M non-overlapping subregions E1,...,EM , so that subregion ES corresponds to S-th leaf (fig. 4). How is the given subregion formed?

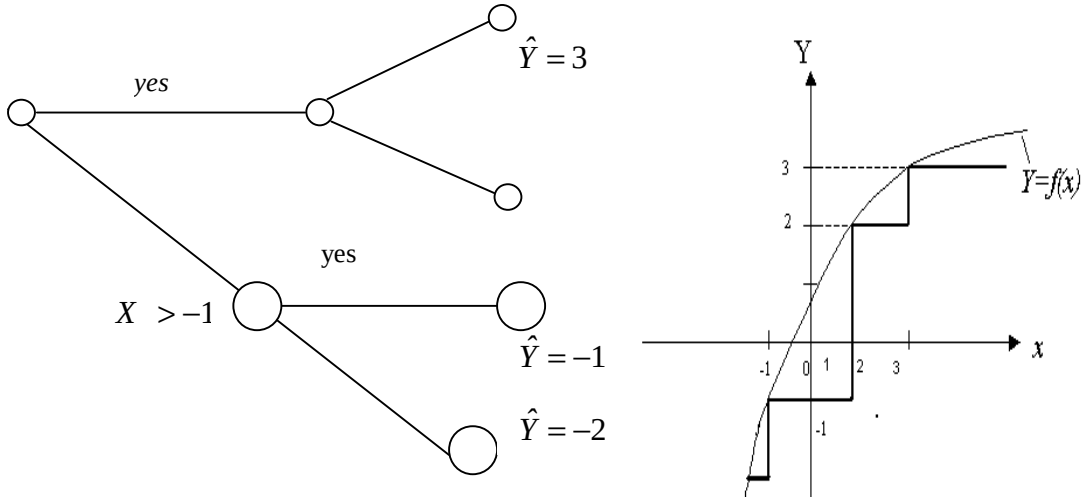

Fig. 3

E^S is defined as Cartesian product $E^S = E_1^S \times E_2^S \times ... \times E_n^S$, where E_j^S - is projection E^S on *j-th* characteristic. E_j^S is obtained at the next way. If the characteristic X_j never situated on the way from the root to *S-th* leaf, then E_j^S coincides with a range of definitions of the characteristic X_j.

Otherwise, E_j^S is equal to intersection of all subregions of the characteristic X_j, which were met on the way from the root to *S-th* leaf.

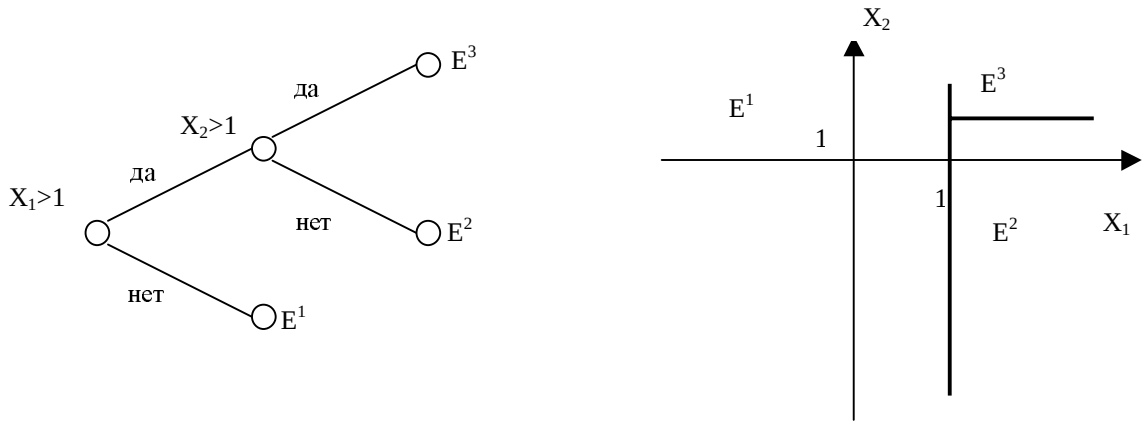

Fig. 4

Let there be some set of experimental observations *Data*=(x^i,y^i), i=1,...,N. Each of these observations belongs to (with respect to X) some of the considered subdomains, i.e. $x^i \in E^S$. We will denote the set of the observations belonging to E^S as $Data^S$, and the number of the observations, we denote as N^S. Let N_i^S denote the number of observations from $Data^S$, belonging to *i-th* class (pattern recognition problem PRP).