

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ
Методические указания к лабораторным работам
(первый семестр)

Составитель: С. М. Наместников

Ульяновск

2022

УДК 621.394.343 (076)

ББК 32.88 я7

ПЗЗ

Рецензент: Deep Learning Engineer компании Huawei, канд. техн. наук, Смирнов П.В.

Одобрено секцией методических пособий научно-методического совета Университета

Методы машинного обучения: методические указания к лабораторным работам (первый семестр) /сост. С. М. Наместников. – Ульяновск : УлГТУ, 2022. – 19 с.

Методические указания по курсу «Методы машинного обучения» для студентов направления 11.04.02 Инфокоммуникационные технологии и системы связи, профиль подготовки " Искусственный интеллект и анализ больших данных в обработке изображений " разработаны в соответствии с программой курса «Методы машинного обучения». Лабораторные работы посвящены исследованию и разработки основных методов машинного обучения с использованием языка Python.

Сборник подготовлен на кафедре «Телекоммуникации».

УДК 621.394.343 (076)

ББК 32.88 я7

© С. М. Наместников, составление, 2022

СОДЕРЖАНИЕ

Лабораторная работа №1

Расчет коэффициентов разделяющей линии и вычисление отступа (margin) для объектов разных классов

Лабораторная работа №2

Обучение линейного алгоритма бинарной классификации образов с помощью градиентного алгоритма

Лабораторная работа №3

Исследование работы L1 и L2-регуляризаторов в задачах регрессии

Лабораторная работа №4

Реализация наивного байесовского классификатора

Лабораторная работа №5

Реализация алгоритма метода опорных векторов для задачи бинарной классификации

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Лабораторная работа №1

Расчет коэффициентов разделяющей линии и вычисление отступа (margin) для объектов разных классов

Цель работы: научиться вычислять коэффициенты разделяющей линии и величину отступа (margin) при бинарной классификации объектов.

Теоретический материал

Теория для выполнения лабораторной работы доступна на следующих страницах сайта:

<https://proproprogs.ru/ml>

в разделах:

- Постановка задачи машинного обучения
- Линейная модель. Понятие переобучения
- Способы оценивания степени переобучения моделей
- Уравнение гиперплоскости в задачах бинарной классификации
- Решение простой задачи бинарной классификации

А также в соответствующих видеоматериалах, размещенных на странице сайта:

<http://tk.ulstu.ru/video.php?id=3>

Задания на лабораторную работу (по вариантам)

1. Используя рисунок своего варианта, необходимо вычислить коэффициенты

$$\omega = [\omega_0, \omega_1, \omega_2]^T$$

разделяющей линии, которая определяется выражением:

$$\omega_1 \cdot x_1 + \omega_2 \cdot x_2 + \omega_0 = 0$$

2. Нормировать полученный вектор коэффициентов так, чтобы его квадратичная норма была равна единице:

$$\|\omega\|^2 = \omega_0^2 + \omega_1^2 + \omega_2^2 = 1$$

3. Вычислить отступы (margin) для зеленых точек (с меткой класса +1) и синих точек (с меткой класса -1). Напомню, что отступ вычисляется по формуле:

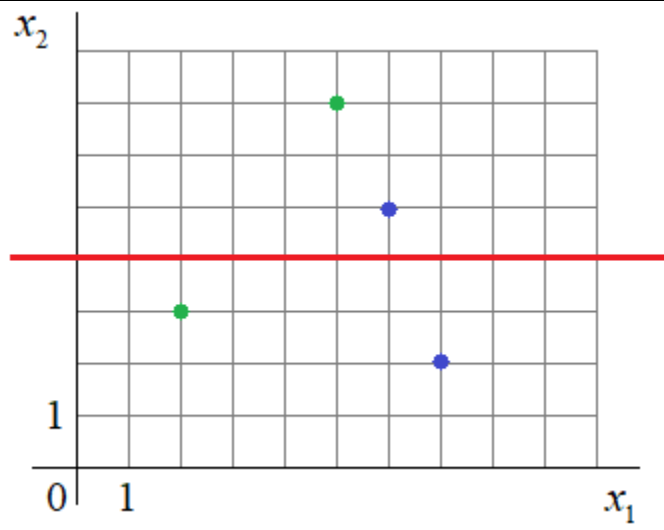
$$M_i = y_i \cdot \langle \omega, x_i \rangle, \quad i=1,2,3,4,$$

где $y_i \in \{-1; +1\}$ - метка класса образа (точки) x_i ; $\langle \omega, x_i \rangle$ - скалярное произведение векторов.

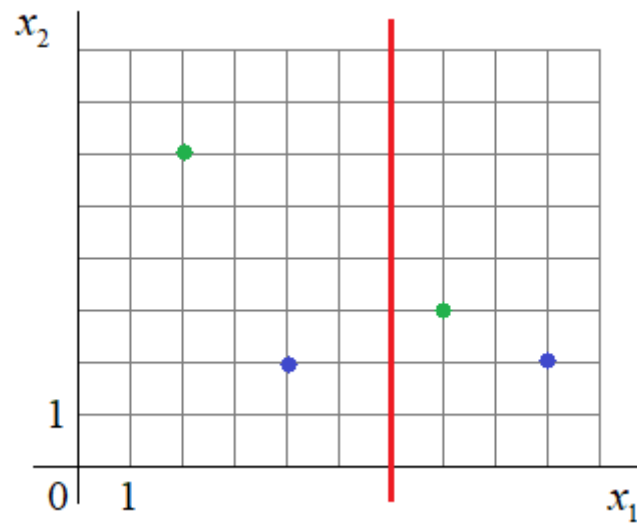
Вектор ω должен быть подобран так, чтобы для наиболее удаленных точек от разделяющей линии отступ был положительным, а для ближних – отрицательным.

Вариант	Графики
1	
2	

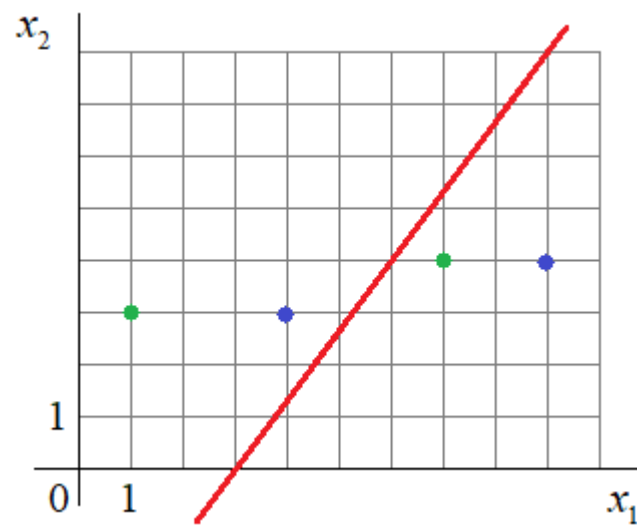
3



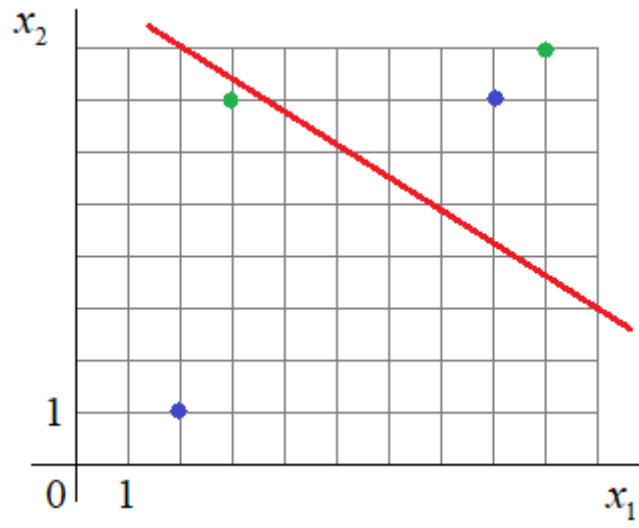
4



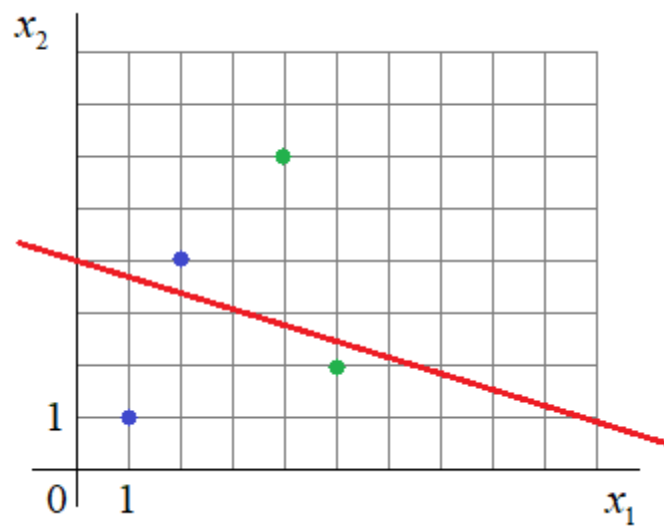
5



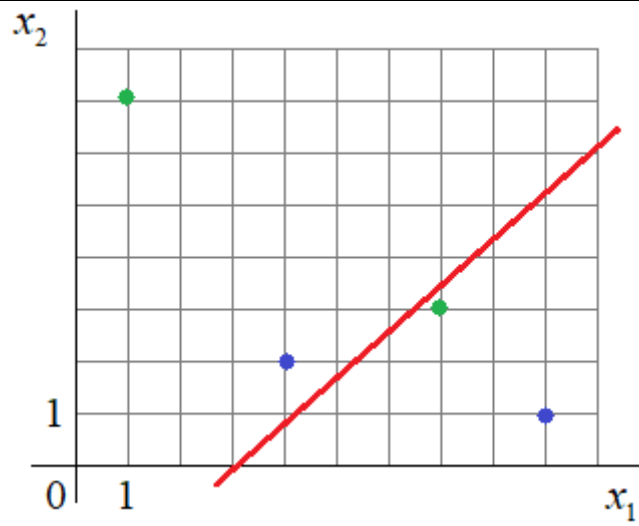
6

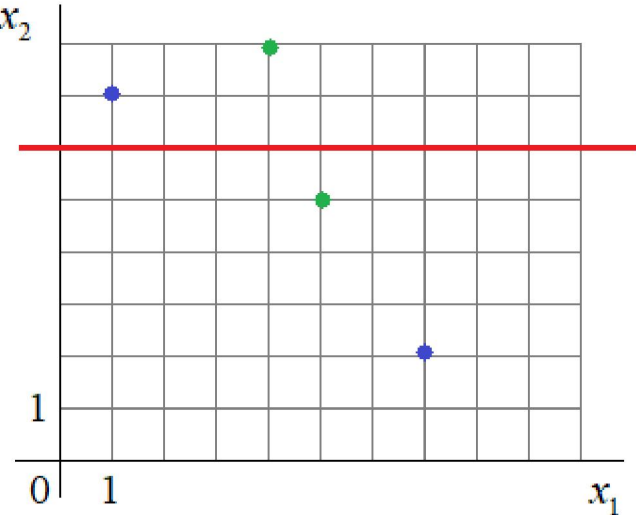
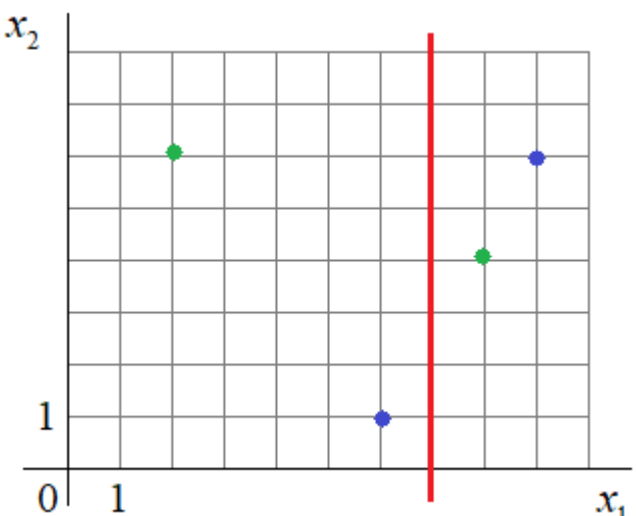


7



8



9	 <p>A scatter plot on a coordinate system with axes x_1 and x_2. The origin is labeled 0, and the x_1 axis has a tick mark at 1. There are five data points: two blue dots and three green dots. A thick red horizontal line is drawn across the plot, separating the points. The blue dots are located at approximately (1, 8) and (7, 4). The green dots are at approximately (3, 9), (4, 6), and (6, 3).</p>
10	 <p>A scatter plot on a coordinate system with axes x_1 and x_2. The origin is labeled 0, and the x_1 axis has a tick mark at 1. There are five data points: two blue dots and three green dots. A thick red vertical line is drawn across the plot, separating the points. The blue dots are located at approximately (4, 1) and (8, 7). The green dots are at approximately (2, 7), (6, 5), and (7, 2).</p>

Содержание отчета

1. Титульный лист с названием лабораторной работы, номером своего варианта, фамилией студента и группы.
2. Расчеты для весов разделяющей линии.
3. Расчеты для отступов.
4. Выводы по полученным результатам.

Лабораторная работа №2

Обучение линейного алгоритма бинарной классификации образов с помощью градиентного алгоритма

Цель работы: научиться реализовывать алгоритм градиентного спуска для задачи обучения линейной модели бинарной классификации образов.

Теоретический материал

Теория для выполнения лабораторной работы доступна на странице сайта:

<https://proproprogs.ru/ml>

в разделах:

- Линейная модель. Понятие переобучения
- Способы оценивания степени переобучения моделей
- Уравнение гиперплоскости в задачах бинарной классификации
- Решение простой задачи бинарной классификации
- Функции потерь в задачах линейной бинарной классификации
- Стохастический градиентный спуск SGD и алгоритм SAG
- Пример использования SGD при бинарной классификации образов

а также в соответствующих видеоматериалах, размещенных на странице сайта:

tk.ulstu.ru/video.php?id=3

Задания на лабораторную работу (по вариантам)

В файле `iris_data.py` даны обучающие выборки (по вариантам) для обучения линейного алгоритма бинарной классификации образов:

http://tk.ulstu.ru/files/iris_data.py

Модель линейного алгоритма должна иметь вид:

$$a(x) = \text{sign}(\langle \omega, x \rangle),$$

где $\omega = [\omega_0, \omega_1, \omega_2]^T$ - весовые коэффициенты модели (определяют ориентацию разделяющей линии); $x = [1, x_1, x_2]^T$ - образ обучающей выборки; $\text{sign}(v) = \begin{cases} -1, & v < 0 \\ +1, & v > 0 \end{cases}$ - знаковая функция.

То есть, метки классов принимают значения $Y = \{-1; +1\}$.

Ваша задача выполнить обучение линейной модели $a(x)$ (найти значения вектора весовых коэффициентов ω) с помощью градиентного алгоритма (программы, написанной на языке Python), который должен минимизировать величину эмпирического риска:

$$Q(X^l) = \sum_{i=1}^l [y_i \neq a(x_i)] \rightarrow \min_{\omega}$$

где $[\cdot]$ - нотация Айзерсона (квадратные скобки возвращают 1, если условие в скобках истинно, и 0 – в противном случае). То есть, эмпирический риск показывает число неверных классификаций.

Так как градиентный алгоритм может минимизировать только гладкие, дифференцируемые функции, то величину $Q(X^l)$ следует сверху ограничить именно таким функционалом:

$$Q(X^l) \leq \tilde{Q}(X^l) = \sum_{i=1}^l L(a(x_i), y_i) \rightarrow \min_{\omega}$$

где $L(a(x_i), y_i) = L(M_i)$ - выбранная функция потерь (здесь $M_i = y_i \cdot \langle \omega, x_i \rangle$ - отступ).

Функция потерь (также, как и набор обучающих данных) определяется вариантом.

Вариант	Функция потерь для реализации градиентного алгоритма
1	$L(M) = \log_2(1 + e^{-M})$ - логарифмическая
2	$Q(M) = (1 - M)^2$ - квадратичная
3	$S(M) = 2 \cdot (1 + e^M)^{-1}$ - сигмоидная
4	$E(M) = e^{-M}$ - экспоненциальная
5	$L(M) = \log_2(1 + e^{-M})$ - логарифмическая
6	$Q(M) = (1 - M)^2$ - квадратичная
7	$S(M) = 2 \cdot (1 + e^M)^{-1}$ - сигмоидная
8	$E(M) = e^{-M}$ - экспоненциальная
9	$L(M) = \log_2(1 + e^{-M})$ - логарифмическая
10	$Q(M) = (1 - M)^2$ - квадратичная

Содержание отчета

1. Титульный лист с названием лабораторной работы, номером своего варианта, фамилией студента и группы.
2. Математические выкладки, необходимые для реализации алгоритма обучения.
3. Текст программы обучения линейной модели с использованием градиентного алгоритма на языке Python.
4. Результаты работы программы в виде графика множества точек обучающей выборки (каждый класс точек должен быть представлен разными маркерами и цветами) и полученной разделяющей линии.
5. Выводы по полученным результатам.

Лабораторная работа №3

Исследование работы L1 и L2-регуляризаторов в задачах регрессии

Цель работы: изучить особенности работы L1 и L2-регуляризаторов на примере задачи аппроксимации функции линейной моделью.

Теоретический материал

Теория для выполнения лабораторной работы доступна на странице сайта:

<https://propoprogs.ru/ml>

в разделах:

- Функции потерь в задачах линейной бинарной классификации
- Стохастический градиентный спуск SGD и алгоритм SAG
- Пример использования SGD при бинарной классификации образов
- Оптимизаторы градиентных алгоритмов: RMSProp, AdaDelta, Adam, Nadam
- L2-регуляризатор. Математическое обоснование и пример работы
- L1-регуляризатор. Отличия между L1- и L2-регуляризаторами

а также в соответствующих видеоматериалах, размещенных на странице сайта:

tk.ulstu.ru/video.php?id=3

Задания на лабораторную работу (по вариантам)

1. Вам необходимо аппроксимировать (описать) функцию своего варианта с помощью линейной модели:

$$a(x) = \omega_0 + \sum_{i=1}^{10} \omega_i x^i,$$

то есть, полиномом 10-й степени. Здесь $\{\omega_i\}$ - весовые коэффициенты, которые требуется найти с помощью градиентного алгоритма по обучающему набору данных.

2. Обучающую выборку следует составить из всех четных индексов сгенерированных значений функции:

$$X^l : \left\{ (x_{2i}, y = f(x_{2i}))_{i=0}^{l // 2} \right\}$$

То есть, сначала формируется первое значение x_0 с целевым значением $y_0 = f(x_0)$, затем, второе: $(x_2, y_2 = f(x_2))$ и так пока не дойдем до конца диапазона.

3. После этого, реализуйте градиентный алгоритм для нахождения весовых коэффициентов вектора ω для квадратической функции потерь (в задачах регрессии, обычно, используют именно такую функцию потерь):

$$Q(X^l) = \frac{1}{2} \sum_{i=1}^l (y_i - a(x_i))^2 \rightarrow \min_{\omega}$$

4. Вычислите прогнозы функции с помощью полученной модели $a(x)$ для всего диапазона значений. (В отчетах, не участвующих в выборке, значения модели должны сильно расходиться с целевыми.)

5. Добавьте в функцию потерь L2-регуляризатор. Повторите процесс обучения и отобразите график для исходной функции и результата ее восстановления с помощью полученной модели $a(x)$.

6. Вместо L2-регуляризатора запишите L1-регуляризатор. Повторите процесс обучения и отобразите график для исходной функции и результата ее восстановления с помощью полученной модели $a(x)$.

P.S. Все программы реализовать на языке Python с использованием пакетов NumPy и Matplotlib.

Вариант	Функция для исследования L1 и L2-регуляризаторов.
1	$y(x) = \frac{1}{10 + x^3}, x \in [0; 20; 0,1]$
2	$y(x) = \frac{\sin(10x)}{x + 0,1}, x \in [0; 3; 0,01]$
3	$y(x) = \sin(5x) \cdot \cos(x), x \in [0; 3; 0,01]$
4	$y(x) = \sin(x) \cdot x^2, x \in [0; 20; 0,01]$
5	$y(x) = \sin(5x) \cdot e^{-x}, x \in [0; 10; 0,01]$
6	$y(x) = \frac{1}{10 + x^2}, x \in [0; 40; 0,1]$
7	$y(x) = \frac{\cos(20x)}{x + 0,1}, x \in [0; 4; 0,01]$
8	$y(x) = \cos(x) \cdot x^2, x \in [0; 7; 0,01]$
9	$y(x) = \cos(5x) \cdot e^x, x \in [0; 6; 0,01]$
10	$y(x) = \ln(x) \cdot e^{-x}, x \in [1; 10; 0,01]$

Содержание отчета

1. Титульный лист с названием лабораторной работы, номером своего варианта, фамилией студента и группы.
2. Математические выкладки для реализации алгоритмов.
3. Тексты программ с результатами их работы.
4. Выводы по полученным результатам.

Лабораторная работа №4

Реализация наивного байесовского классификатора

Цель работы: научиться строить наивный байесовский классификатор и с его помощью выполнять бинарную классификацию образов.

Теоретический материал

Теория для выполнения лабораторной работы доступна на странице сайта:

<https://proproprogs.ru/ml>

в разделах:

- Логистическая регрессия. Вероятностный взгляд на машинное обучение
- Вероятностный взгляд на L1 и L2-регуляризаторы
- Формула Байеса при решении конкретных задач
- Байесовский вывод. Наивная байесовская классификация
- Гауссовский байесовский классификатор
- Линейный дискриминант Фишера

а также в соответствующих видеоматериалах, размещенных на странице сайта:

tk.ulstu.ru/video.php?id=3

Задания на лабораторную работу (по вариантам)

1. Необходимо построить (реализовать на языке Python) наивный байесовский классификатор на основе, следующих данных обучающей выборки (для своего варианта):

http://tk.ulstu.ru/files/iris_data.py

Полагать, что признаки независимы и распределены по гауссовскому закону (нормальной плотности распределения вероятностей).

2. Для данной обучающей выборки подсчитать число и процент неверных классификаций.

3. Отобразить обучающую выборку в виде графика точек на плоскости (объекты разных классов должны быть иметь разные маркеры и цвет).

Содержание отчета

1. Титульный лист с названием лабораторной работы, номером своего варианта, фамилией студента и группы.
2. Все расчеты, связанные с построением наивного байесовского классификатора.
3. Программа, реализующая наивный байесовский классификатор.
4. Графики и результаты работы программы.
5. Выводы по полученным результатам.

Лабораторная работа №5

Реализация алгоритма метода опорных векторов для задачи бинарной классификации

Цель работы: реализовать метод опорных векторов (SVM) для задачи бинарной классификации.

Теоретический материал

Теория для выполнения лабораторной работы доступна на странице сайта:

<https://proproprogs.ru/ml>

в разделах:

- Введение в метод опорных векторов (SVM)
- Реализация метода опорных векторов (SVM)
- Метод опорных векторов (SVM) с нелинейными ядрами

а также в соответствующих видеоматериалах, размещенных на странице сайта:

tk.ulstu.ru/video.php?id=3

Задания на лабораторную работу (по вариантам)

1. Необходимо построить (реализовать на языке Python с применением пакета Scikit-Learn) линейный вариант метода опорных векторов, для следующих данных обучающей выборки (для своего варианта):

http://tk.ulstu.ru/files/iris_data.py

2. Для данной обучающей выборки подсчитать число и процент неверных классификаций.

3. Отобразить обучающую выборку в виде графика точек на плоскости (объекты разных классов должны быть иметь разные маркеры и цвет), а также полученную (в результате обучения) разделяющую линию.

Содержание отчета

1. Титульный лист с названием лабораторной работы, номером своего варианта, фамилией студента и группы.
2. Программа, реализующая метод опорных векторов.

3. Графики и результаты работы программы.
4. Выводы по полученным результатам.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Николенко С., Кадури́н А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.
2. Рашид, Тарик. Создаем нейронную сеть.: Пер. с англ. — СПб.: ООО «Альфа-книга», 2017. — 272 с.: ил.
3. Хайкин, Саймон. Нейронные сети: полный курс, 2-е издание.: Пер. с англ. — М.: Издательский дом «Вильямс», 2006. — 1104 с.: ил.
4. Васильев К.К., Оптимальная обработка сигналов в дискретном времени: Учебн. пособие. — М.: Радиотехника, 2016. — 288 с.: ил.
5. Christopher M. Bishop, Neural Networks for Pattern Recognition. — Clarendon Press Oxford, 1995 - 498 с.
6. Harrison Kinsley, Neural Networks from Scratch in Python - 666 с.