Математические методы
(Исследование операций, Методы оптимизации)

Деревья решений

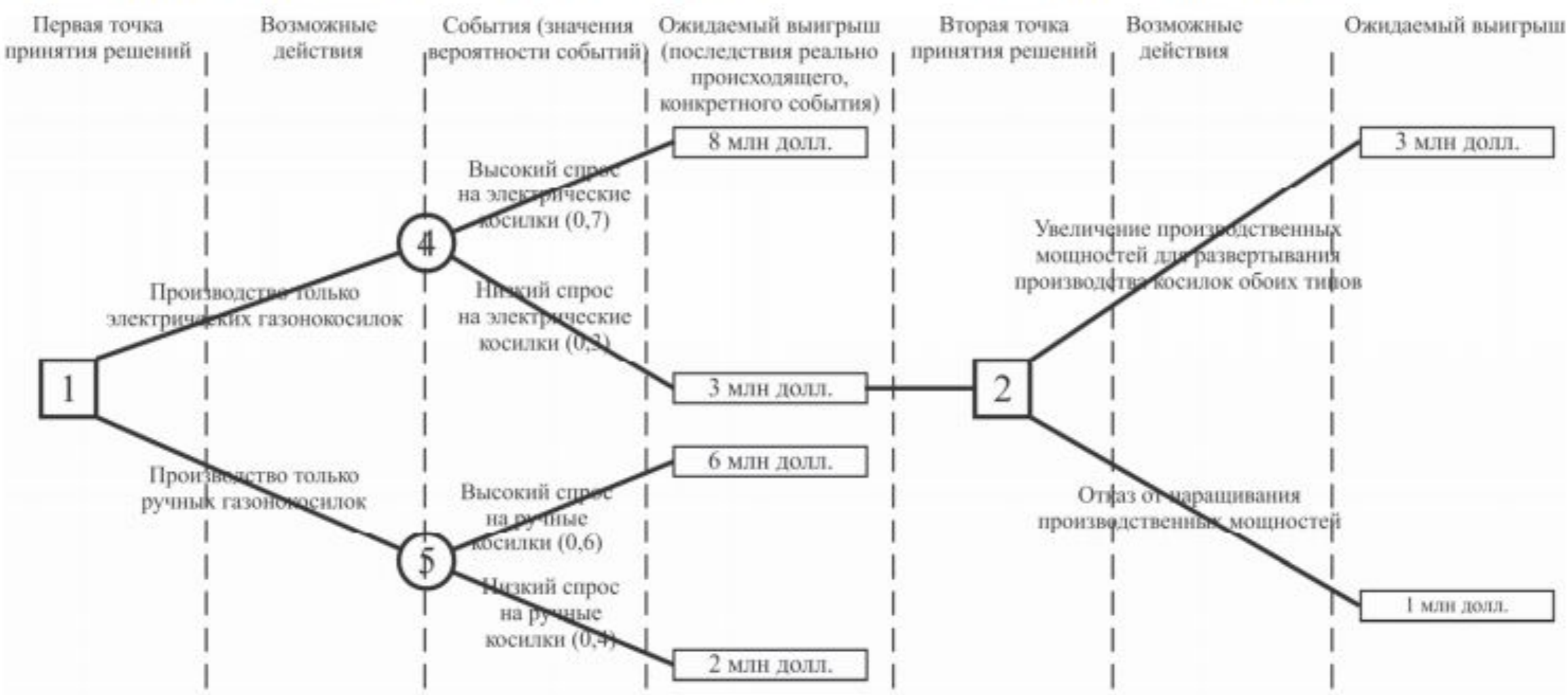
Актуальность

- **Принятие решений** является наиболее важным видом деятельности, осуществляемой менеджерами, и представляет собой единовременный акт окончательного выбора **одного из возможных вариантов действий по достижению целей организации**.
- Необходимость принятия решений обусловлена тем, что организации **под влиянием изменений внешней среды вынуждены адаптироваться к изменяющимся условиям функционирования** с помощью обратных связей – информации о состоянии объекта управления, представленной в виде отклонений параметров объекта управления от целей, эти отклонения называются проблемой.

Актуальность

- Когда нужно принять **несколько решений в условиях неопределенности**, когда каждое решение зависит от результата предыдущего решения или результатов испытаний, то применяют схему, называемую «деревом решений».
- Это графическое изображение процесса принятия решений, в котором **отражены альтернативные решения, альтернативные состояния среды, соответствующие вероятности и плюсы различных комбинаций.**

«Дерево решений» – это графическая схема того, к какому выбору в будущем приведет нас принятое сегодня решение



Особенности построения

- рисуют слева направо;
- участки, где принимаются решения, обозначают квадратами, участки проявления последствий – кругами;
- возможные решения обозначают пунктирными линиями, возможные последствия – сплошными линиями;
- «дерево решений» не может содержать в себе циклические элементы, т.е. каждый новый «лист» впоследствии может лишь «расщепляться», отсутствуют сходящиеся пути

Преимущества

- простота в понимании и интерпретации;
- не требует подготовки данных. Прочие методы анализа данных требуют нормализации данных, добавления фиктивных переменных, удаления пропущенных данных;
- использует модель «белого ящика»;
- позволяет работать с большим объемом информации без специальных подготовительных процедур.

Пример применения метода «дерево решений»

Задача. Для финансирования проекта бизнесмену нужно занять 15 000 руб. сроком на один год.

Банк может одолжить ему эти деньги под 15 % годовых или вложить в дело со 100%-ным возвратом суммы, но под 9 % годовых. Из прошлого опыта банкиру известно, что 4 % таких клиентов ссуду не возвращают.

Что делать? Давать ему заем или нет?

Решение

- Максимизируем ожидаемый в конце года чистый доход, который представляет собой разность суммы, полученной в конце года и инвестированной в его начале. Таким образом, если заем был выдан и возвращен, то
- Чистый доход = ((15 000 + 15 %) – 15 000) = 2250 руб.
- Если вложиться в другое дело, то чистый доход = ((15 000 + 9 %) – 15 000) = 1350 руб.
- Далее рассчитывается ожидаемый чистый доход с учетом вероятностей:
- ЧД 1 = ((15 000 + 2250)·0,96 + 0·0,04) – 15 000 = 1560 руб.
- ЧД 2 = (15 000 + 1350)·1,0 – 15 000 = 1350 руб.

Чистый доход в конце года

Чистый доход в конце года, руб.

Возможные исходы	Возможные решения		Вероятность
	Выдавать заем	Не выдавать (инвестировать)	
Клиент заем возвращает	2250	1350	0,96
Клиент заем не возвращает	–15 000	1350	0,04
Ожидаемый чистый доход	1560	1350	

Поскольку ожидаемый чистый доход больше для варианта А, то принимается решение выдать заем.

Дерево решений

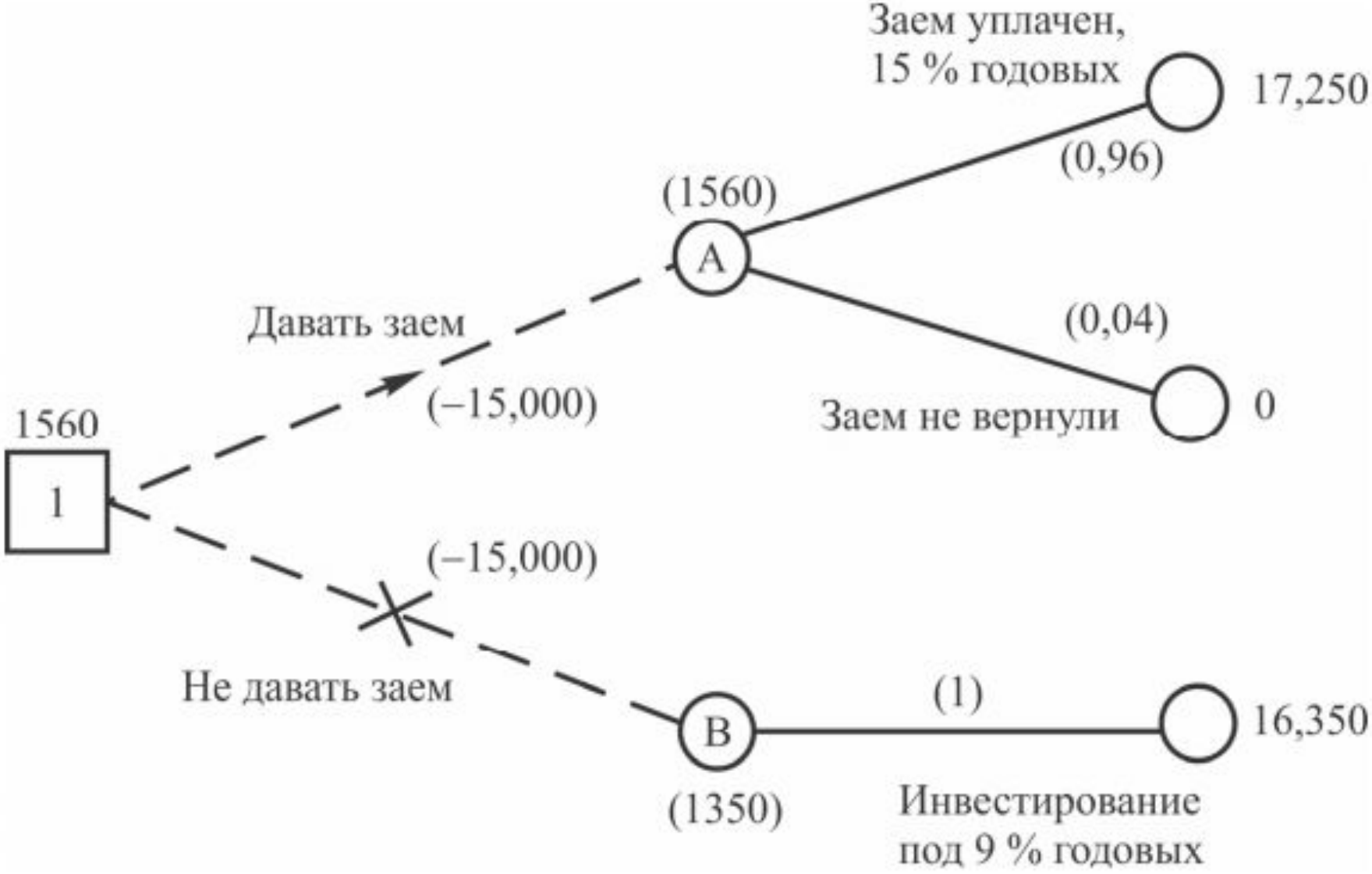


Рис. 2. «Дерево решений» банкира

Выводы

- необходимость принятия решения пронизывает все, что делает управляющий, ставя цели и добиваясь их достижения, поэтому эффективность принимаемых руководством решений во многом определяет эффективность функционирования всего предприятия;
- метод «дерево решений» – один из наиболее точных методов принятия управленческих решений (высока точность прогноза, сопоставимая с другими методами – статистикой, нейронными сетями);
- этот метод наглядно показывает альтернативные решения, альтернативные состояния среды, соответствующие вероятности и выигрыши для любых комбинаций.

Нечеткие деревья решений

- Однако может возникнуть случай, когда точно классифицировать объект по тому или иному признаку довольно трудно.
- Эти ситуации разрешаются благодаря возможностям **нечеткой логики**, когда говорят **не просто о принадлежности к кому-то классу**, признаку, атрибуту, **а о её степени**.
- При использовании нечетких деревьев решений (fuzzy decision trees) не теряются знания о том, что объект может обладать свойствами **как одного признака, так и другого в той или иной мере.**

Особенности построения нечеткого дерева решений

- отличительной чертой деревьев решений является то, что каждый пример определенно принадлежит конкретному узлу. В нечетком случае это не так.
- **Для каждого атрибута необходимо выделить несколько его лингвистических значений и определить степени принадлежности примеров к ним**;
- вместо количества примеров конкретного узла нечеткое дерево решений **группирует их степень принадлежности**

Алгоритм

Коэффициент – это соотношение примеров $D_j \in S^N$ узла N для целевого значения i, вычисляемый как:

$$P_i^N = \sum_{S^N} \min\left(\mu_N\left(D_j\right), \mu_i\left(D_j\right)\right), \quad (1)$$

где $mu_N\left(D_j\right)$ – степень принадлежности примера D_j к узлу N, $mu_i\left(D_j\right)$ – степень принадлежности примера относительно целевого значения i, S^N – множество всех примеров узла N. Затем находим коэффициент P^N, обозначающий общие характеристики примеров узла N. В стандартном алгоритме дерева решений определяется отношение числа примеров, принадлежащих конкретному атрибуту, к общему числу примеров. Для нечетких деревьев используется отношение $\frac{P_i^N}{P^N}$, для расчета которого учитывается степень принадлежности.

Алгоритм

Выражение

$$E(S^N) = -\sum_i \frac{P_i^N}{P^N} \cdot \log_2 \frac{P_i^N}{P^N}, \quad (2)$$

даёт оценку среднего количества информации для определения класса объекта из множества P^N.

На следующем шаге построения нечеткого дерева решений алгоритм вычисляет энтропию для разбиения по атрибуту A со значениями a_j:

$$E(S^N, A) = \sum_j \frac{P^{N|j}}{P^N} \cdot E(S^{N|j}), \quad (3)$$

где узел $N \mid j$ – дочерний для узла N.

Алгоритм

Алгоритм выбирает атрибут $A^{\times}$ с максимальным приростом информации:

$$G(S^N, A) = E(S^N) - E(S^N, A), \quad (4)$$

$$A^{\times} = argmax_A \, G(S, A), \quad (5)$$

Узел N разбивается на несколько подузлов $N \mid j$. Степень принадлежности примера D_k узла $N \mid j$ вычисляется пошагово из узла N как

$$\mu_{N|j}(e_k) = \min(\mu_{N|j}(D_k), \mu_{N|j}(D_k, a_j)), \quad (6)$$

где $\mu_{N|j}(D_k, a_j))$ показывает степень принадлежности D_k к атрибуту a_j.

Подузел $N \mid j$ удаляется, если все примеры в нем имеют степень принадлежности, равную нулю. Алгоритм повторяется до тех пор, пока все примеры узла не будут классифицированы либо пока не будут использованы для разбиения все атрибуты.

Принадлежность к целевому классу для новой записи находится по формуле

$$\sigma_j = \frac{\sum_l \sum_k P_k^l \cdot \mu_l(D_j) \cdot \chi_k}{\sum_l (\mu_l(D_j) \cdot \sum_k P_k^l)}, \quad (7)$$

где

P_k^l – коэффициент соотношения примеров листа дерева l для значения целевого класса k,

$\mu_l(D_j)$ – степень принадлежности примера к узлу l,

χ_k – принадлежность значения целевого класса k к положительному значению исхода классификации.

Пример

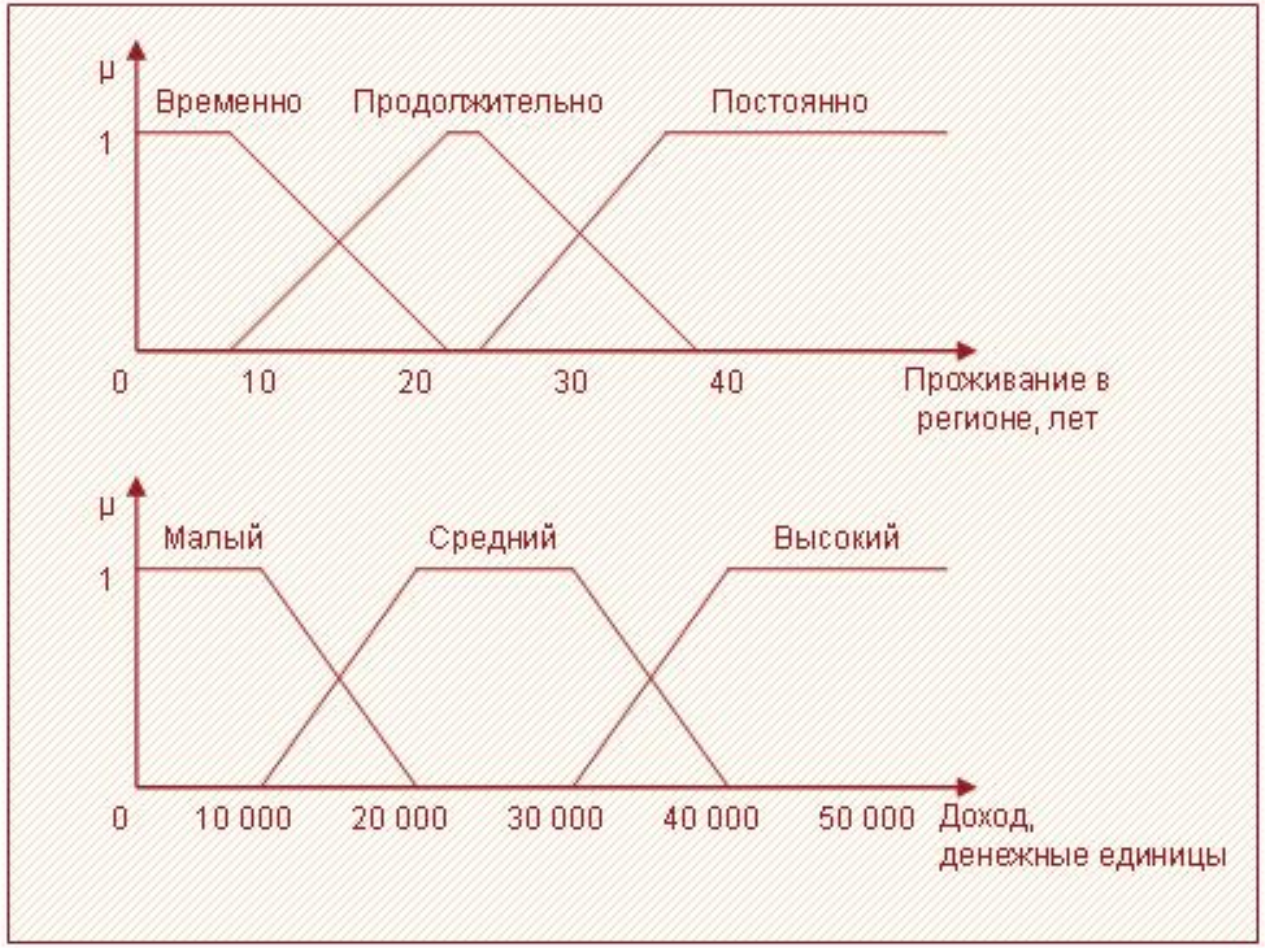
- В таблице 1 представлены **данные о клиентах** банка: **проживание в регионе** (в годах), **доход** (в денежных единицах) и **рейтинг выдачи** ему кредита.
- Необходимо построить нечеткое дерево решений, с помощью которого определить рейтинг выдачи кредита для клиента, который проживает в регионе 25 лет, и доход его составляет 32 000 (будет решаться задача регрессии).

Пример

№	Проживание в регионе	Доход	Рейтинг
D1	0	10 000	0,0
D2	10	15 000	0,0
D3	15	20 000	0,1
D4	20	30 000	0,3
D5	30	25 000	0,7
D6	40	35 000	0,9
D7	40	50 000	1,0

Нечеткие шкалы

Предположим, что атрибут **"проживание в регионе"** может принимать значения "временно", "продолжительно", "постоянно", а атрибут **"доход"** – "малый", "средний" и "высокий". Степень принадлежности каждого примера к значениям атрибутов представлена в таблице 2. Общий вид функции для атрибутов показан на рисунке 1.



Степень принадлежности примеров к атрибутам

№	Временно	Продолжительно	Постоянно	Малый	Средний	Высокий
D1	1	0	0	1	0	0
D2	0,8	0,2	0	0,6	0,4	0
D3	0,5	0,5	0	0,1	0,9	0
D4	0,2	0,8	0	0	1	0
D5	0	0,5	0,5	0	1	0
D6	0	0	1	0	0,6	0,4
D7	0	0	1	0	0	1

Расчеты

Рейтинг
0,0
0,0
0,1
0,3
0,7
0,9
1,0

В начале необходимо найти значение – общая энтропия.

$$P_{Да} = 0 + 0 + 0{,}1 + 0{,}3 + 0{,}7 + 0{,}9 + 1{,}0 = 3,$$

$$P_{нет} = 1 + 1 + 0{,}9 + 0{,}7 + 0{,}3 + 0{,}1 + 0 = 4,$$

$$P = P_{Да} + P_{нет} = 3 + 4 = 7,$$

$$E(S^N) = -\frac{3}{7} \cdot \log_2 \frac{3}{7} - \frac{4}{7} \cdot \log_2 \frac{4}{7} \approx 0{,}985 \text{ бит.}$$

Расчет нечетких показателей атрибута «проживание» (на примере «временноеоживание»)

Теперь рассчитаем $E(S^N$, **проживание в регионе**).

$\mathbf{P_{да}^{временно}}$ = min(0;1) + min(0;0,8) + min(0,1;0,5) + min (0,3;0,2) + min(0,7;0) + min(0,9;0) + min(1;0) = 0 + 0 + 0,1 + 0,2 + 0 + 0 + 0 = 0,3

$\mathbf{P_{нет}^{временно}}$ = min(1;1) + min(1;0,8) + min(0,9;0,5) + min (0,7;0,2) + min(0,3;0) + min(0,1;0) + min(0;0) = 1 + 0,8 + 0,5 + 0,2 + 0 + 0 + 0 = 2,5

$\mathbf{p^{временно}}$ = 0,3 + 2,5 = 2,8

E(проживание в регионе, временно)= $-\frac{0,3}{2,8} \cdot \log_2 \frac{0,3}{2,8} - \frac{2,5}{2,8} \cdot \log_2 \frac{2,5}{2,8} \approx 0,491$ бит.

Для продолжительного и постоянного проживания в регионе проводятся аналогичные вычисления. Результат сведем в таблицу 3.

Комментарии:
2,5/7, 2/7, 2,5/7 - получаются путем суммирования по столбцу «временно», «продолжительно», «постоянно»

0,985 бит отпределено на первом этапе расчетов
0,491 на предыдущем слайде

Для продолжительного и постоянного проживания в регионе проводятся аналогичные вычисления. Результат сведем в таблицу 3.

Таблица 3 – Итог расчетов для атрибута "проживание в регионе"

	Временно	Продолжительно	Постоянно
$P_{да}$	0,3	0,9	2,4
$P_{нет}$	2,5	1,7	0,4
E в битах	0,491	0,931	0,592

Отсюда находим энтропию:

$$E(S^N, \text{проживание в регионе}) = \frac{2,5}{7} \cdot 0,491 + \frac{2}{7} \cdot 0,931 + \frac{2,5}{7} \cdot 0,592 = 0,653 \text{ бит.}$$

Рассчитаем прирост информации для данного атрибута.

$$G(S^N, \text{проживание в регионе}) = 0,985 - 0,653 = 0,332 \text{ бит.}$$

Расчеты для нового атрибута («доход»)

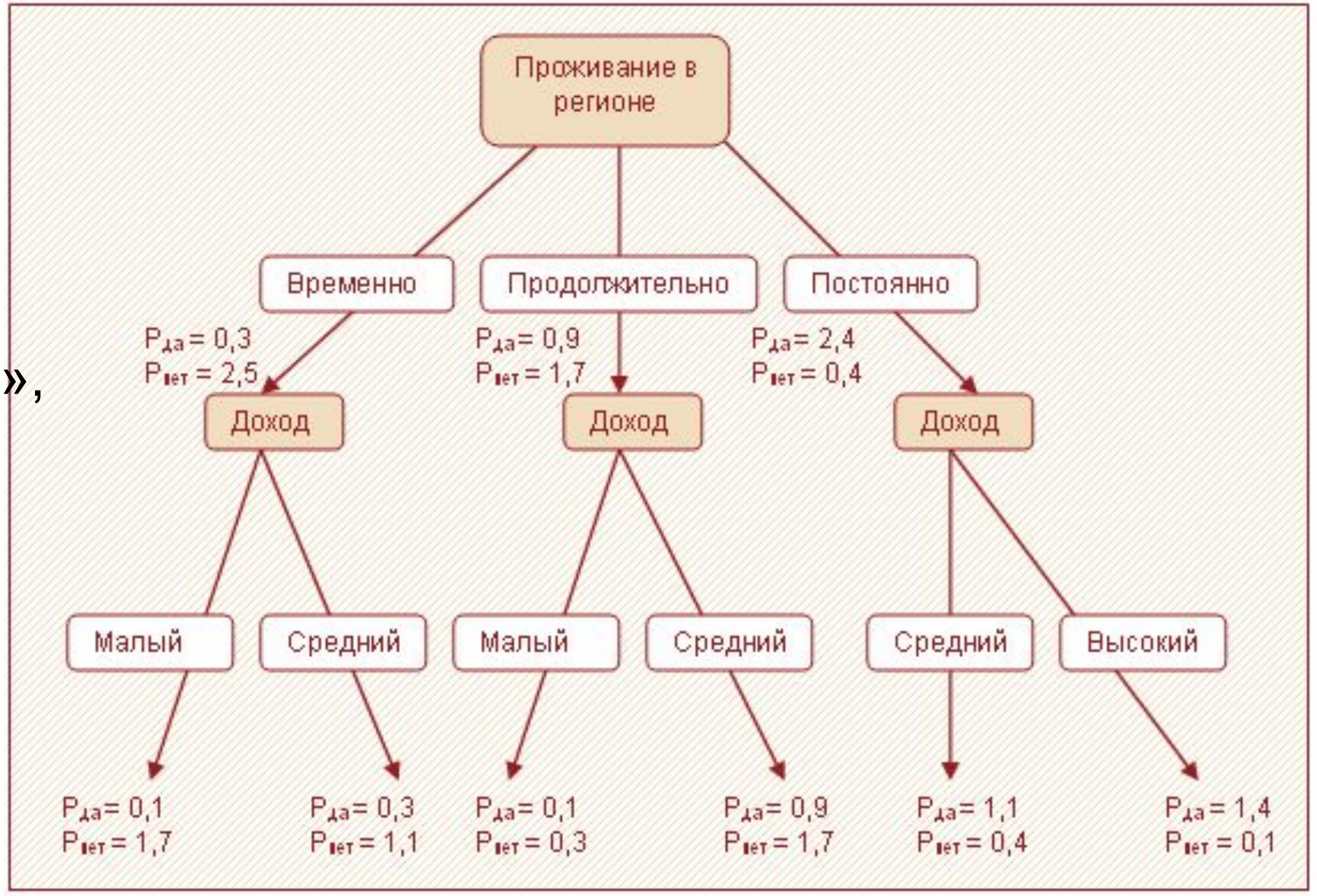
- Проводя подобные вычисления для **атрибута "доход"**, получаем
- E(SN, доход) = 0,691 бит,
- G(SN, доход) = 0,294 бит.
- Максимальный прирост информации обеспечивает **атрибут "проживание в регионе"** (0,332 бит>0,294 бит), следовательно, разбиение начнется с него.
- На следующем шаге алгоритма необходимо для каждой записи рассчитать степень принадлежности к каждому новому узлу

Принадлежность записей новым узлам дерева

Проживание в регионе	временно			продолжительно			постоянно		
доход	малый	средний	высокий	малый	средний	высокий	малый	средний	высокий
D1	1	0	0	0	0	0	0	0	0
D2	0,6	0,4	0	0,2	0,2	0	0	0	0
D3	0,1	0,5	0	0,1	0,5	0	0	0	0
D4	0	0,2	0	0	0,8	0	0	0	0
D5	0	0	0	0	0,5	0	0	0,5	0
D6	0	0	0	0	0	0	0	0,6	0,4
D7	0	0	0	0	0	0	0	0	1

Построение нечеткого дерева решений

Комментарии:
1,7; 1,1; и т.д. - получаются путем суммирования по столбцу «временно», «продолжительно», «постоянно»



Теперь определим кредитный рейтинг для клиента, проживающего в регионе 25 лет, и с доходом 30 000.

За положительный исход в данной задаче принято одобрение в выдаче кредита, поэтому $X_{да} = 1,0$ $X_{нет} = 0,0$. Новый клиент принадлежит к двум узлам: [проживание в регионе = продолжительно и доход = средний] и [проживание в регионе = постоянно и доход = средний], со степенями 0,8 и 0,2 соответственно. Подставляя полученные значения в формулу (7), рассчитываем кредитный рейтинг:

$$\delta = \frac{0,9 \times 0,8 \times 1,0 + 1,7 \times 0,8 \times 0,0 + 1,1 \times 0,8 \times 1,0 + 0,4 \times 0,2 \times 0,0}{(0,9 + 1,7) \times 0,8 + (1,1 + 0,4) \times 0,2} = 0,395$$

В итоге мы получили кредитный рейтинг, равный 0,395. Он означает, что степень принадлежности записи к тому, что кредит клиенту будет выдан, равна 0,395, а к невыдаче – 0,605. Следовательно, этому клиенту банком будет отказано.

Комментарии: 0,8 и 0,2 взяты как нечеткая оценка 25 лет по графику нечетких шкал. Обведенные зеленым с предыдущего слайда