СПЕЦКУРС
Логический анализ данных в распознавании
(Logical data analysis in recognition)

лектор д.ф.-м.н. Елена Всеволодовна Дюкова

Спецкурс посвящён вопросам применения аппарата дискретной математики в задачах интеллектуального анализа данных. Излагаются общие принципы, лежащие в основе логического подхода к задачам машинного обучения. Описываются методы конструирования процедур классификации по прецедентам с использованием понятий теории булевых функций и теории покрытий булевых матриц. Рассматриваются основные модели логических процедур классификации, вопросы сложности их реализации и качества решения прикладных задач.

Спецкурс для бакалавров 2-4 курсов ВМК МГУ им. М.В. Ломоносова.

По спецкурсу издано учебное пособие:
http://www.ccas.ru/frc/papers/djukova03mp.pdf

Лекция 10
Алгоритмы классификации на основе решающих деревьев

- Решающие деревья – это один из наиболее популярных инструментов для решения задач классификации по прецедентам.
- Для простоты рассмотрим задачу распознавания с бинарными признаками $\boldsymbol{x_1, \dots, x_n}$, с непересекающимися классами $\boldsymbol{K_1, \dots, K_l}$ и обучающими объектами $\boldsymbol{S_1, \dots, S_m}$.
- Ориентированное бинарное дерево называется бинарным решающим деревом (БРД), если:
- Каждой внутренней вершине дерева сопоставлен один из признаков $\boldsymbol{x_1, \dots, x_n}$
- Каждой висячей вершине сопоставлен один из классов $\boldsymbol{K_1, \dots, K_l}$
- Каждая дуга помечена одним из чисел 0 или 1, причем из каждой внутренней вершины выходят две дуги, помеченные разными числами.

• Нетрудно видеть, что ветви БРД с вершинами $\boldsymbol{x_{j_1}, \dots, x_{j_r}}$ можно сопоставить ЭК над переменными $\boldsymbol{x_1, \dots, x_n}$ вида $\boldsymbol{x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}}$, где $\boldsymbol{\sigma_i}$ – метка дуги, выходящей из вершины $\boldsymbol{x_{j_i}}$, $\boldsymbol{i = 1, 2, \dots, r}$. С другой стороны, каждой висячей вершине, а следовательно, каждой ветви БРД сопоставлен один из классов $\boldsymbol{K_1, \dots, K_l}$. Таким образом, каждой ветви БРД соответствует пара $\boldsymbol{(B, K)}$, где $\boldsymbol{B}$ – ЭК над переменными $\boldsymbol{x_1, \dots, x_n}$, $\boldsymbol{K \in \{K_1, \dots, K_l\}}$.

• Пусть БРД имеет $\boldsymbol{\mu}$ ветвей и этим ветвям соответствуют пары $\boldsymbol{\left(B_1, K_{j_1}\right), \dots, \left(B_\mu, K_{j_\mu}\right)}$, $\boldsymbol{N_{B_i}}$ – интервал истинности ЭК $\boldsymbol{B_i}$, $\boldsymbol{i \in \{1, 2, \dots, \mu\}}$, $\boldsymbol{S}$ – распознаваемый объект. Если среди конъюнкций $\boldsymbol{B_i, \dots, B_\mu}$ существует конъюнкция $\boldsymbol{B_i}$ такая, что $\boldsymbol{S \in N_{B_i}}$, то объект **S** относится к классу $\boldsymbol{K_{j_i}}$. В противном случае происходит отказ от распознавания объекта **S**.

• Пусть некоторой ветви дерева сопоставлена пара $(\boldsymbol{B}, \boldsymbol{K})$, где $\boldsymbol{B} = \boldsymbol{x}_{\boldsymbol{j_1}}^{\boldsymbol{\sigma_1}} \dots \boldsymbol{x}_{\boldsymbol{j_r}}^{\boldsymbol{\sigma_r}}$. Рассматриваемая ветвь называется корректной, если наборы из $\{\boldsymbol{S_1}, \dots, \boldsymbol{S_m}\} \cap \boldsymbol{N_B}$ являются описаниями объектов класса $\boldsymbol{K}$. В этом случае интервал истинности $\boldsymbol{N_B}$ конъюнкции $\boldsymbol{B}$ называется допустимым. Бинарное решающее дерево, у которого каждая ветвь корректная, называется корректным. Нетрудно видеть, что в корректном дереве ветвь $(\boldsymbol{B}, \boldsymbol{K})$ с конъюнкцией $\boldsymbol{B} = \boldsymbol{x}_{\boldsymbol{j_1}}^{\boldsymbol{\sigma_1}} \dots \boldsymbol{x}_{\boldsymbol{j_r}}^{\boldsymbol{\sigma_r}}$ порождает представительный набор класса $\boldsymbol{K}$ вида $(\boldsymbol{\sigma}, \boldsymbol{H}), \boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \dots, \boldsymbol{\sigma_r}), \boldsymbol{H} = (\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_r}})$.

• Синтез решающего дерева представляет собой итерационный процесс. Как правило, для построения очередной вершины дерева выбирается признак, наилучшим образом удовлетворяющий некоторому критерию ветвления. По значениям этого признака и осуществляется ветвление, далее указанная процедура повторяется для каждой из ветвей. Описанный подход обуславливает основные достоинства метода, а именно, решающее правило строится быстро, получается достаточно простым и хорошо интерпретируемым.

- Одним из наиболее известных алгоритмов классификации на основе БРД является **алгоритм построения допустимого разбиения (далее АДР).**
- Критерий ветвления в АДР представляет собой набор условий с разным приоритетом. При выборе очередной вершины сначала проверяется условие с наибольшим приоритетом. Ищется признак с наименьшим номером, для которого это условие выполняется. Если ни один признак не удовлетворяет рассматриваемому условию, то проверяется следующее по порядку условие с более низким приоритетом. Однако, в случае, когда два или более признака удовлетворяют рассматриваемому критерию ветвления в равной или почти равной мере, выбор одного из этих признаков осуществляется практически случайным образом. При этом в зависимости от выбранного признака построенные деревья могут существенно отличаться как по составу используемых признаков, так и по своим распознающим качествам.

- Алгоритм АДР строит разбиение множества $\boldsymbol{E^n}$ всех двоичных наборов длины $\boldsymbol{n}$ на непересекающиеся интервалы, такие, что в каждом из них находятся описания обучающих объектов только из одного класса.
- На первом шаге множество $\boldsymbol{E^n}$ разбивается на два непересекающихся интервала $\boldsymbol{N_1^1}$ и $\boldsymbol{N_2^2}$ ранга $\boldsymbol{n-1}$ таких, что $\boldsymbol{N_1^1 \cup N_2^2 = E^n}$ и можно указать пару наборов из $\boldsymbol{E^n}$, принадлежащих разным интервалам и описывающих обучающие объекты из разных классов. Интервалы $\boldsymbol{N_1^1}$ и $\boldsymbol{N_2^2}$ являются интервалами истинности конъюнкций $\boldsymbol{x_{j_1}}$ и $\boldsymbol{\bar{x}_{j_1}}$ ранга 1.
- Пусть на шаге $\boldsymbol{i}$ построено разбиение $\boldsymbol{E^n}$ на некоторое множество интервалов. Если в этом разбиении не существует интервала, содержащего наборы, описывающие объекты из разных классов, то алгоритм заканчивает свою работу. В противном случае в построенном разбиении выбирается интервал, содержащий наборы, описывающие объекты из разных классов. Пусть $\boldsymbol{n - r}$ – ранг выбранного интервала и пусть ему соответствует конъюнкция $\boldsymbol{x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}}$. Тогда этот интервал разбивается на два непересекающихся интервала $\boldsymbol{N_1^{i+1}}$ и $\boldsymbol{N_2^{i+1}}$ ранга $\boldsymbol{n-r-1}$

• Пусть $\boldsymbol{n - r}$ – ранг выбранного интервала и пусть ему соответствует конъюнкция $\boldsymbol{x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}}$. Тогда этот интервал разбивается на два непересекающихся интервала $\boldsymbol{N_1^{i+1}}$ и $\boldsymbol{N_2^{i+1}}$ ранга $\boldsymbol{n - r - 1}$, таких, что можно указать пару наборов $\tilde{\boldsymbol{\alpha}}$ и $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\alpha}} \in \boldsymbol{N_1^{i+1}}$, $\tilde{\boldsymbol{\beta}} \in \boldsymbol{N_2^{i+1}}$, которые описывают объекты из разных классов. Интервалами $\boldsymbol{N_1^{i+1}}$ и $\boldsymbol{N_2^{i+1}}$ соответствуют конъюнкции вида $\boldsymbol{x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r} \overline{x}_{j_{r+1}}}$ и $\boldsymbol{x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r} x_{j_{r+1}}}$.

• Нетрудно видеть, что в процессе работы описанного выше алгоритма строится множество допустимых интервалов $\boldsymbol{N_1, \dots, N_\mu}$ таких, что $\bigcup_{q=1}^{\mu} \boldsymbol{N_q} = \boldsymbol{E^n}$ и $\bigcap_{q=1}^{\mu} \boldsymbol{N_q} = \emptyset$. Множество интервалов $\boldsymbol{N_1, \dots, N_\mu}$ называется *допустимым разбиением*.

• Статистические оценки надежности распознавания БРД на контрольной выборке говорят о том, что наиболее надежным является решающее дерево с наименьшим числом листьев, а среди БРД, имеющих наименьшее число листьев, наиболее надежными являются те, у которых объекты обучающей выборки наиболее равномерно распределены по интервалам соответствующего разбиения. Известно, что задача построения корректного БРД с минимальным числом листьев является NP-полной. Поэтому при выборе признака для ветвления используются эвристические соображения, называемые критерием выбора признака. Например, при выборе признака последовательно проверяются условия полной, частичной и максимальной отделимости, приведенные ниже.

• Пусть для разбиения интервала $\boldsymbol{N}$ на пару интервалов $\boldsymbol{N_0}$ и $\boldsymbol{N_1}$ выбирается признак $\boldsymbol{x_j}$. Имеем $\boldsymbol{N = N_0 \cup N_1,\ N_0 \cap N_1 = \emptyset}$. Обозначим через $\boldsymbol{A_0(j)}$ и $\boldsymbol{A_1(j)}$ соответственно множество $\boldsymbol{\{S_1, \dots, S_m\} \cap N_0}$ и $\boldsymbol{\{S_1, \dots, S_m\} \cap N_1}$ (здесь и далее $\boldsymbol{\{S_1, \dots, S_m\}}$ - обучающая выборка).

• Будем говорить, что признак $\boldsymbol{x_j}$ удовлетворяет *условию полной отделимости*, если каждое из множеств $\boldsymbol{A_0(j)}$ и $\boldsymbol{A_1(j)}$ содержит описания объектов только одного класса. Будем говорить, что признак $\boldsymbol{x_j}$ удовлетворяет *условию частичной отделимости*, если одно из множеств $\boldsymbol{A_0(j)}$ или $\boldsymbol{A_1(j)}$ содержит описания объектов только одного класса, а другое этим свойством не обладает. Пусть $\boldsymbol{D(u)}$, $\boldsymbol{u \in 1, 2, \dots, n}$, – число пар объектов вида $\boldsymbol{(S_{i_1}, S_{i_2})}$ таких, что $\boldsymbol{S_{i_1}}$, $\boldsymbol{S_{i_2}}$ - обучающие объекты из разных классов и $\boldsymbol{S_{i_1} \in A_0(u), S_{i_2} \in A_1(u)}$. Будем говорить, что признак $\boldsymbol{x_j}$ удовлетворяет *условию максимальной отделимости*, если $\boldsymbol{D(j) = max_{u \in \{1,\dots,n\}} D(u)}$.

• Рассмотрим вариант алгоритма построения допустимого разбиения с использованием перечисленных выше условий. Данный алгоритм является рекурсивным. Положим на первом шаге $\boldsymbol{T^* = T_{mn}}$, где $\boldsymbol{T_{mn}}$ – таблица из $\boldsymbol{m}$ строк и $\boldsymbol{n}$ столбцов, в которой строка с номером $\boldsymbol{i}$, $\boldsymbol{i \in \{1, \dots, m\}}$, представляет собой описание обучающего объекта $\boldsymbol{S_i}$ (таблица обучения). Далее на каждом шаге рекурсии выполняется следующая последовательность действий.

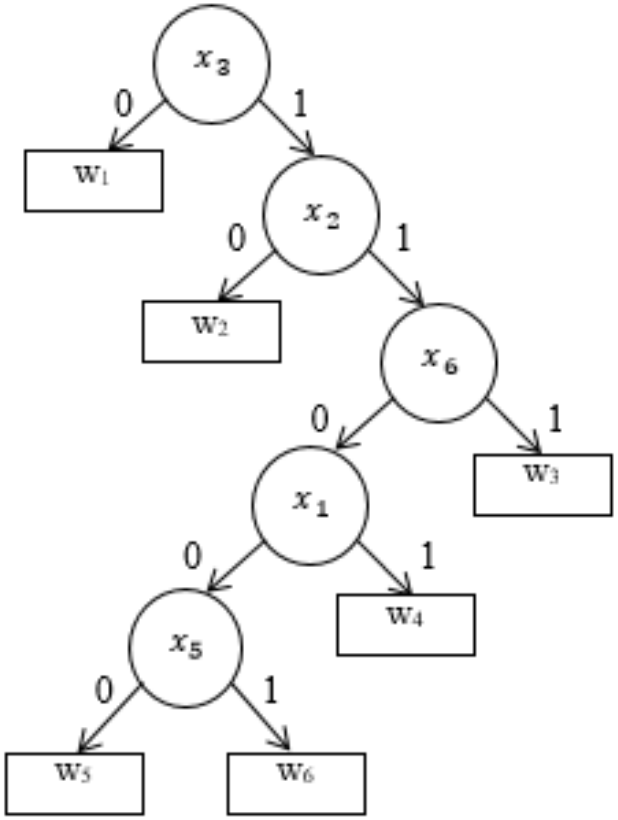
- Если все строки в $\boldsymbol{T}^*$ описывают объекты одного класса, то создается висячая вершина с меткой этого класса и рекурсия останавливается. В противном случае вызывается процедура выбора признака по таблице $\boldsymbol{T}^*$.

- Пусть в результате выбран признак $\boldsymbol{x_j}$. Создается внутренняя вершина с меткой $\boldsymbol{x_j}$ и для каждого значения $\boldsymbol{a} \in \{\boldsymbol{0}, \boldsymbol{1}\}$ признака $\boldsymbol{x_j}$ делается следующее. Создается дуга с меткой $\boldsymbol{a}$, исходящая из построенной вершины. Строится подматрица $\boldsymbol{T}^*(\boldsymbol{j}, \boldsymbol{a})$ матрицы $\boldsymbol{T}^*$ путём удаления столбца, соответствующего признаку $\boldsymbol{x_j}$, и удаления строк, которые в пересечении с этим столбцом дают $\boldsymbol{a}$. Далее полагается $\boldsymbol{T}^* = \boldsymbol{T}^*(\boldsymbol{j}, \boldsymbol{a})$ и производится следующий шаг рекурсии.

- Процедура выбора признака по таблице $\boldsymbol{T}^*$ представляет собой следующую последовательность шагов.

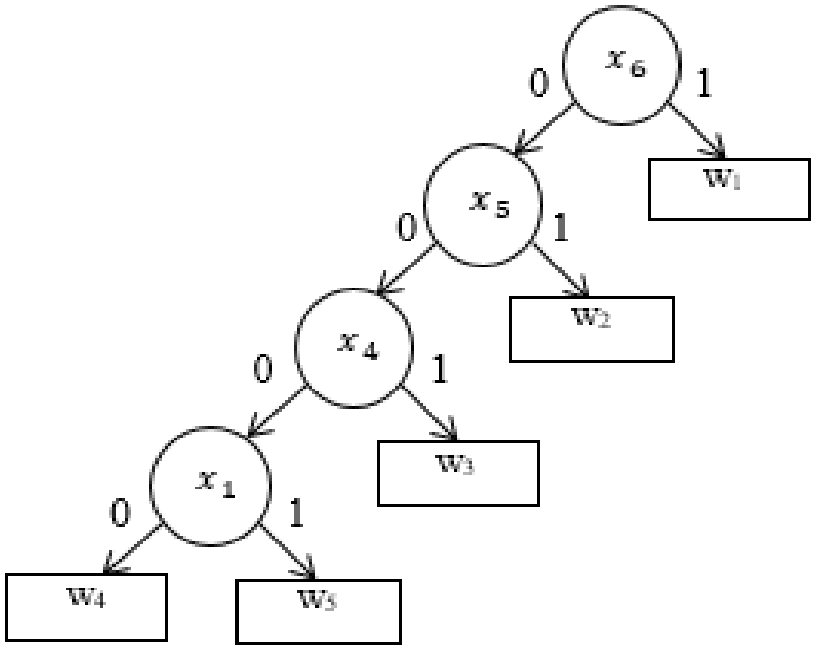
- Шаг 1. Для каждого признака проверяется условие полной отделимости. Если находятся признаки, для которых оно выполняется, то среди них выбирается первый по порядку признак и процедура останавливается.
- Шаг 2. Для каждого признака проверяется условие частичной отделимости. Если находятся признаки, для которых оно выполняется, то среди них выбирается первый по порядку признак и процедура останавливается.
- Шаг 3. Выбирается первый по порядку признак, для которого выполняется условие максимальной отделимости.

- Рассмотрим пример. Пусть обучающая выборка состоит из шести объектов и имеет вид:
- **(1, 0, 0, 0, 1, 0),**
- **(0, 1, 1, 1, 1, 0),**
- **(1, 1, 1, 0, 0, 0),**
- **(1, 0, 1, 1, 0, 0),**
- **(1, 1, 1, 0, 1, 1),**
- **(0, 1, 1, 0, 0, 0).**
- Пусть первые три объекта принадлежат классу первому классу, а последующие три объекта принадлежат второму классу.
- Ниже на рисунках 1 и 2 приведены различные БРД, построенные с помощью алгоритма АДР.



$$w_1 = (\bar{x}_3, K_1), w_2 = (\bar{x}_2 x_3, K_2), w_3 = (x_2 x_3 x_6, K_2), w_4 = (x_1 x_2 x_3 x_6, K_1),$$
$$w_5 = (\bar{x}_1 x_2 x_3 \bar{x}_5 x_6, K_2), w_6 = (\bar{x}_1 x_2 x_3 x_5 x_6, K_1)$$

Рис. 1



$$w_1 = (x_6, K_2), w_2 = (x_5\bar{x}_6, K_1), w_3 = (x_4\bar{x}_5\bar{x}_6, K_2), w_4 = (\bar{x}_1\bar{x}_4\bar{x}_5\bar{x}_6, K_2),$$
$$w_5 = (x_1\bar{x}_4\bar{x}_5\bar{x}_6, K_1)$$

Рис. 2

- Как видно из приведенного выше описания алгоритма АДР, если проверяемому условию удовлетворяет более одного признака, то выбор одного из этих признаков происходит фактически случайно. Чтобы исправить указанный недостаток предлагается несколько модифицировать решающее дерево. Для этого введем второй тип внутренних вершин, которым соответствует набор равноценных признаков, т.е. признаков, удовлетворяющих критерию ветвления в равной или почти равной мере.
- Пусть внутренней вершине второго типа $\boldsymbol{v}$, называемой далее полной вершиной, соответствует набор признаков $\left\{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_q}}\right\}, \boldsymbol{1 \le q \le n}$, тогда из этой вершины исходит $\boldsymbol{q}$ дуг $\boldsymbol{t_1}, \dots, \boldsymbol{t_q}$, каждая из которых связывает вершину $\boldsymbol{v}$ с вершиной первого типа. Причем, дуга $\boldsymbol{t_u}$, $\boldsymbol{u = 1, 2, \dots, q}$, входит в вершину первого типа с меткой $\boldsymbol{x_{j_u}}$. Полученную конструкцию будем называть полным решающим деревом.

• Заметим, что в полном решающем дереве, также как и в обычном, каждой висячей вершине соответствует некоторая конъюнкция. Однако в полном решающем дереве описание распознаваемого объекта может попасть в интервалы истинности конъюнкций, соответствующих разным ветвям дерева. В случае, если таким ветвям дерева соответствуют разные классы, выбор класса осуществляется простым голосование, т.е. объект зачисляется в тот класс, который соответствует большинству из указанных ветвей.

• Алгоритм построения полного решающего дерева (ПРД) также является рекурсивным. Обозначим через $\boldsymbol{T}^*$ матрицу, рассматриваемую на текущем шаге алгоритма, на первом шаге $\boldsymbol{T}^* = \boldsymbol{T_{mn}}$. Шаг рекурсии представляет собой следующую последовательность действий.

- Если все строки в таблице $\boldsymbol{T}^*$ описывают объекты одного класса, то создается висячая вершина с меткой этого класса и рекурсия останавливается. В противном случае вызывается процедура выбора в таблице $\boldsymbol{T}^*$ набора равноценных признаков для ветвления. Пусть в результате выбран набор признаков $\left\{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_q}}\right\}$. Тогда создается полная вершина с меткой $\left\{\boldsymbol{x_{j_1}}, \dots, \boldsymbol{x_{j_q}}\right\}$ и для каждого признака $\boldsymbol{x_{j_u}}, \boldsymbol{u} \in \{\boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{q}\}$, строится вершина с меткой $\boldsymbol{x_{j_u}}$. Для каждого значения $\boldsymbol{a} \in \{\boldsymbol{0}, \boldsymbol{1}\}$ признака $\boldsymbol{x_{j_u}}$ создается дуга, исходящая из вершины $\boldsymbol{x_{j_u}}$. Строится подматрица $\boldsymbol{T}^*(\boldsymbol{j_u}, \boldsymbol{a})$, которая образуется удалением столбца матрицы $\boldsymbol{T}^*$, соответствующего признаку с номером $\boldsymbol{j_u}$, и строк, в которых $\boldsymbol{a_{ij_u}} \neq \boldsymbol{a}$, $\boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \dots, \boldsymbol{m}$. Далее для рассматриваемой ветви полагается $\boldsymbol{T}^* = \boldsymbol{T}^*(\boldsymbol{j_u}, \boldsymbol{a})$ и производится следующий шаг рекурсии.
- Процедура выбора по таблице $\boldsymbol{T}^*$ набора признаков для ветвления работает аналогично процедуре выбора признака для ветвления в классическом БДР.

- Описанный алгоритм также как АДР легко обобщается на случай, когда признаки целочисленные. Сравнение этого алгоритма с АДР, а также с широко используемым алгоритмом построения решающего дерева, известным под названием C4.5, показало целесообразность использования идеи полного решающего дерева. Результаты тестирования алгоритмов описаны в работе:
- *Djukova E.V., Peskov N.V. A Classification Algorithm Based on the Complete Decision Tree // J. Pattern Recognition and Image Analysis, 2007. Vol. 17. No. 3, pp. 363–367.*

- Затем была поставлена задача модифицировать на основе ПРД алгоритм C4.5 (*Генрихов И.Е., Дюкова Е.В. Классификация на основе полных решающих деревьев. Ж. вычисл. матем. и матем. физ. 2012. Т. 52, № 4. С. 750-761*). Впервые были построены и исследованы классификаторы на основе ПРД с энтропийным критерием ветвления, использующие взвешенное голосование по голосующим ветвям дерева (вместо голосования по большинству). С их помощью были эффективно решены вопросы обработки вещественнозначной информации. Построенные классификаторы существенно отличаются от алгоритма C4.5.
- Теоретическое и экспериментальное исследование обобщающей способности классификаторов на основе ПРД, т.е. способности этих классификаторов правильно классифицировать объекты, не вошедшие в обучающую выборку, свидетельствует о том, что обобщающая способность ПРД выше обобщающей способности классического РД и не уступает другим современным конструкциям РД, среди которых Random Forest.

УПРАЖНЕНИЯ

- 1. Классифицировать объекты (1, 1, 1, 1, 1, 1) и (0, 0, 0, 0, 0, 0), используя решающие деревья на рисунках 1 и 2.
- 2. Для рассмотренного в лекции примера построить полное решающее дерево.