

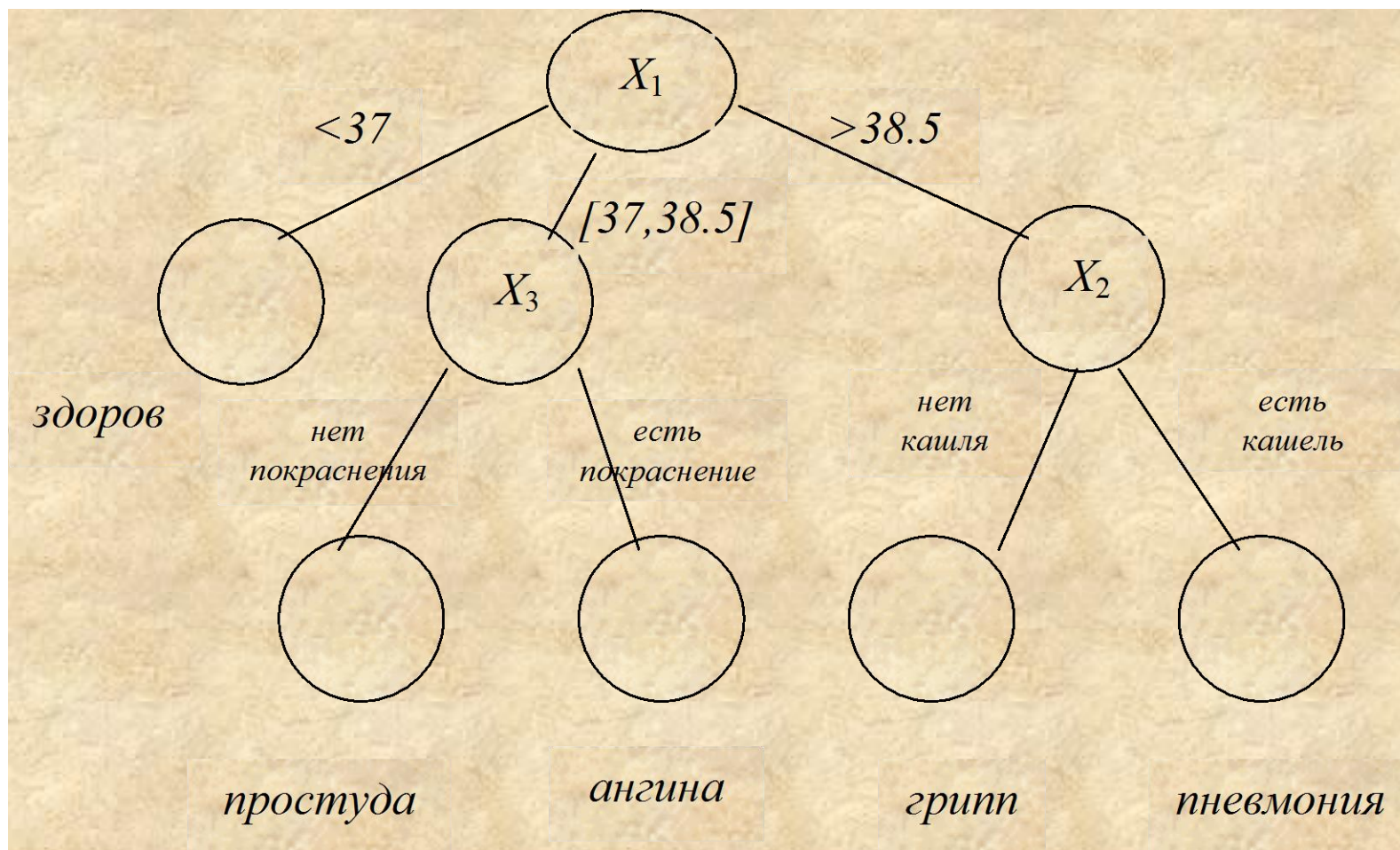
Деревья решений.

Лекция 5.

Задачи, решаемые деревьями решений

- задачи прогнозирования (задачи регрессионного анализа и анализа временных рядов);
- задачи кластеризации (задача кластерного анализа);
- задачи классификации (распознавания образов с учителем, задача дискриминантного анализа)

Что такое дерево решений?



Что такое дерево решений?

- Дерево состоит из корня, ветвей, узлов (мест разветвления), листьев.
- Самая первая (левая или верхняя) вершина называется корнем.
- Цепочка "корень - ветвь - вершина - ... - вершина" заканчивается вершиной, которую называют "листом".
- Из каждой внутренней вершины (т.е. не листа) может выходить две или более ветвей.
- Каждому такому узлу сопоставлена некоторая характеристика, а ветвям - области значения этой характеристики, причем эти области дают разбиение множества значений данной характеристики.

Что такое дерево решений?

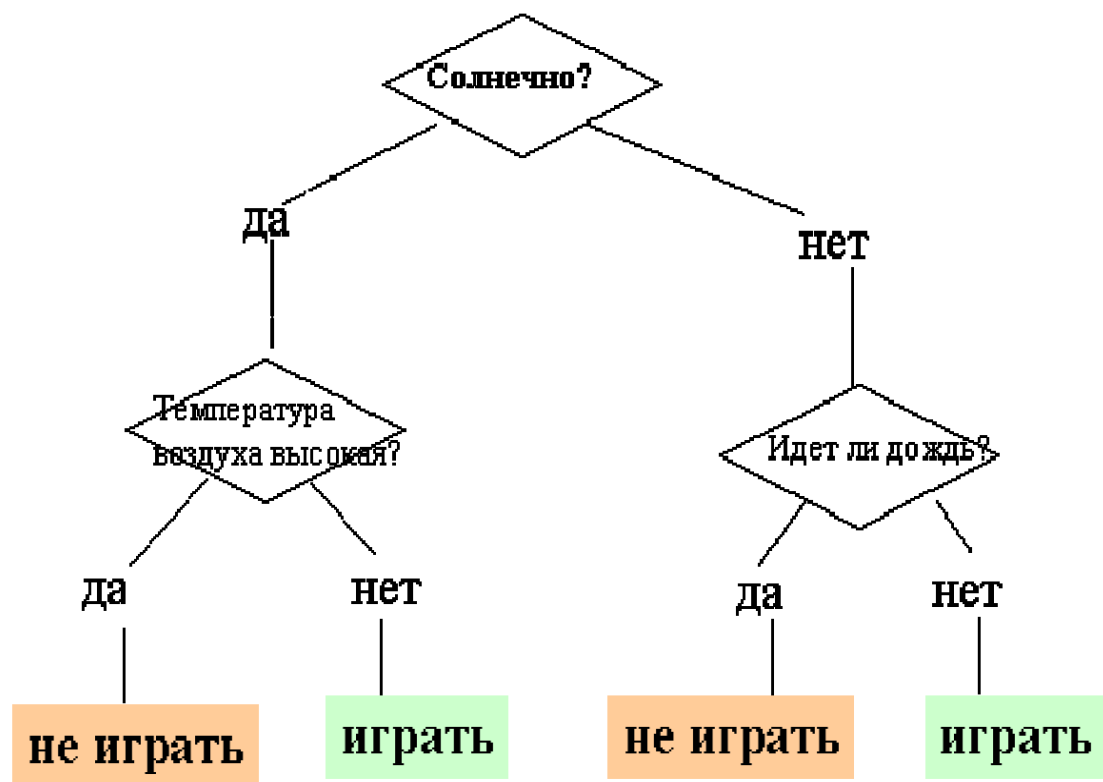


Рис. 1. Дерево решений "Играть ли в гольф?"

Что такое дерево решений?

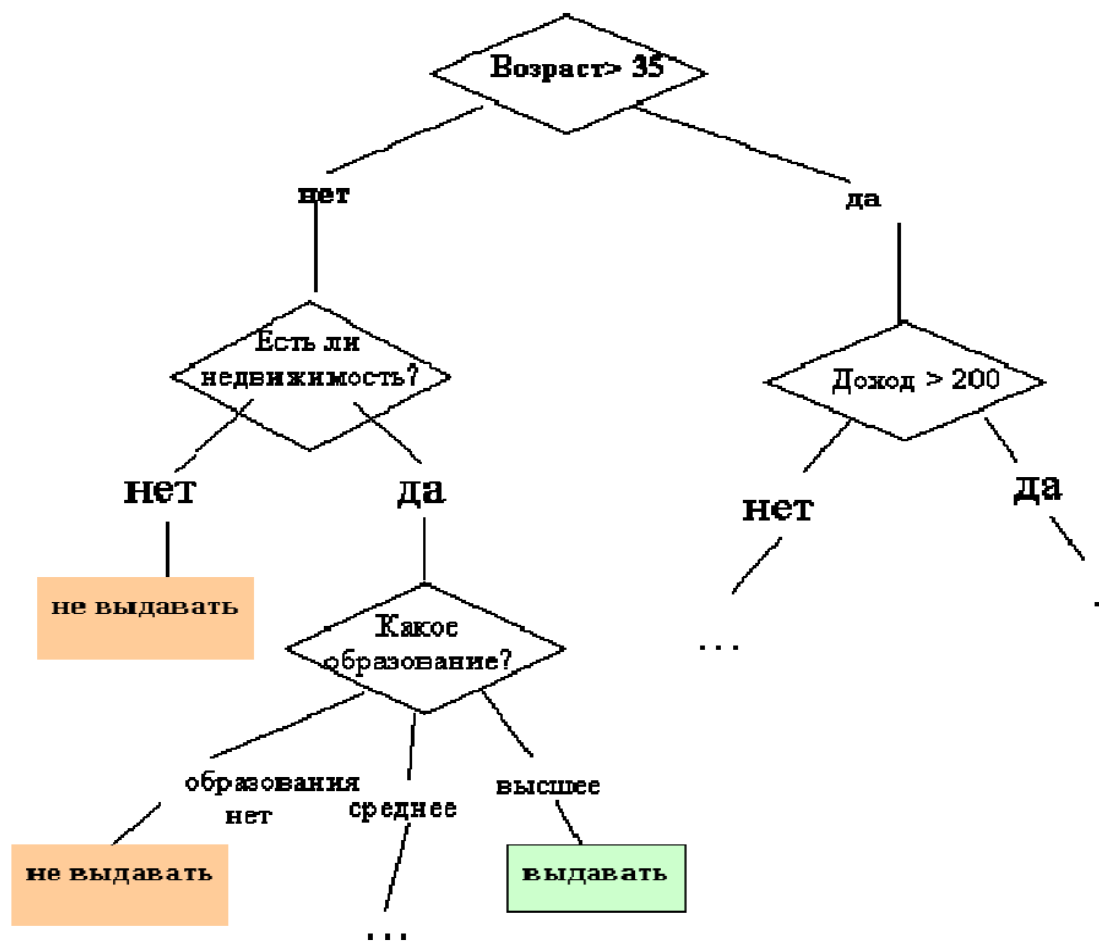
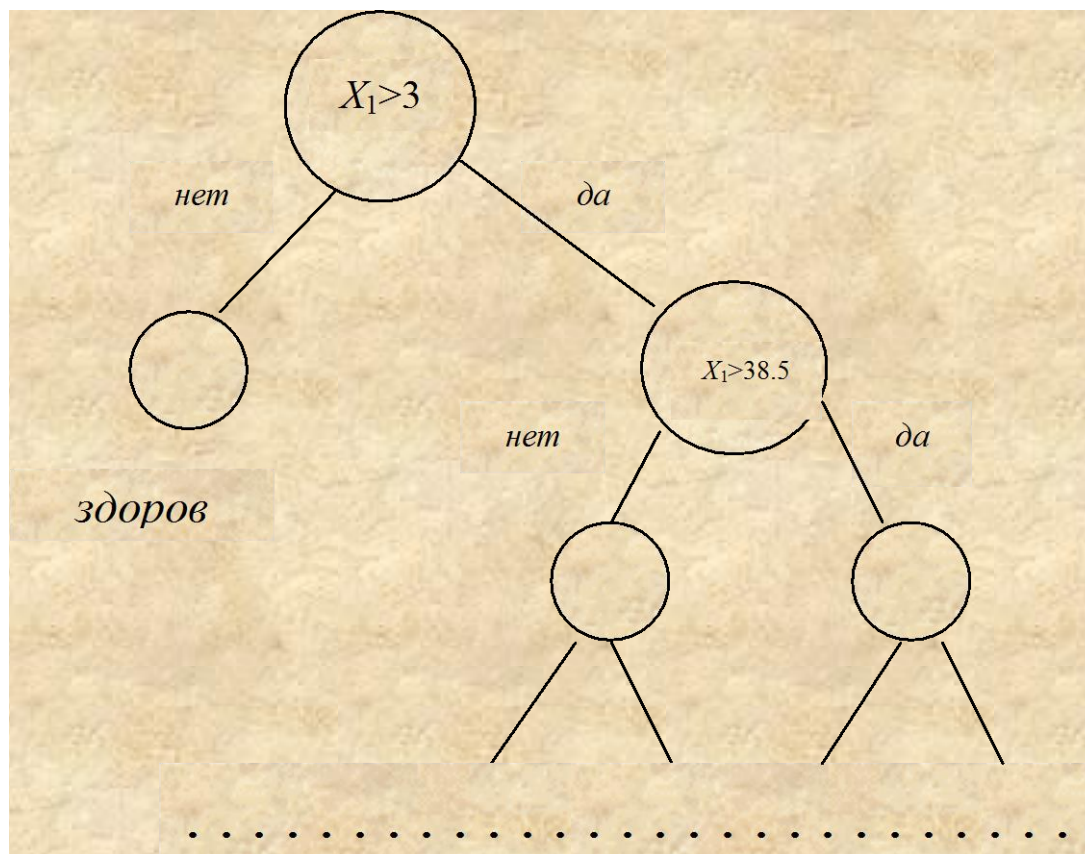


Рис. 2. Дерево решений "Выдавать ли кредит?"

Что такое дерево решений?

- Пример дихотомического дерева



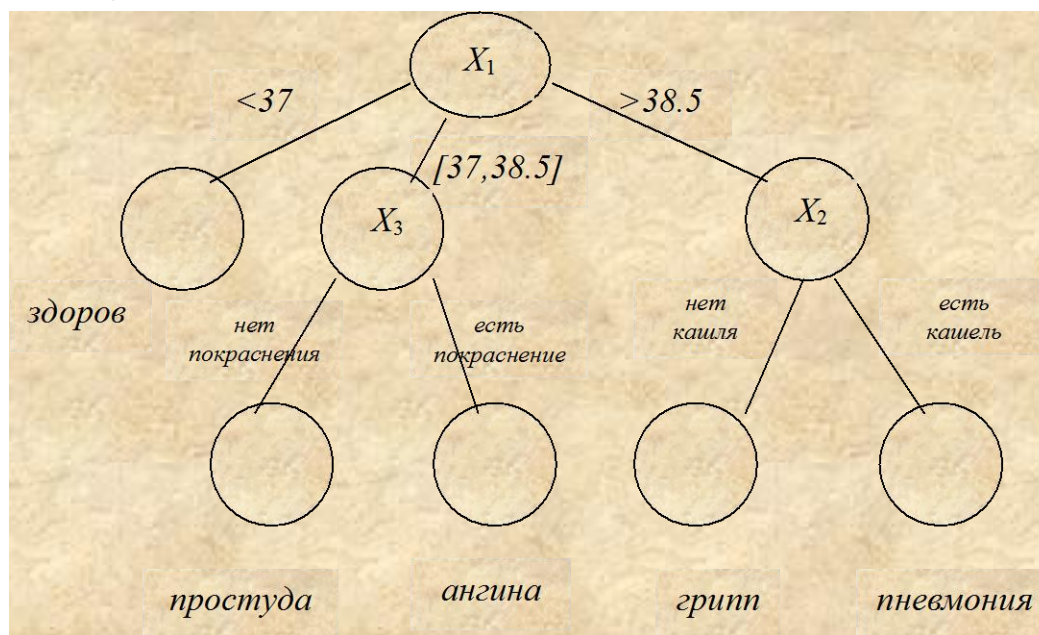
Что такое дерево решений?

Можно выделить следующие особенности деревьев решений.

- Деревья решений позволяют обрабатывать одновременно как количественные, так и качественные характеристики.
- Дереву решений соответствует набор логических высказываний о значениях характеристик. Каждое высказывание получается при прохождении пути из корневой вершины в лист.
- Дерево решений представляет собой логическую модель закономерностей исследуемого явления.

Что такое дерево решений?

1. Если $X_1 < 37$, то $Y = \text{«здоров»}$.
2. Если $X_1 \in [37, 38.5]$ и $X_3 = \text{«нет покраснения горла»}$, то $Y = \text{«простуда»}$...



Что такое дерево решений?

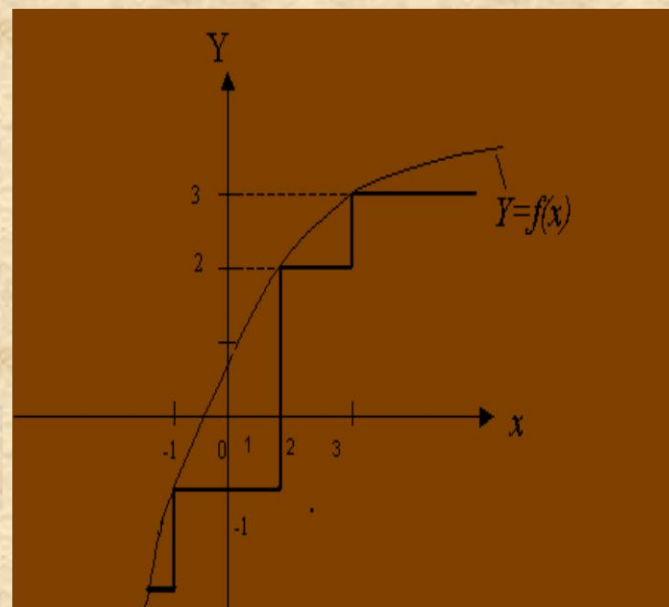
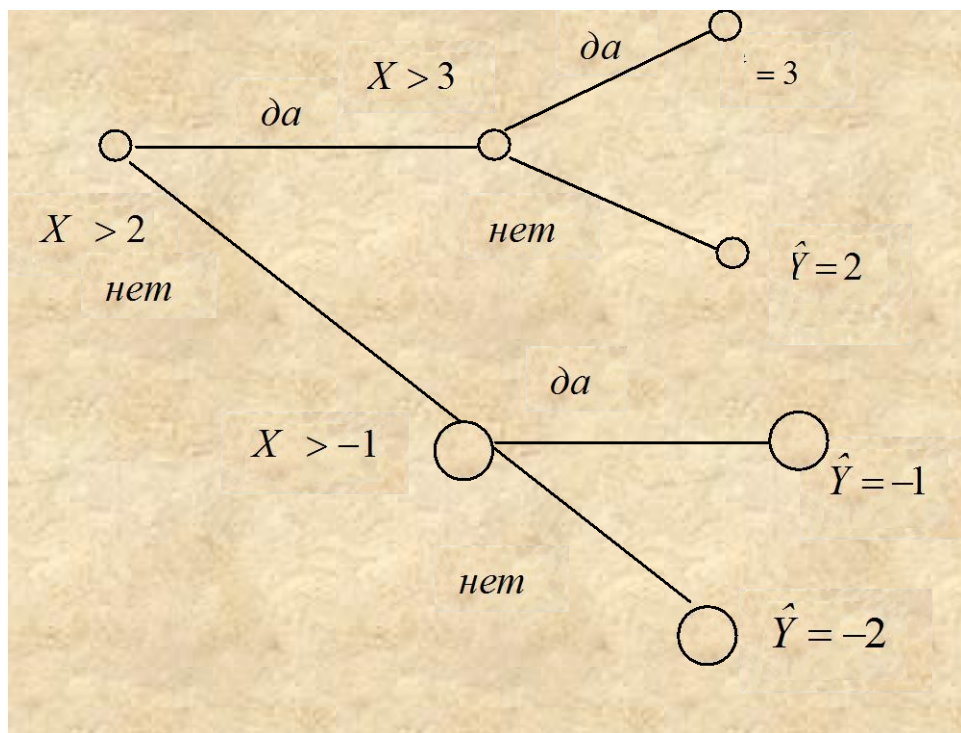
Существуют, однако, более сложные виды деревьев, каждой внутренней вершине которых соответствуют более сложные высказывания о значениях не одной, а нескольких характеристик.

Например, эти высказывания могут определяться линейной комбинацией количественных характеристик:

например, выражением $10x_1 + 5x_2 - 1 > 0$.

Что такое дерево решений?

пример дерева регрессии

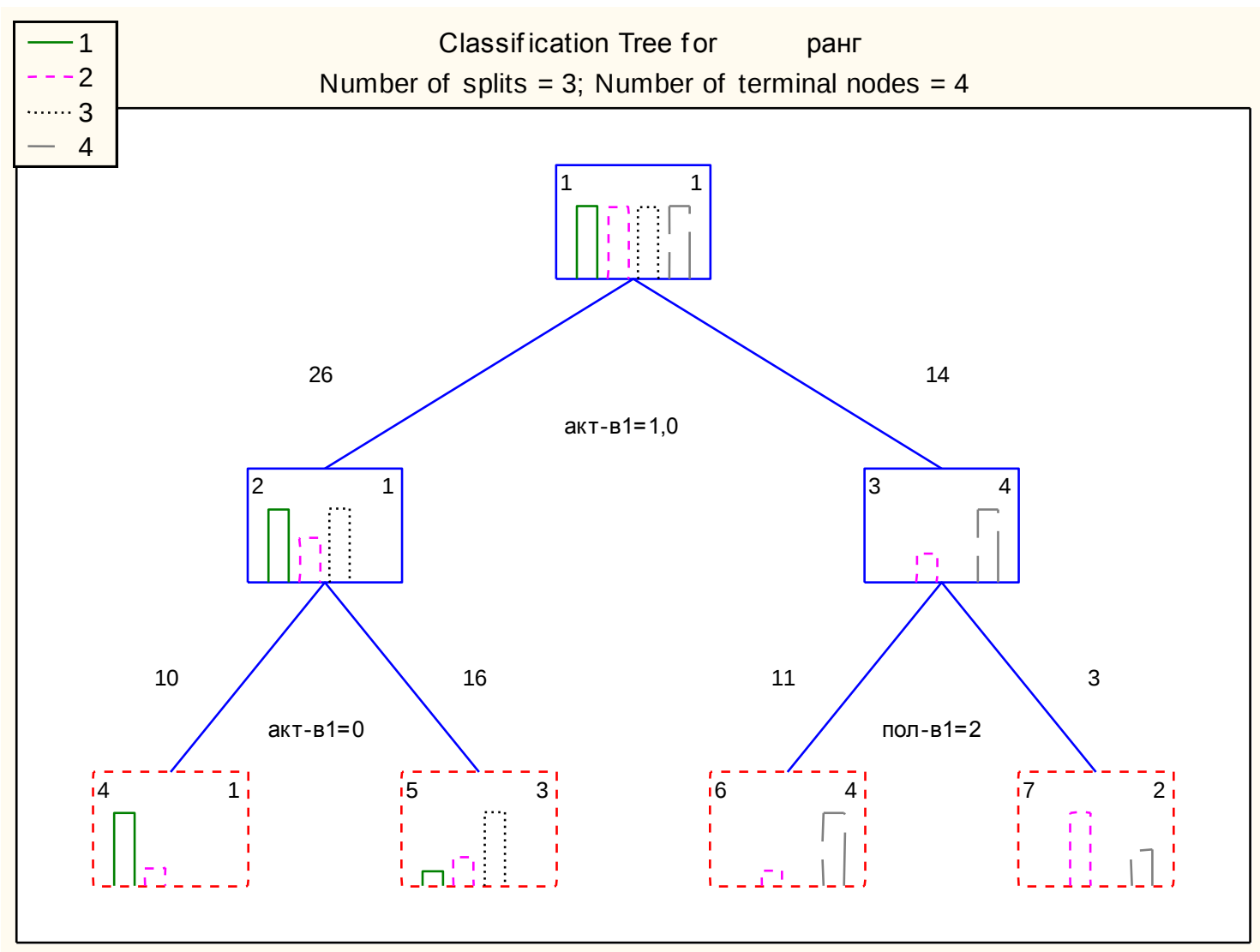


Что такое дерево решений?

Исходные данные:

- пользователи Интернет – магазина (объем выборки - 40);
- разбиты на 4 класса по показателям: возраст, пол, частота заказа литературы, активность (степень участия в Интернет-опросах).

Что такое дерево решений?



Показатели качества дерева

- показатели точности

Доля объектов, неверно отнесенных деревом к чужому классу.

$$\hat{P}_{err} = \frac{N_{err}}{N}$$

$$N_{err} = \sum_{S=1}^M \sum_{\substack{i=1 \\ i \neq \hat{Y}(S)}}^K N_i^S,$$

где K – число классов, M – число листьев

Показатели качества

- показатели сложности
 - число листьев дерева;
 - число внутренних вершин;
 - максимальная дина пути от корня к листу;
 - длина внешнего пути – число ветвей, дополняющих данное дерево до полного дерева

Показатели качества

- Показатели сложности и точности взаимосвязаны: чем сложнее дерево, тем оно, как правило, точнее.
- Менее сложное дерево, при прочих равных условиях, более предпочтительно. Это объясняется стремлением получить более простую модель исследуемого явления, облегчить последующую интерпретацию («объяснение» модели).
- Кроме того, из теоретических исследований следует, что в случае небольшого (в сравнении с числом характеристик) объема выборки слишком сложное дерево становится неустойчивым, т.е. дает более высокую погрешность для новых наблюдений.

Методы построения деревьев решений

- методы построения строго оптимального дерева по заданному критерию качества (задача дискретного программирования и методы: полный перебор, метод ветвей и границ, метод динамического программирования);
- методы построения приближенно оптимального дерева.

Этапы построения деревьев решений

1. Выбор критерия точности прогноза
2. Выбор типа ветвления
3. Определение момента прекращения ветвлений
4. Определение оптимальных размеров дерева
5. Оценка прогнозной способности дерева

1. Выбор критерия точности прогноза

- **Критерии качества**

р - показатель точности

М – число листьев

a – некоторый параметр

$$Q = p + \alpha M$$

- **Минимум потерь (в Statistica)**

Показатели:

1. цена ошибки классификации
2. априорная вероятность попадания объекта в класс
3. веса наблюдений

2. Выбор типа ветвления

- **Дискриминантное одномерное ветвление**
- **Дискриминантное многомерное ветвление по линейным комбинациям**
- **Полный перебор деревьев с одномерным ветвлением по методу CART**

2. Выбор типа ветвления

- **Дискриминантное одномерное ветвление**

- для ветвления исходной терминальной вершины на две используют алгоритм кластеризации 2-х средних в случае количественных предикторов.
- в случае качественных предикторов создаются фиктивные переменные, соответствующие уровням этого предиктора, и применяется алгоритм кластеризации 2-х средних.

2. Выбор типа ветвления

Полный перебор деревьев с одномерным ветвлением по методу CART

- Используется для качественных и порядковых предикторных переменных
- Перебираются все возможные варианты ветвления по каждой предикторной переменной, и находится тот из них, который дает наибольший рост для критерия согласия:
 1. Мера Джини однородности вершины
 2. мера Хи-квадрат Бартлетта
 3. мера G-квадрат measure

2. Выбор типа ветвления

Мера Джини (Гини, Gini)

Сумма всех попарных произведений относительных размеров классов, представленных в данной вершине.

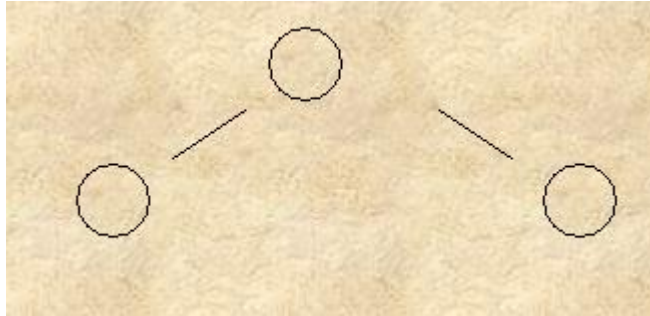
- **Пример вычисления меры Джини:**

В вершине 10 (1 класс); 20 (2 класс); 30 (3 класс); всего 60 объектов.

$$P = 10/60 * 20/60 + 10/60 * 30/60 + 20/60 * 30/60 = 11/36.$$

Если размеры всех классов одинаковы (по 20 значений), то получается 12/36

2. Выбор типа ветвления



Число ошибок для этих вариантов совпадает, однако ясно, что первый вариант более предпочтителен, т.к. в нем одна из новых вершин уже не нуждается в дальнейшем ветвлении

Вариант 1

$$\begin{aligned} N_1^1 &= 0, & N_2^1 &= 20, \\ N_1^2 &= 20, & N_2^2 &= 10, \\ N_{err} &= 10 \end{aligned}$$

Мера Джини = 0,222 – вариант лучше

Вариант 2

$$\begin{aligned} N_1^1 &= 5, & N_2^1 &= 25, \\ N_1^2 &= 15, & N_2^2 &= 5, \\ N_{err} &= 10 \end{aligned}$$

Мера Джини = 0,327

Оценка прогнозной способности дерева

- Кросс-проверка на тестовой выборке
- V-кратная кросс-проверка
- Глобальная кросс-проверка

Оценка прогнозной способности дерева

- Кросс-проверка на тестовой выборке

Дерево строится по обучающей выборке, а его способность к прогнозированию проверяется путем предсказания классовой принадлежности элементов тестовой выборки. Если значение цены на тестовой выборке окажется больше, чем на обучающей выборке, то необходимо строить дерево большего размера

Оценка прогнозной способности дерева

• **V-кратная кросс-проверка**

Задаваемое пользователем значение V определяет число случайных подвыборок, которые формируются из обучающей выборки и становятся тестовыми.

Дерево классификации нужного размера строится V раз.

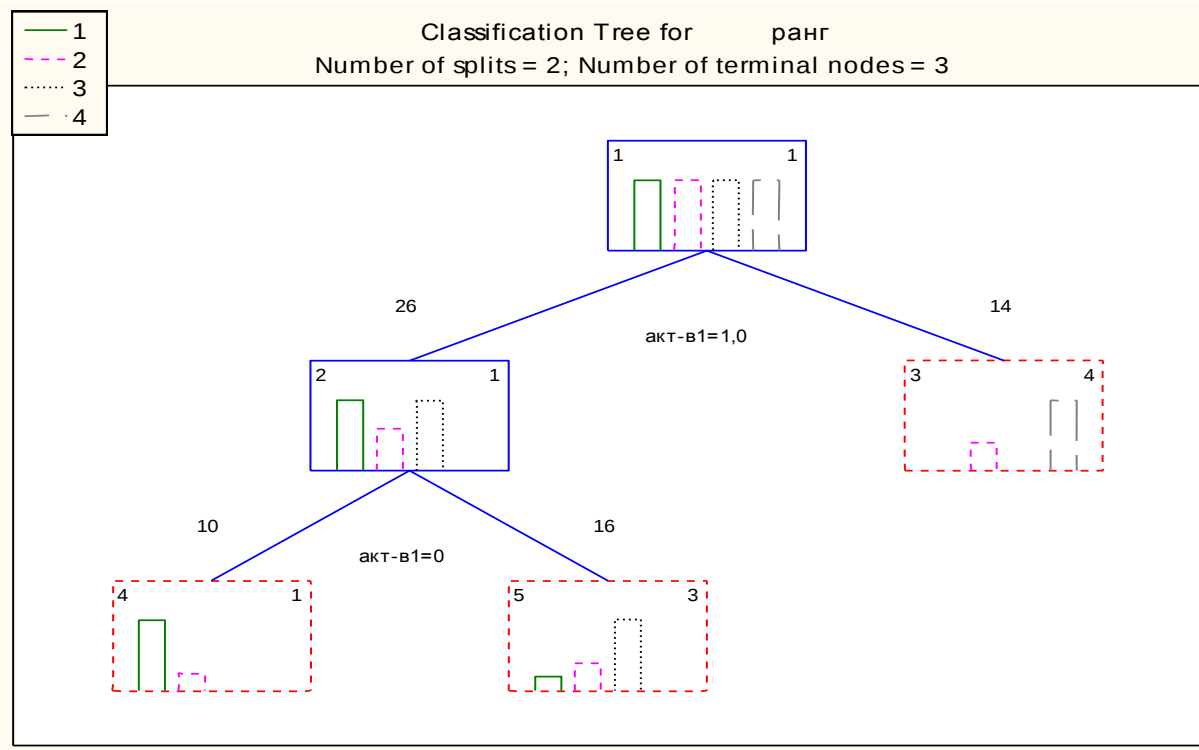
4. Определение оптимального размера дерева

- **Прямая остановка по методу FAST**
- **Кросс-проверочное отсечение по минимальной цене-сложности**

4. Определение оптимального размера дерева

- **Кросс-проверочное отсечение по минимальной цене СЛОЖНОСТИ**
 1. При каждом новом ветвлении рассчитывается цена обучающей выборки и цена V- кратной кросс-проверки.
 2. Вычисляется функция $f = r.cost + h * d.$
 3. Параметр h увеличивается и пересчитываются функции f .
 4. Выбирается дерево из последовательности оптимально усеченных деревьев.

Пример кросс-проверочного отсеечения



Первый шаг: $h=0$	Второй шаг: $h=0,1$	Третий шаг $h=0,2$	Четвертый шаг $h=0,3$
$f5=0,2+0*5=0,2$ - максимальное $f4=0,3$ $f3=0,3$ $f2=0,5$ $f1=0,75$	$f5=0,2+0,1*5=0,7$ $f4=0,3+0,1*4=0,7$ $f3=0,3+0,1*3=0,6$ – становится максимальным $f2=0,5+0,1*2=0,7$ $f1=0,75+0,1=0,85$	$f3=0,3+0,2*3=0,9$ $f2=0,5+0,2*2=0,9$ – становится максимальным $f1=0,95$	$f2=0,5+0,3*2=1,1$ $f1=1,05$ становится максимальным

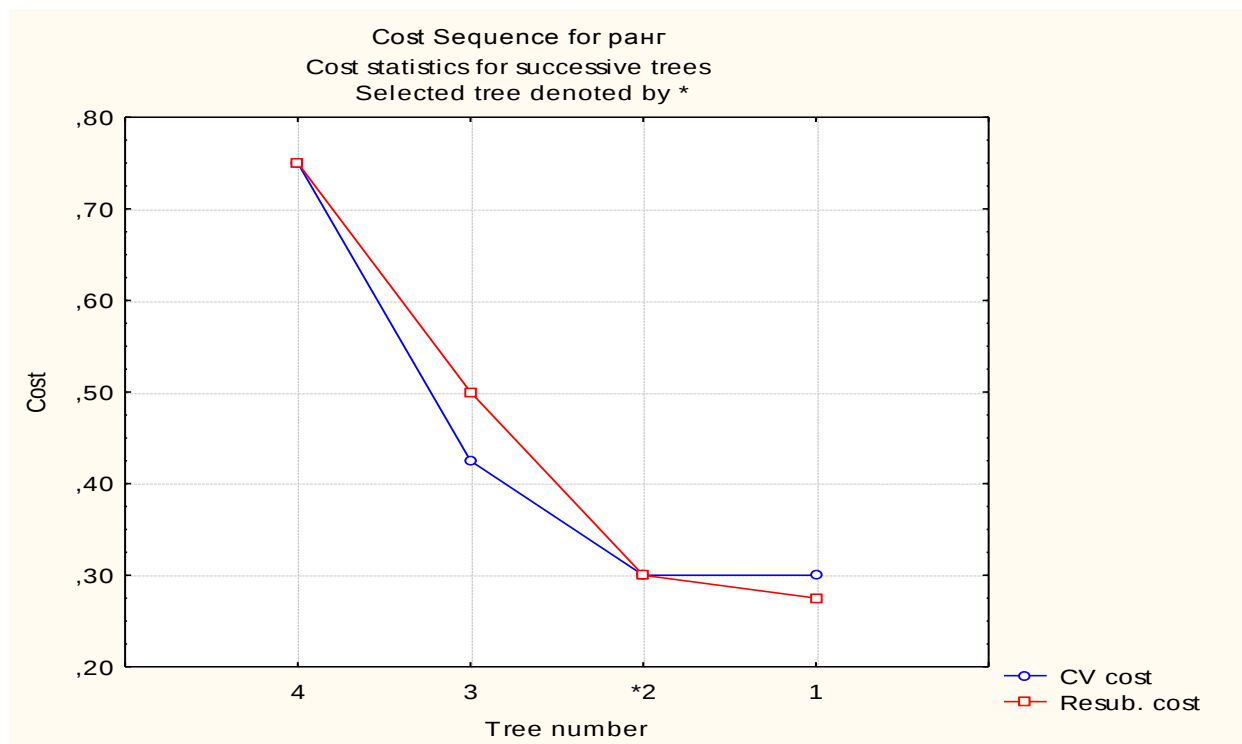
Пример кросс-проверочного отсечения

В результате получается последовательность максимальных деревьев, обладающих рядом свойств:

1. Деревья вложенные.
2. По мере приближения к корневой вершине отсекается все меньше вершин.
3. Деревья последовательности усекаются оптимально

Пример кросс-проверочного отсеечения

	Terminal	CV	Std.	Resub.	Node
1	5	0,200000	0,146098	0,2	0,000000
*2	3	0,300000	0,031623	0,300000	0,025000
3	2	0,425000	0,045415	0,500000	0,200000
4	1	0,750000	0,000000	0,750000	0,250000



типичные графики *цены обучения* и *цены кросс-проверки* для цепочки последовательно усекаемых деревьев

КОНЕЦ