Классификация данных

Лекция 5

Определение

Классификация – это процесс определения принадлежности объектов к определенным классам.

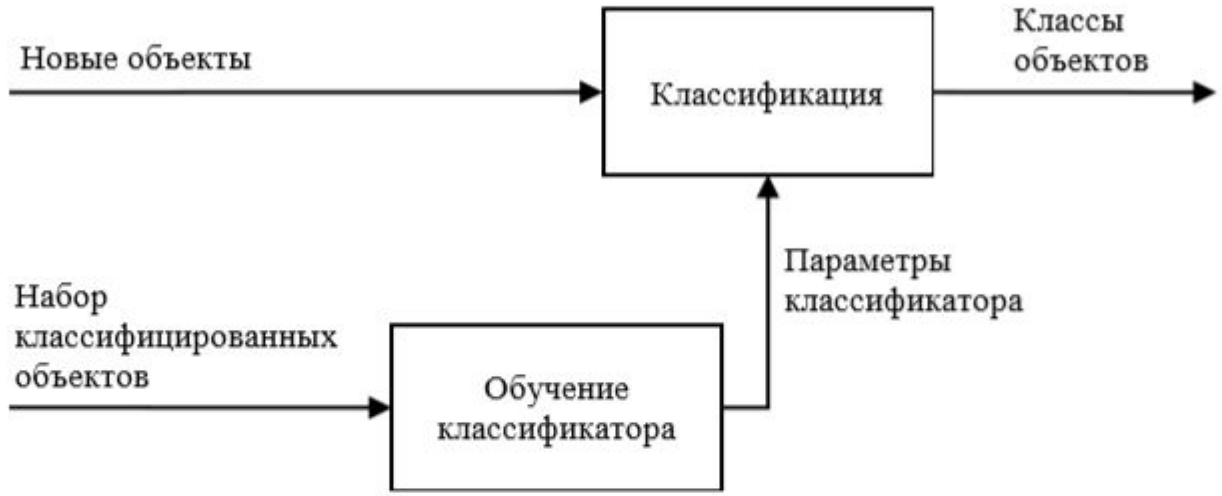
- классификация относится к типу задач обучения с учителем (Supervised Learning в терминах Machine Learning).
- Предполагается, что имеется некоторая выборка данных, в которой представлены объекты нескольких классов.
- При этом выборка содержит как свойства объектов, так и признак принадлежности объекта к какому-либо классу.

Применение задач классификац

- Существует много практических задач классификации.
- В промышленности при оценке качества продукции возникает задача подразделения изделий на годные и бракованные.
- В банковском секторе при выдаче кредитов возникает задача подразделения заемщиков на кредитоспособных и некредитоспособных.
- В медицине при оценке состояния здоровья возникает задача постановки диагноза.

Два этапа

- Применение классификации производится в два этапа.
- 1 – выполняется обучение классификатора на некотором наборе данных, а
- 2 – непосредственная классификация новых объектов

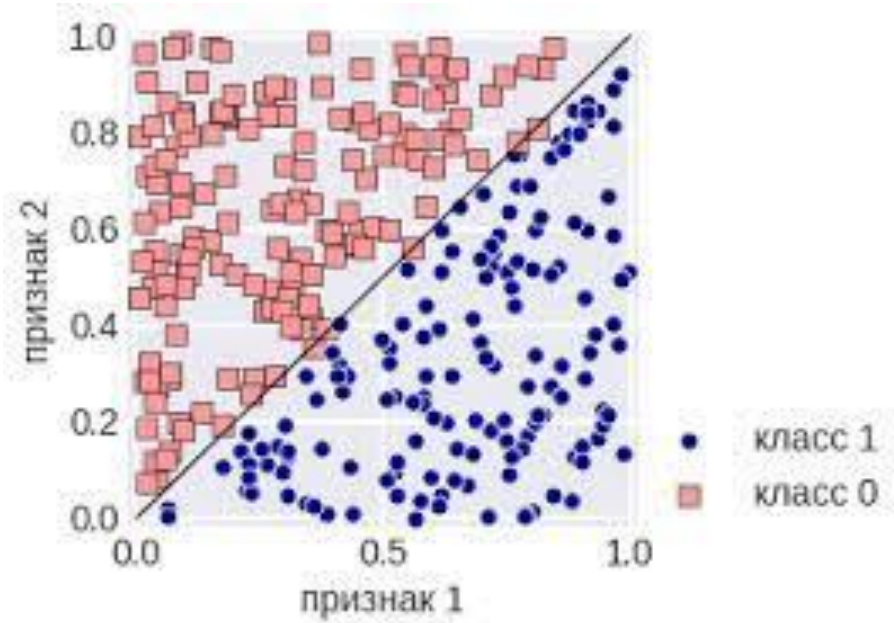


Виды классификации

- Различают бинарную и множественную классификацию.
- Бинарная классификация предполагает наличие двух классов, множественная – трех и более классов.
- Классификация выполняется с помощью специальных методов (алгоритмов). Известно большое количество алгоритмов классификации.

Бинарная классификация

- определение принадлежности некоего объекта к одному из двух возможных классов.



Примеры бинарной классификации

- является ли сообщение электронной почты «нормальным» или представляет собой спам;
- здоров или болен пациент;
- является ли заемщик банка надежным или ненадежным;
- качественная или бракованная деталь.

Методы бинарной классификации

- логистическая регрессия (Logistic Regression);
- «наивный» байесовский классификатор (Naive Bayes Classifier);
- метод опорных векторов (Support Vector Machine, SVM);
- нейронная сеть (Neural Network).

Логистическая регрессия

– один из методов бинарной классификации данных.

Алгоритм применения логистической регрессии:

1 Подготовка обучающей выборки – кодирование классов числами.
2 Задание функций штрафа.
3 Задание целевой функции.
4 Задание начальных значений коэффициентам функции.
5 Численное решение.

Численное решение логистической регрессии

$$z = x \cdot \theta ; \quad (1)$$

$$h(x_j) = \frac{1}{1 + e^{-z}}; \quad (2)$$

$$CF(h_j, y_j) = -(1 - y_j) \cdot \ln(1 - h_j) - y_j \cdot \ln(h_j). \quad (3)$$

Другой вариант решения

- В ряде случаев использование численных методов может приводить к ошибкам вычислений, поэтому иногда удобнее использовать формулу в другом варианте:

$$CF(h_j, y_j) = \begin{cases} -\ln(1-h_j), & y_j = 0 \\ -\ln(h_j), & y_j = 1 \end{cases}.$$

Оптимизационная задача

- Оптимизационная задача по-прежнему формулируется как задача минимизации функции штрафа:

$$CF = \sum_j CF(h_j, y_j) \to \min .$$

Численное решение задачи логистической регрессии с помощью Microsoft Excel
Шаг 1

1. В соответствии с предложенным выше алгоритмом представим исходные данные и расчетные формулы (режим

	A	B	C	D	E	F	G
1	X0	X1	X2	y	z	h(x)	cost(h,y)
2	1	1	6	0	2,000	0,881	2,127
3	1	3	4	0	2,000	0,881	2,127
4	1	2	2	1	-1,000	0,269	1,313
5	1	3	3	1	1,000	0,731	0,313
6	1	4	3	1	2,000	0,881	0,127
7							6,007
8							
9		theta0	-5				
10		theta1	1				
11		theta2	1				

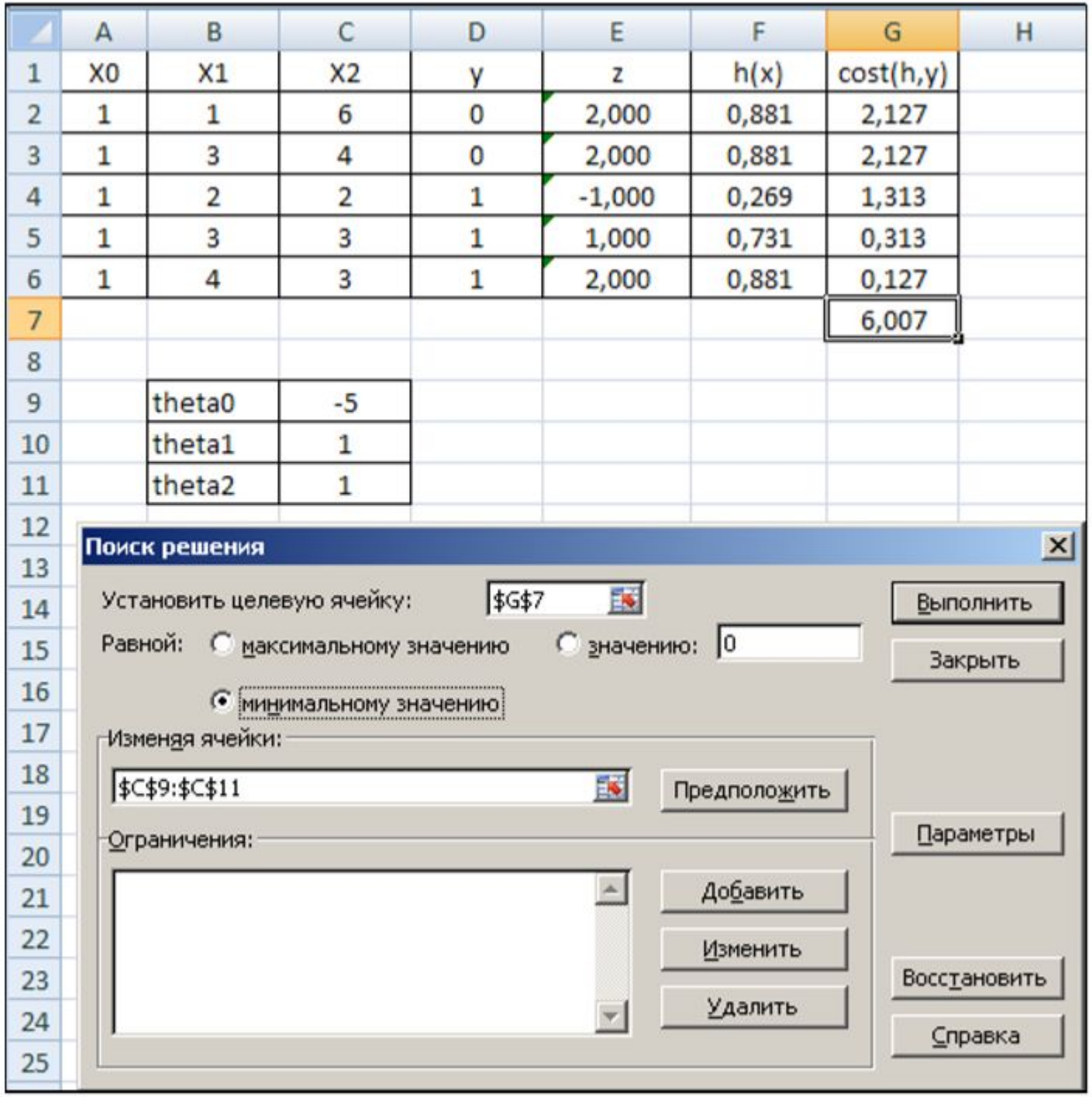
Логистическая регрессия в Excel (режим формул)

	A	B	C	D	E	F	G
1	X0	X1	X2	y	z	h(x)	cost(h,y)
2	1	1	6	0	=МУМНОЖ(A2:C2;$C$9:$C$11)	=1/(1+EXP(-E2))	=-LN(1-F2)
3	1	3	4	0	=МУМНОЖ(A3:C3;$C$9:$C$11)	=1/(1+EXP(-E3))	=-LN(1-F3)
4	1	2	2	1	=МУМНОЖ(A4:C4;$C$9:$C$11)	=1/(1+EXP(-E4))	=-LN(F4)
5	1	3	3	1	=МУМНОЖ(A5:C5;$C$9:$C$11)	=1/(1+EXP(-E5))	=-LN(F5)
6	1	4	3	1	=МУМНОЖ(A6:C6;$C$9:$C$11)	=1/(1+EXP(-E6))	=-LN(F6)
7							=СУММ(G2:G6)

Шаг 2-3

2 Выполним численное решение с помощью инструмента «Поиск решения»

3 В результате численного решения будут определены параметры функции линейного разделения. Визуальная проверка показывает корректность разделения двух классов



	A	B	C	D	E	F	G	H
1	X0	X1	X2	y	z	h(x)	cost(h,y)	
2	1	1	6	0	2,000	0,881	2,127	
3	1	3	4	0	2,000	0,881	2,127	
4	1	2	2	1	-1,000	0,269	1,313	
5	1	3	3	1	1,000	0,731	0,313	
6	1	4	3	1	2,000	0,881	0,127	
7							6,007	
8								
9		theta0	-5					
10		theta1	1					
11		theta2	1					
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								

Поиск решения

Установить целевую ячейку: $G$7

Равной: максимальному значению / значению: 0

минимальному значению

Изменяя ячейки: $C$9:$C$11

Ограничения:

Выполнить

Закрыть

Предположить

Параметры

Добавить

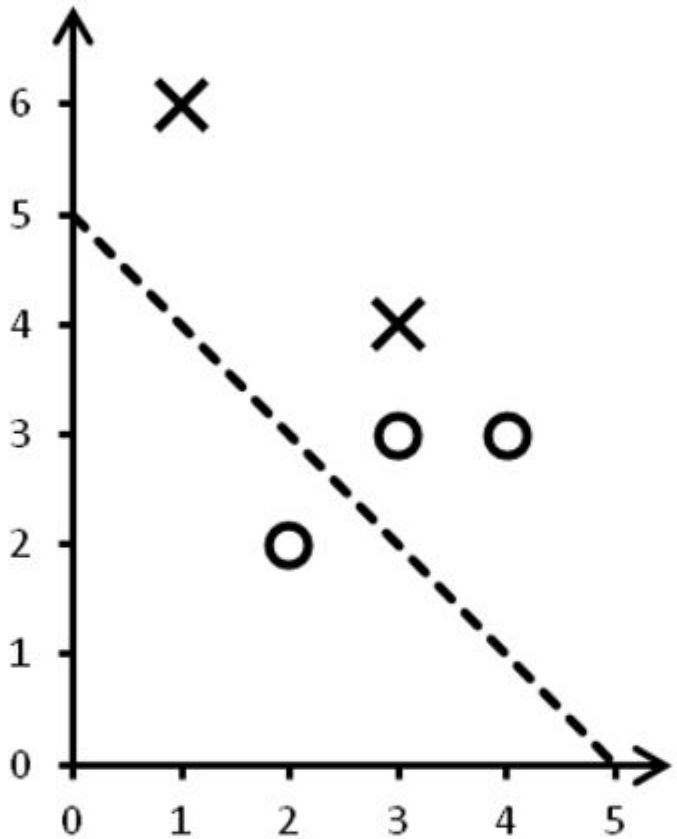
Изменить

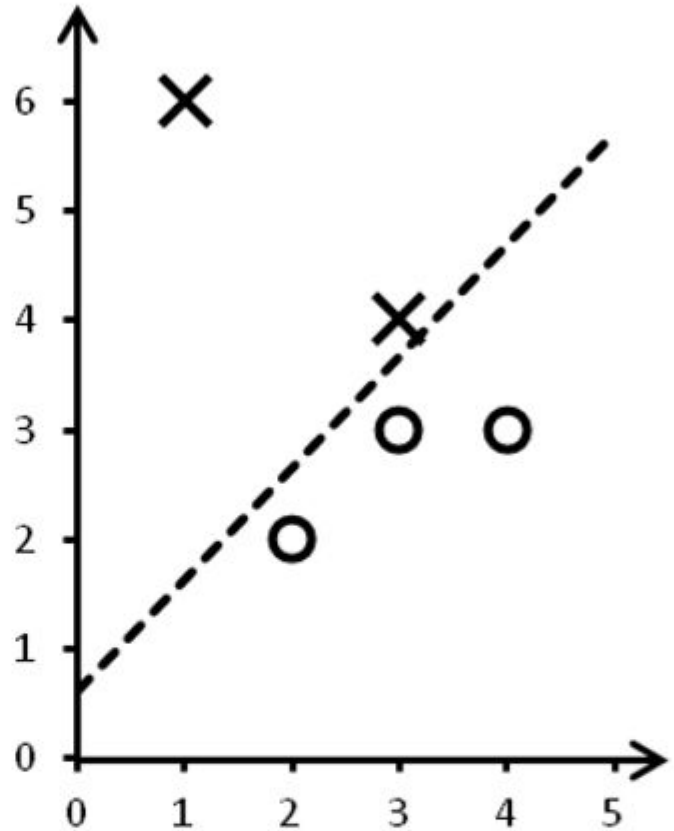
Удалить

Восстановить

Справка

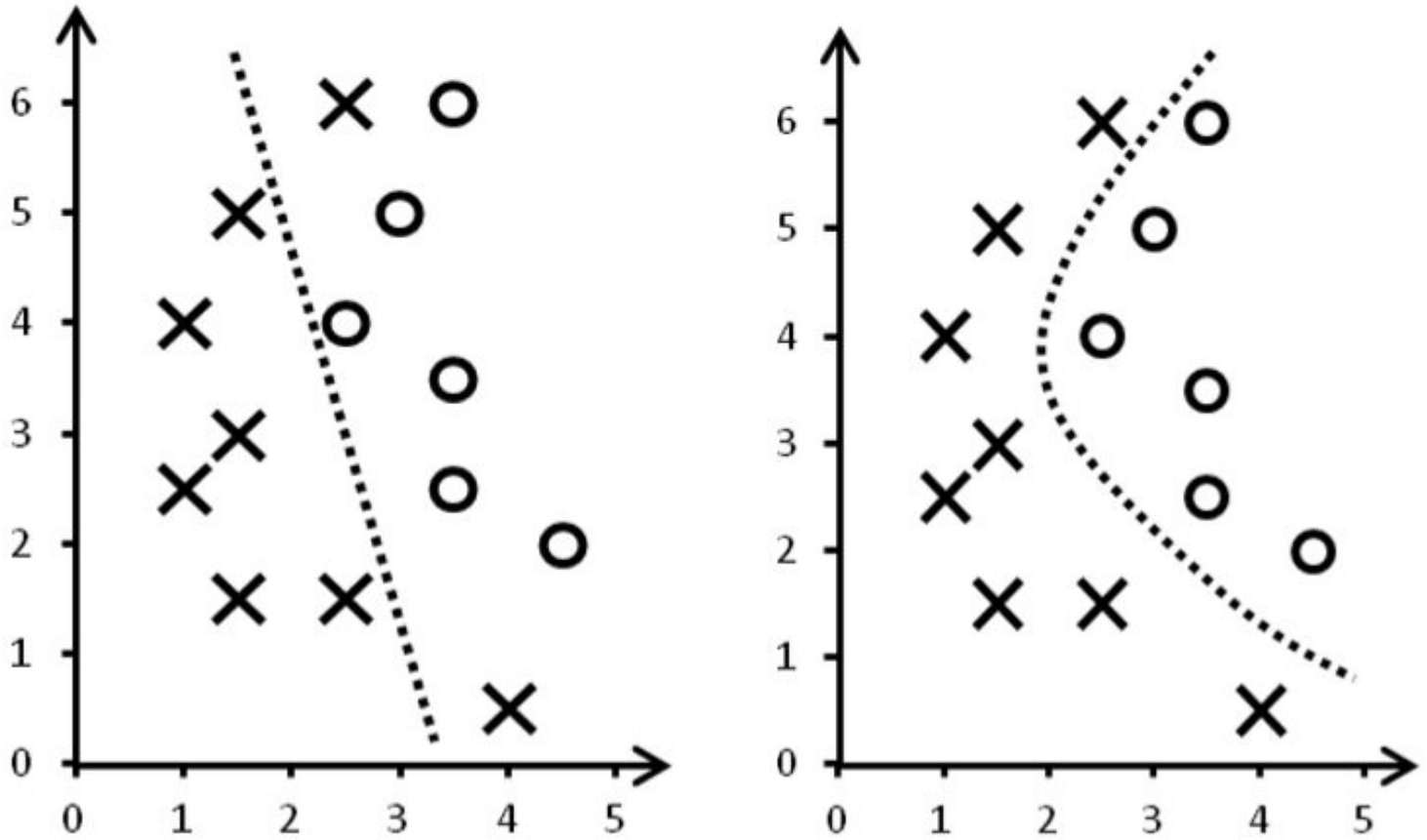
Визуальное представление классов





Проблема линейной разделимости

- Зачастую в реальных задачах бинарной классификации данные не могут быть разделены на два класса линейной функцией



Способы решения проблемы

- Возможны следующие способы решения этой проблемы:

- применение нелинейной функции гипотезы;

- принципиальная замена логистической регрессии другим методом, например, нейросетевым классификатором.

Качество классификации

- Очевидно, что при бинарной классификации возможны четыре сочетания реального класса каждого из объектов выборки данных и предположения алгоритма о классе объекта
- Правильно классифицированные алгоритмом объекты относятся либо к группе «true positives», либо к группе «true negatives». Неправильно классифицированные алгоритмом объекты относятс
группе «false negati

		Реальность	
		+	-
Предположение алгоритма	+	True positives (TP)	False positives (FP) Ошибка I рода
	-	False negatives (FN) Ошибка II рода	true negatives (TN)

Последствия ошибок классификации

- Реальные алгоритмы допускают ошибки классификации двух видов:
- ошибки I рода;
- ошибки II рода.

		Реальность	
		Нормальное письмо	Письмо с вирусом
Предположение алгоритма	Нормальное письмо	Письмо пропущено в почтовый ящик	Письмо пропущено в почтовый ящик. Последствие: заражение компьютера вирусом
	Письмо с вирусом	Письмо отброшено. Последствие: пользователь не получит важную информацию	Письмо отброшено

Ошибки классификации объектов могут привести к последующим неправильным решениям и нежелательным последствиям

Методы оценки качества классификации

- Существует несколько методов оценки качества классификации. Одним из методов является оценка с помощью F-критерия, выполняемая в четыре этапа:

1 Подсчет количества каждого сочетания случаев.

2 Расчет точности (precision)

$$P=\frac{TP}{TP+FP}$$

3 Расчет чувствительности (recall)

$$R=\frac{TP}{TP+FN}.$$

4 Расчет F-критерия

$$F=\frac{2\cdot P\cdot R}{P+R}.$$

Пример

Предположим, что в электронный почтовый ящик пришло 10 сообщений, часть из которых является нормальными, а часть – спамом

№	Вид сообщения	«Мнение» антивируса
1	письмо	письмо
2	спам	письмо
3	письмо	спам
4	спам	письмо
5	письмо	спам
6	письмо	письмо
7	спам	спам
8	письмо	письмо
9	письмо	спам
10	письмо	письмо

Возможные варианты

- Рассчитаем количество всех четырех сочетаний

		Реальность	
		письмо	Спам
«Мнение» антивируса	письмо	4	2
	спам	3	1

- В соответствии с формулами из слайда 21

$$P = \frac{4}{4+2} \approx 0{,}667\,;\ R = \frac{4}{4+3} \approx 0{,}571\,;\ F = \frac{2 \cdot 0{,}667 \cdot 0{,}571}{0{,}667 + 0{,}571} \approx 0{,}615\,.$$

- Для идеального алгоритма, не совершающего ошибок, F=0 .
- Для проверки качества классификатора можно использовать репозиторий открытых наборов данных

Множественная классификация

- Задачей множественной классификации является определение принадлежности некоего объекта к одному из нескольких (трех или более) возможных классов, например постановка диагноза пациенту

методами множественной классификации

- Наиболее известными методами множественной классификации являются:

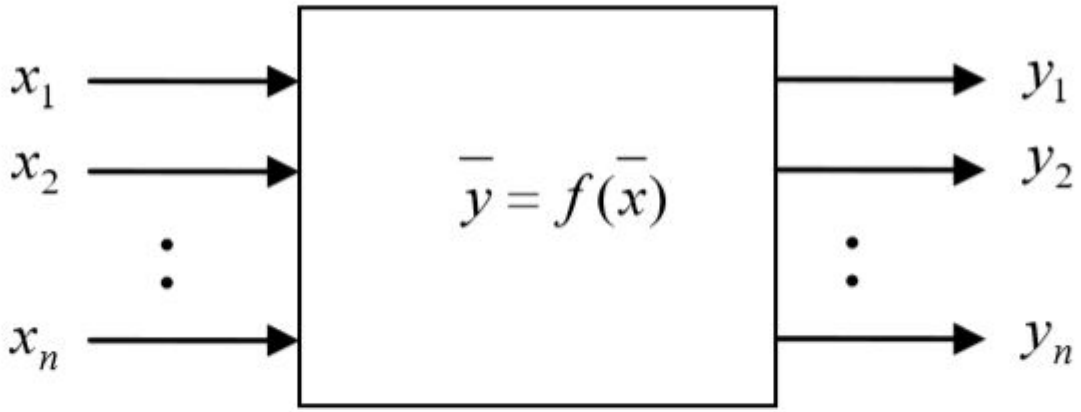
- метод «один против всех» (One vs All);

- нейронная сеть (Neural Network).

Искусственная нейронная сеть (ИНС)

- – математическая модель нервной системы живого организма. Было обнаружено, что свойства ИНС позволяют использовать их для решения широкого круга прикладных задач, в том числе задач классификации.
- Исторически первой была искусственная нейронная сеть под названием «перцептрон Розенблатта» (1957).
- В общем случае ИНС имеет несколько входов и выходов.
- На входы подаются некоторые значения (сигналы).
- Результатом работы нейронной сети являются значения (сигналы) на её выходе

Модель нейронной сети

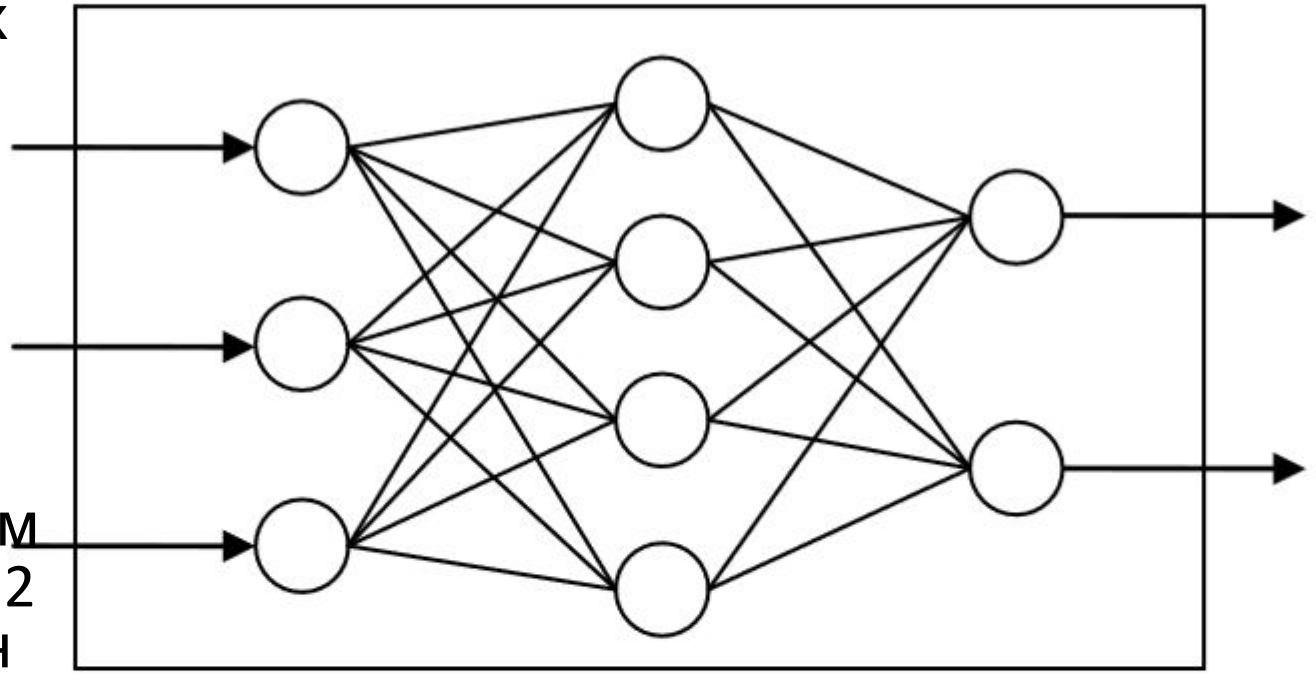


- ИНС можно рассматривать как векторную функцию векторного аргумента:

$$\bar{y} = h(\bar{x}).$$

Структура нейронной сети

- Нейронная сеть состоит из элементов – нейронов, связанных друг с другом
- нейроны объединяются в группы, называемые слоями.
- Различают три вида слоёв: входной, выходной и скрытый.
- На рисунке нейронная сеть, содержащая 3 нейрона во входном слое, 4 нейрона в скрытом слое и 2 нейрона во входном слое. Нейрон является базовым составляющим элементом нейронной сети.

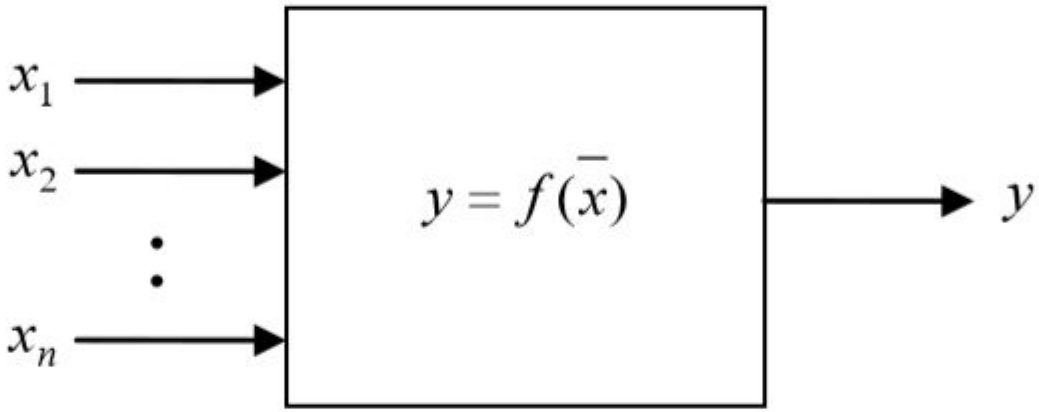


Модель нейрона

- В общем случае нейрон имеет несколько входов и один выход
- Нейрон можно рассматривать ка скалярную функцию векторного аргумента:

$$y = f(\bar{x}).$$

- Предполагается, что каждому входу нейрона соответствует некоторый весовой коэффициент



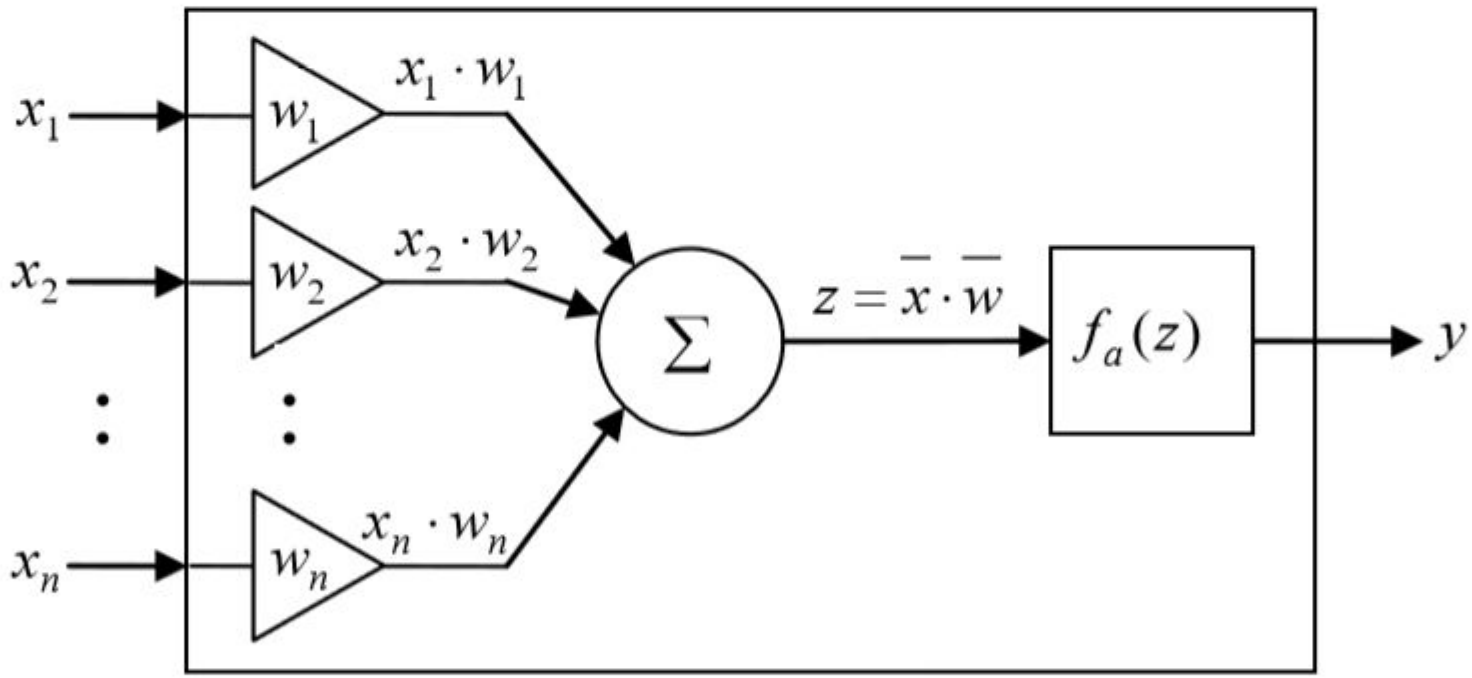
Структура нейрона

- Значения на входе нейрона можно представить в виде вектора

$$\bar{x} = \{x_1, x_2, \ldots, x_n\},$$

- а весовые коэффициенты – в виде вектора

$$\bar{w} = \{w_1, w_2, \ldots, w_n\}.$$



Вычисление значения

- Вычисление значения на выходе нейрона осуществляется в два этапа. На первом этапе рассчитывается взвешенная сум

$$z = x_1 \cdot w_1 + x_2 \cdot w_2 + \ldots + x_n \cdot w_n = \sum_{i=1}^{n} x_i \cdot w_i = \overline{x} \cdot \overline{w}.$$

- На втором этапе рассчитывается значение функции активации . Наиболее часто применяется логистическая (сигмоидная) функци

$f_a(z)$

$$f_a(z) = \frac{1}{1+e^{-z}}.$$

Свойства функции

Свойства функции нейронной сети определяются:

- структурой нейронной сети, то есть характером взаимосвязей между нейронами;

- свойствами нейронов: их весовыми коэффициентами и функциями активации.

Обучение ИНС

- Как и логистическая регрессия, нейронная сеть приобретает свои свойства в результате так называемого «обучения».
- Обучение ИНС – процесс подстройки весовых коэффициентов нейронов ИНС.
- Обучение производится на так называемой «обучающей выборке», представляющей собой набор «вопросов» и соответствующих «правильных ответов».
- Качество обучения определяется степенью соответствия ответов сети («гипотез») «правильным ответам».

Взвешенная сумма квадратов отклонений

- Показателем качества обучения является значение функции штрафа, определяемой взвешенной суммой квадратов отклонений:

$$CF_j = \frac{1}{n}\sum_{i=1}^{n}(h(x_i^{(j)}) - y_i^{(j)})^2 \ ;$$

$$CF = \frac{1}{m}\sum_{j=1}^{m} CF_j \ .$$

Обучение сети

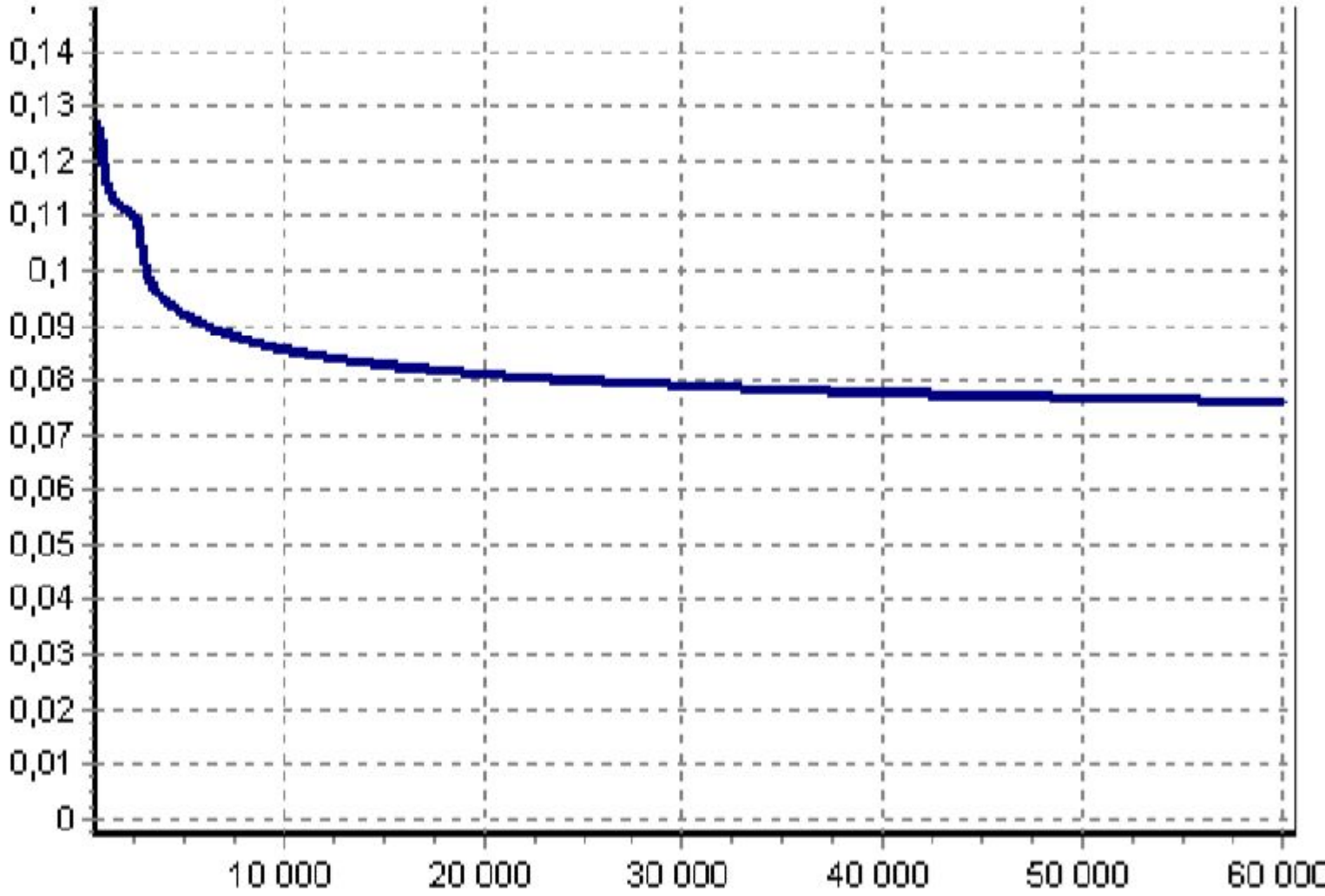
- В процессе обучения весовые коэффициенты нейронов ИНС изменяются согласно определенным правилам.
- Обучение производится шагами (эпохами).
- На одном шаге (в течение одной эпохи) происходит одно обновление коэффициентов *W*.
- Обучение заканчивается в момент, когда значение функции штрафа достигает заданного пользователем порога.
- Также обучение может быть остановлено, если был превышен заданный лимит числа шагов.

Алгоритмы обучение сети

- Обучение сети производится с помощью специальных алгоритмов.
- В основе большинства алгоритмов лежат градиентные методы обучения.
- Исторически первым был так называемый «алгоритм обратного распространения ошибки» (error backpropagation).
- В дальнейшем были предложены еще несколько алгоритмов, наиболее известными из которых являются QPROP и RPROP.
- В ходе обучения возможно проявление двух нежелательных эффектов: эффекта недообученности и эффекта переобученности.

Функция штрафа при недообученности

- Эффект недообученности, как в регрессионном анализе, проявляется в виде недостаточного качества классификации объектов из обучающей выборки.
- Графически это иллюстрируется как приближение функции штрафа к некоему постоянному значению



Избежания эффекта недообученности

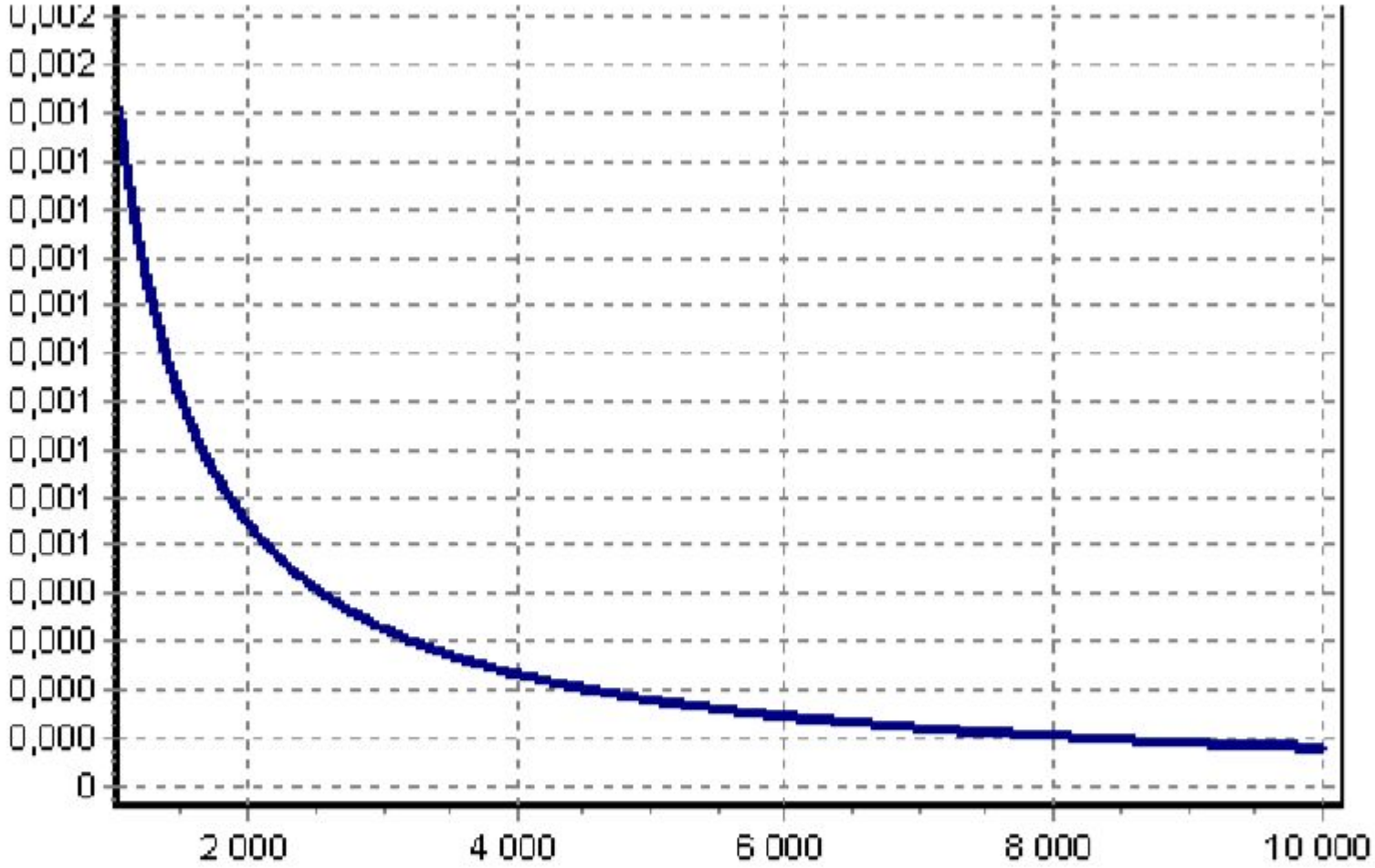
- Для избежания эффекта недообученности можно использовать следующие способы:
  1) увеличение числа нейронов в скрытом слое ИНС;
  2) увеличение числа скрытых слоев.

Эффект переобученности

- Можно выделить три признака переобучения:
- 1) относительно быстрое убывание функции штрафа в процессе обучения;
- 2) нулевое или близкое к нулю значение функции штрафа;
- 3) абсолютно точная при предъявлении объектов из обучающей выборки.

Функция штрафа при переобучении

- Одним из признаков переобученности является нулевое значение функции штрафа после обучения ИНС



Избежания эффекта переобученности

- Переобучение приводит к потере классификатором способности к обобщению.
- Для избежания эффекта переобученности можно использовать следующие способы:
- 1) уменьшение числа нейронов в скрытом слое ИНС;
- 2) уменьшение числа скрытых слоев.

Заключение

В лекции были рассмотрены вопросы классификации

Виды классификации как бинарная и множественная

Также рассмотрены алгоритмы их построения

Источники

- Поручиков, М. А., Анализ данных: учеб. пособие– Самара: Изд-во Самарского университета, 2016. – 88 с.
- Data analysis with Excel. https://www.tutorialspoint.com/excel_data_analysis/excel_data_analysis_tutorial.pdf
- Guerrero H. (2019) Modeling and Simulation: Part 1. In: Excel Data Analysis. Springer, Cham. https://doi.org/10.1007/978-3-030-01279-3_7