Data Mining

классификация и регрессия

Классификация

Постановка задачи

- Если входные и выходные переменные модели непрерывные — перед нами задача регрессии.
- Если выходная переменная одна и она является дискретной (метка класса), то речь идет о задаче классификации.

Этап первый

- Поскольку метка класса каждого примера предварительно задана, построение классификационной модели часто называют обучением с учителем.
- В процессе обучения формируются правила, по которым производится отнесение объекта к одному из

Этап второй

- модель применяется для классификации новых, ранее неизвестных объектов и наблюдений

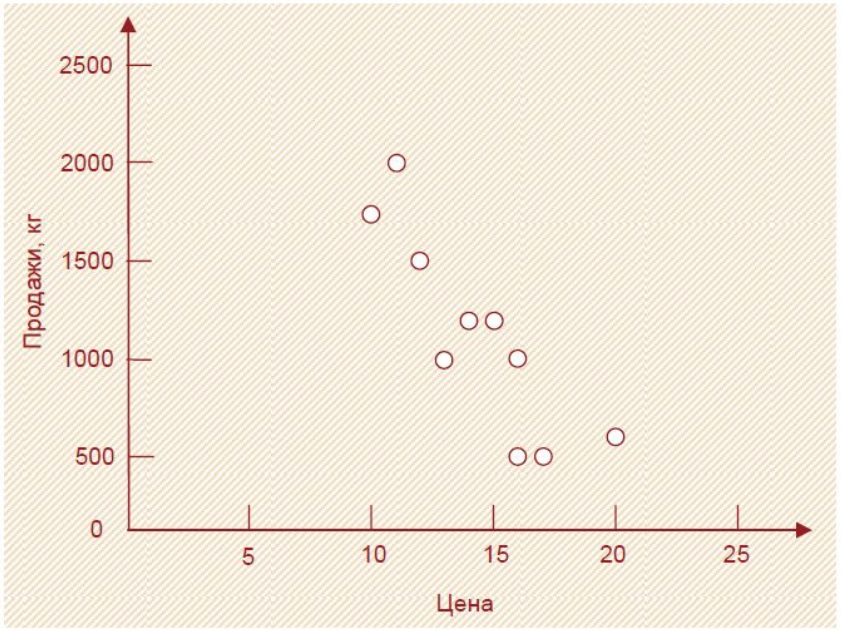
Линейная регрессия

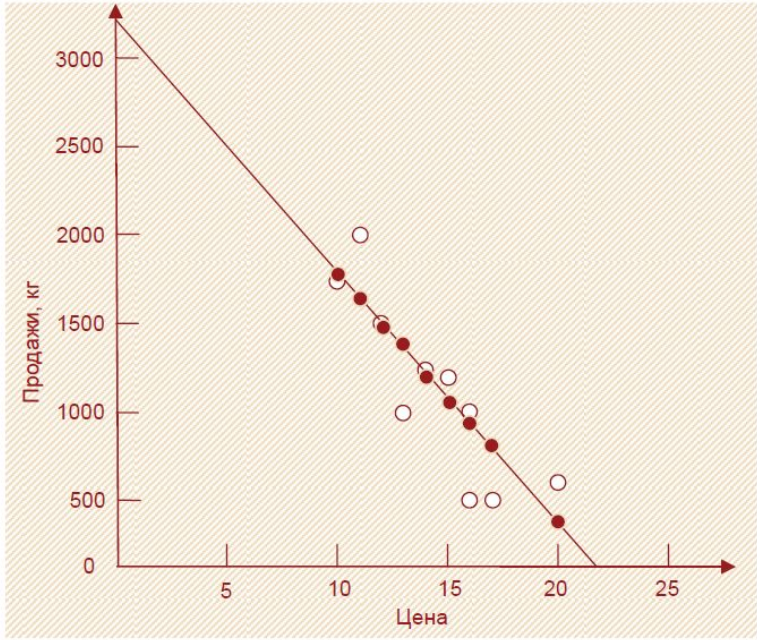
- Задача линейной регрессии заключается в нахождении коэффициентов уравнения линейной регрессии, которое имеет вид: $y = b_0 + b_1 x_1 + b_2 x_2 + ...+ b_n x_n$,

- где *y — выходная (зависимая) переменная модели;*
- $x_1, x_2, \ldots, x_n$ *— входные (независимые)* переменные;
- b_i *— коэффициенты линейной регрессии (*b_0 *— свободный член).*

- Задача линейной регрессии заключается в подборе коэффициентов b_i *уравнения таким* образом, чтобы на заданный входной вектор $\mathbf{X} = (x_1, x_2, \ldots, x_n)^T$ ***регрессионная модель*** формировала желаемое выходное значение *y*

№ месяца	Цена за 1 кг, x	Количество проданного картофеля y, кг	Количество проданного картофеля, оцененное с помощью регрессии $\hat{y}$, кг
1	13	1000	1323,4
2	20	600	305,6
3	17	500	741,8
4	15	1200	1032,6
5	16	1000	887,2
6	12	1500	1468,8
7	16	500	887,2
8	14	1200	1178,0
9	10	1700	1759,6
10	11	2000	1614,2





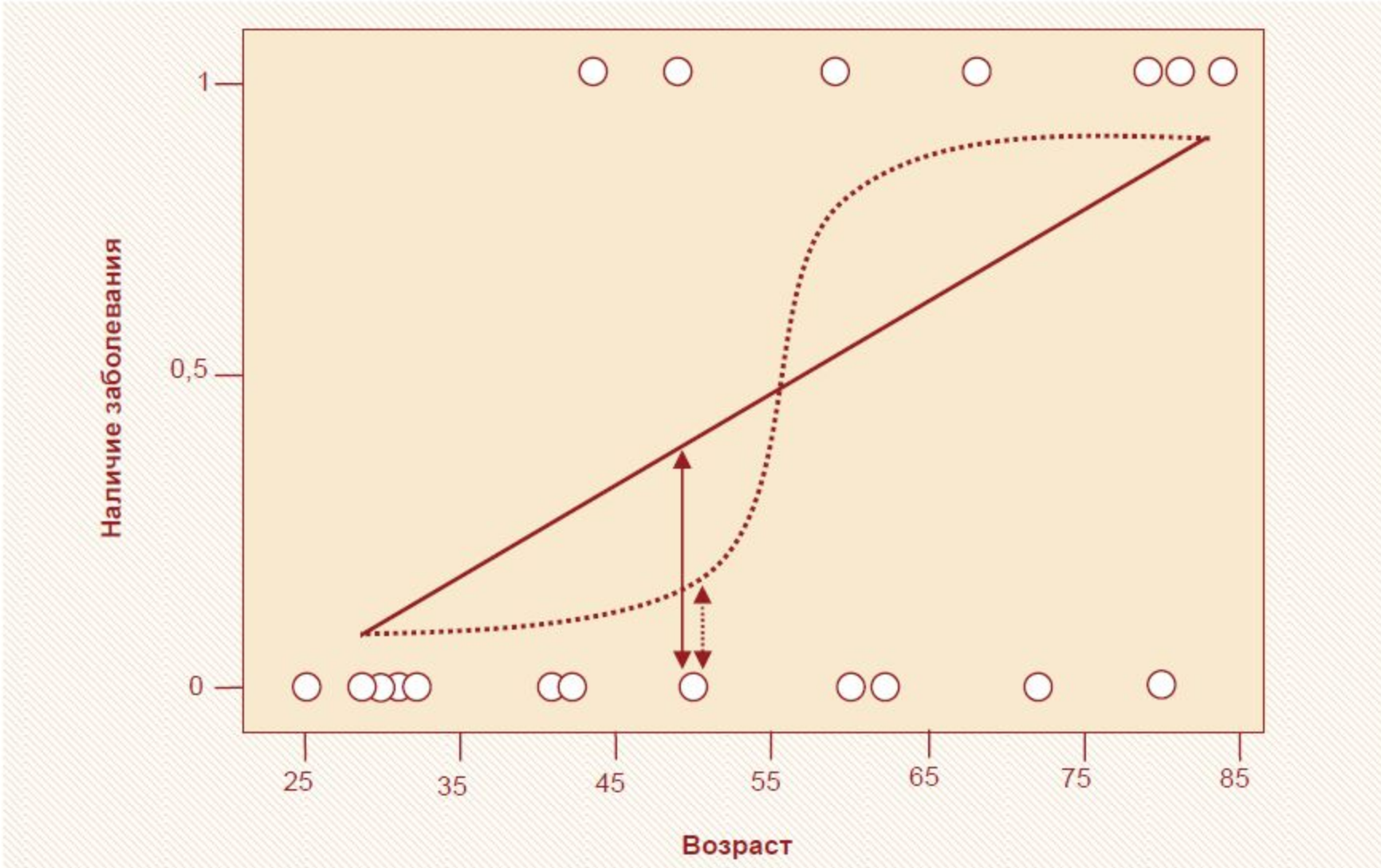
- *Линия регрессии — это прямая наилучшего приближения для набора пар значений входной и* выходной переменной

Логистическая регрессия

- задачи предсказания вероятности некоторого события в зависимости от значений набора независимых переменных

- логистическая регрессия служит не для предсказания значений зависимой переменной, а для **оценки вероятности** того, что зависимая переменная примет заданное значение.

№ пациента	Возраст, x	Наличие заболевания, y
1	25	0
2	29	0
3	30	0
4	31	0
5	32	0
6	41	0
7	41	0
8	42	0
9	44	1
10	49	1
11	50	0
12	59	1
13	60	0
14	62	0
15	68	1
16	72	0
17	79	1
18	80	0
19	81	1
20	84	1


Наличие заболевания
1
0,5
0
25
35
45
55
65
75
85
Возраст

Деревья решений

Методы, основанные на обучении

- Дерево решений — это древовидная иерархическая модель, где в каждом узле производится проверка определенного атрибута (признака) с помощью правила

- *Деревья решений — иерархические древовидные структуры, состоящие из решающих правил вида «если… то…» и позволяющие выполнять классификацию объектов. В дереве каждому объекту соответствует единственный узел, дающий решение.*

- Деревья решений — это модели, основанные на обучении. Процесс обучения сравнительно прост в настройке и управлении.

- Процесс обучения деревьев решений быстр и эффективен.

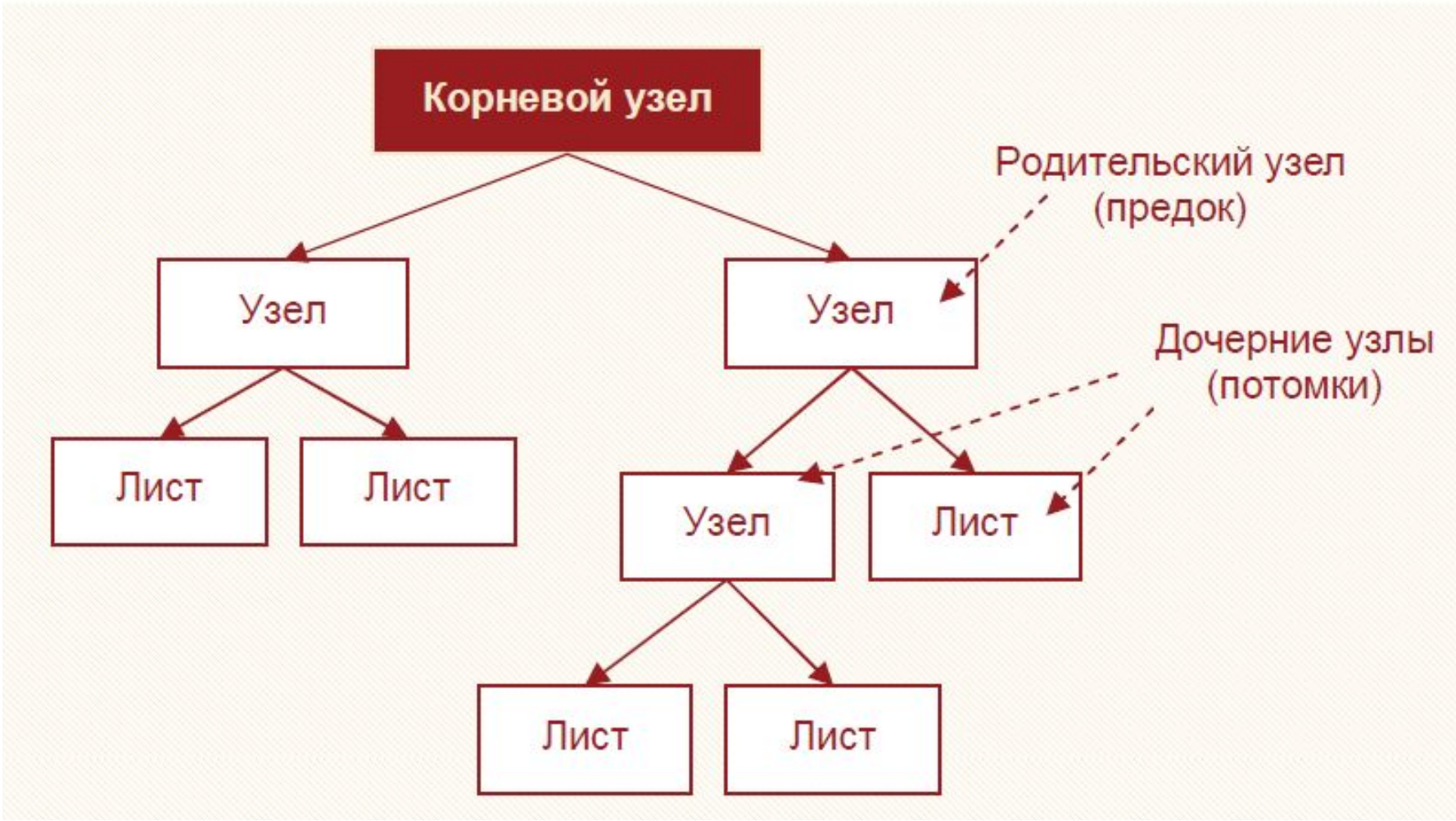
- Деревья решений универсальны — способны решать задачи как классификации, так и регрессии.

- Деревья решений обладают высокой объясняющей способностью и интерпретируемостью.

Построение дерева

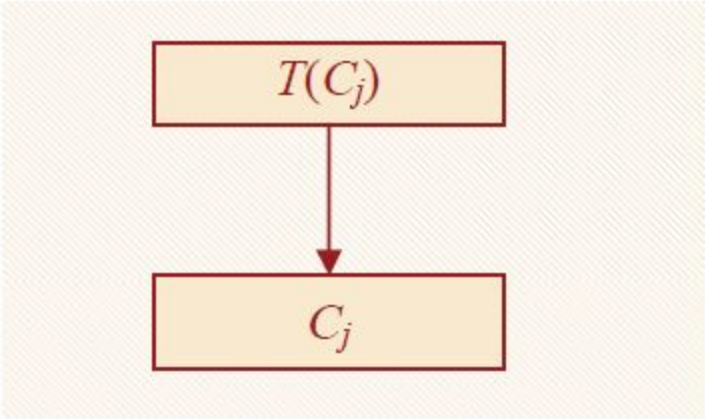
- *Описание атрибутов*
- *Предварительное определение классов*
- *Различимость классов*
- *Полнота данных*

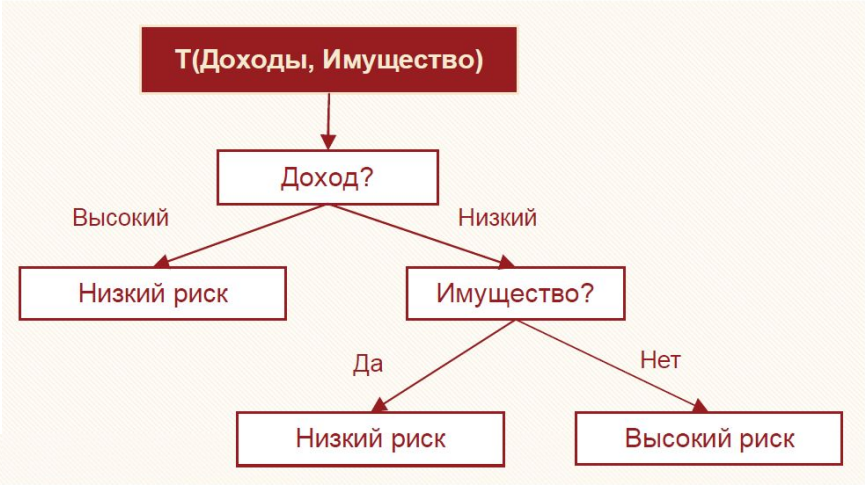
Структура дерева решений



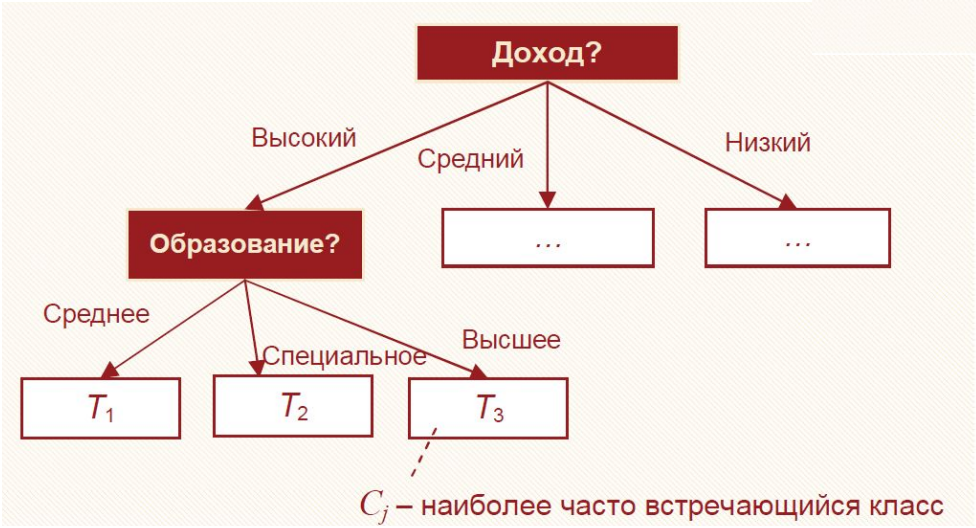
Принцип «разделяй и властвуй»

1





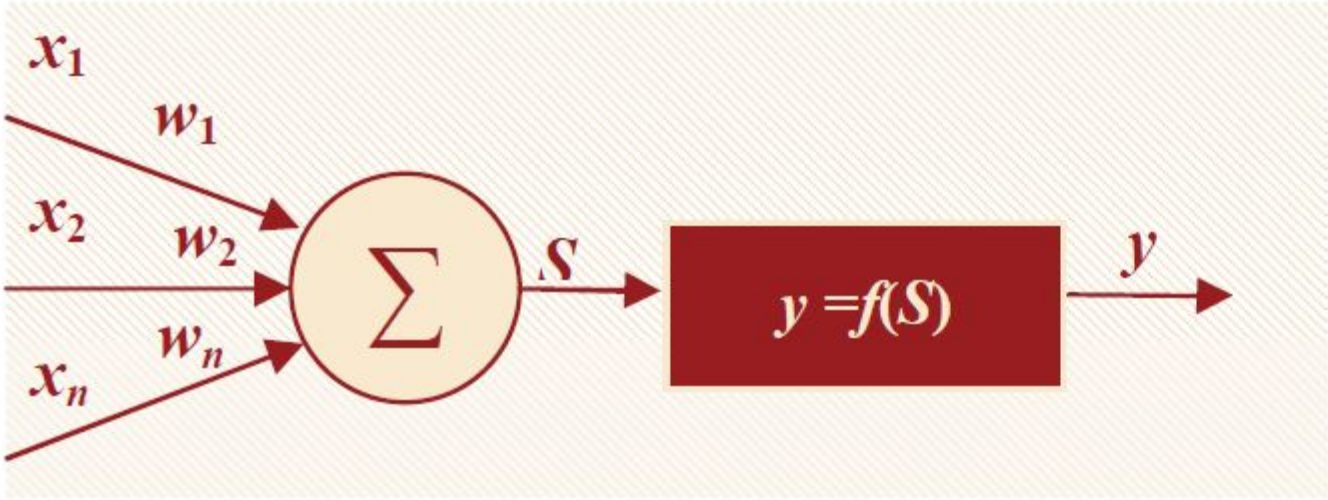
2



Искусственные нейронные сети

Методы, основанные на обучении

- *Искусственная нейронная сеть представляет собой параллельно-распределенную систему процессорных элементов (нейронов), способных выполнять простейшую обработку данных, которая может настраивать свои параметры в ходе обучения на эмпирических данных. Накопленные знания нейронной сети сосредоточены в весах межэлементных связей.*



свойства нейронных сетей

1. *Нелинейность.*
2. *Обучение на примерах*
3. *Параллельная обработка данных.*
4. *Адаптивность.*
5. *Отказоустойчивость.*

1. **Входные данные хорошо интерпретируются**
2. **Желаемые результаты также хорошо интерпретируются**
3. **Доступный опыт**

Выбор числа нейронов в многослойном персептроне

- Число нейронов во входном и выходном слоях жестко определяется числом входных и выходных переменных модели соответственно.

- Число нейронов в скрытых слоях и число скрытых слоев выбираются таким образом, чтобы количество образованных ими связей было как минимум в два-три раза меньше числа обучающих примеров.

Обучение сети