

**МИНОБРНАУКИ РОССИИ**  
**Ярославский государственный университет им. П.Г. Демидова**

Кафедра цифровых технологий и машинного обучения

**УТВЕРЖДАЮ**

Декан физического факультета  
  
\_\_\_\_\_  
(подпись) И.С. Огнев

«23» мая 2023 г.

**Рабочая программа дисциплины**  
**«Обработка больших данных и системы искусственного**  
**интеллекта»**

Направление подготовки  
11.04.02 Инфокоммуникационные технологии и системы связи

Направленность (профиль)  
Сети, системы и устройства телекоммуникаций

Форма обучения  
очная

Программа одобрена  
на заседании кафедры  
от «17» апреля 2023 года, протокол № 8

Программа одобрена НМК  
физического факультета  
протокол № 5 от «25» апреля 2023 года

Ярославль

## 1. Цели освоения дисциплины

Целью освоения дисциплины **«Обработка больших данных и системы искусственного»** является изучение студентами эффективных алгоритмов современных систем искусственного интеллекта, включая методы машинного и глубокого обучения, а также получение опыта их практического применения.

В процессе преподавания дисциплины решаются следующие задачи:

- ознакомление с методами обучения с учителем;
- ознакомление с методами обучения без учителя;
- изучение алгоритмов глубокого обучения;
- практическое использование алгоритмов машинного обучения.

## 2. Место дисциплины в структуре ОП бакалавриата

Данная дисциплина относится к дисциплинам части, формируемой участниками образовательных отношений.

Она основывается на знаниях, полученных при изучении дисциплин математического, естественнонаучного цикла и цикла профессиональных дисциплин: «Теория вероятностей и математическая статистика» и «Дискретная математика».

Знания и навыки, полученные при изучении данной дисциплины, будут востребованы при изучении дисциплины «Сети связи», ряда дисциплин по выбору, при выполнении курсовых и выпускных квалификационных работ, а также в последующей трудовой деятельности обучающихся.

Следует отметить стремительную динамику постоянного совершенствования систем на базе методов искусственного интеллекта, что требует от процесса преподавания постоянной доработки и переработки соответствующих разделов.

## 3. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины направлен на формирование следующих элементов компетенций в соответствии с ФГОС ВО, ОП ВО и приобретения следующих знаний, умений, навыков и (или) опыта деятельности:

Формируемая компетенция (код и формулировка)	Индикатор достижения компетенции (код и формулировка)	Перечень планируемых результатов обучения
<b>Профессиональные компетенции</b>		
ПК-1 Способен осуществлять сбор и обработку исходных данных для решения поставленных профессиональных задач в области радиофизики, осуществлять поиск, анализ и выбор методов их решения	ИД ПК-1.2 Проводит анализ и обоснованный выбор методов решения профессиональных задач в области радиофизики	<b>Знать:</b> <ul style="list-style-type: none"><li>– методы машинного обучения с учителем и без учителя;</li><li>- подходы, позволяющие выполнять настройку параметров алгоритмов машинного обучения;</li><li>- примеры практических задач, в которых возможно эффективное использование методов машинного обучения.</li></ul> <b>Уметь:</b> <ul style="list-style-type: none"><li>– реализовывать алгоритмы машинного обучения с использованием средств</li></ul>

		<p>компьютерного моделирования;</p> <ul style="list-style-type: none"> <li>- выполнять настройку параметров алгоритмов машинного обучения;</li> <li>- анализировать правильность работы алгоритмов машинного обучения.</li> </ul> <p><b>Владеть навыками:</b> построения, анализа и компьютерного моделирования систем машинного обучения.</p>
--	--	--

#### 4. Объем, структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 3 зачетных единицы, 108 акад. часов.

##### Очное отделение

№ п/п	Темы (разделы) дисциплины, их содержание	Семестр	Виды учебных занятий, включая самостоятельную работу студентов, и их трудоемкость (в академических часах)						Формы текущего контроля успеваемости Форма промежуточной аттестации (по семестрам)
			Контактная работа						
			лекции	практические	лабораторные	консультации	аттестационные испытания	самостоятельная работа	
1	Введение и обзор материала курса	2	1					1,7	Тестирование, задание для самостоятельной работы
2	Линейная регрессия с одной переменной. Линейная регрессия со множеством переменных. Классификация. Логистическая регрессия	2	2		11			1	Тестирование, задание для самостоятельной работы
3	Искусственные нейронные сети (представление)	2	2		6	1		1	Тестирование, задание для самостоятельной работы
4	Искусственные нейронные сети (обучение)	2	2		7	1		1	Тестирование, задание для самостоятельной работы
5	Рекомендации по применению алгоритмов машинного обучения. Построение систем машинного обучения. Оптическое распознавание символов. Формирование базы данных	2	1					1	Тестирование, задание для самостоятельной работы
6	Кластеризация	2	2		5			1	Тестирование, задание

									для самостоятельной работы
7	Анализ главных компонент	2	2		5			1	Тестирование, задание для самостоятельной работы
8	Детектирование лиц на основе алгоритма Виола/Джонса	2	1					1	Тестирование, задание для самостоятельной работы
9	Машинное обучение на больших базах данных	7	2					1	Тестирование, задание для самостоятельной работы
10	Глубокое обучение. Свёрточные нейронные сети	2	2			2		11	Тестирование, задание для самостоятельной работы
	<b>Всего</b>		<b>17</b>		<b>34</b>	<b>4</b>		<b>20,7</b>	
		2					0,3	32	Экзамен
	<b>Всего с зачетом</b>		<b>17</b>		<b>34</b>	<b>4</b>	<b>0,3</b>	<b>52,7</b>	

## Содержание разделов дисциплины

### Тема № 1

#### Введение и обзор материала курса

Что такое машинное обучение? Примеры задач, решаемые в области машинного обучения (обучение с учителем и без учителя, стимулированное обучение, эволюционное обучение).

### Тема № 2

Линейная регрессия с одной переменной. Линейная регрессия со множеством переменных. Классификация. Логистическая регрессия

Общая постановка задачи регрессии. Что такое гипотеза, параметры модели и стоимостная функция на примере задачи линейной регрессии? Минимизация стоимостной функции (нормальные уравнения и численная оптимизация). Метод градиентного спуска. Масштабирование признаков и настройка скорости обучения алгоритма.

Общая постановка задачи классификации. Логистическая регрессия и ее стоимостная функция. Граница принятия решения. Многоклассовая классификация на основе логистической регрессии (подходы «один против всех» и «один против одного»).

### Тема № 3

#### Искусственные нейронные сети (представление)

Проблема нелинейной классификации. Биологические нейроны и мозг. Модель нейрона и искусственные нейронные сети (ИНС). Нейронные сети прямого распространения. Функции активации ИНС. Реализации логических операций на основе ИНС. Классификация рукописных цифр. Классификация объектов в сложных сценах. Многоклассовая классификация для ИНС.

### Тема № 4

#### Искусственные нейронные сети (обучение)

Регуляризация и проблема переобучения. Регуляризованная стоимостная функция для линейной и логистической регрессии, нейронной сети прямого распространения.

Алгоритм обратного распространения ошибки. Градиентная проверка. Проблема симметричности весов ИНС.

### **Тема № 5**

Рекомендации по применению алгоритмов машинного обучения. Построение систем машинного обучения. Оптическое распознавание символов. Формирование базы данных.

### **Тема № 6**

#### **Кластеризация**

Что такое кластеризация? Разновидности алгоритмов кластеризации данных. Алгоритм К-средних (стоимостная функция, случайная инициализация, выбор числа кластеров). Примеры использования алгоритма К-средних для обработки цифровых изображений.

### **Тема № 7**

#### **Анализ главных компонент**

Задача сокращения размерности данных. Общая формулировка анализа главных компонент и его реализация на практике. Выбор числа главных компонент. Рекомендации по применению анализа главных компонент. Распознавание лиц с использованием анализа главных компонент (пространство лиц, собственные лица, структура алгоритма распознавания лиц).

### **Тема № 8**

#### **Детектирование лиц на основе алгоритма Виола/Джонса**

Общее описание проблемы детектирования лиц. Признаки, используемые в алгоритме детектирования лиц Виола/Джонса. Каскад классификаторов. Интегральные изображения. Слабые классификаторы. Комитет классификаторов и бустинг. Данные для обучения каскада классификаторов и тестирование алгоритма Виола/Джонса.

### **Тема № 9**

#### **Машинное обучение на больших базах данных**

Общая постановка задачи обучения на больших базах данных. Стохастический градиентный спуск и минигрупповой градиентный спуск. Онлайн-обучение. Технология MapReduce и распараллеливание данных.

### **Тема № 10**

#### **Глубокое обучение. Свёрточные нейронные сети**

Проблема глубокого обучения. Архитектура свёрточной нейронной сети (СНС) (свёрточный слой, слой прореживания, полносвязные слои). Функции активации СНС. Основные типы архитектур СНС. Обучение свёрточной нейронной сети. Аугментация данных. Предварительное обучение и автоэнкодеры. Примеры использования СНС для решения различных практических задач.

### **Список лабораторных работ**

1. Линейная регрессия.
2. Логистическая регрессия.
3. Многоклассовая классификация и нейронные сети.
4. Обучение нейронных сетей.
5. Регуляризованная линейная регрессия. Недообучение и переобучение.
6. Кластеризация с использованием алгоритма K-средних.
7. Анализ главных компонент.

## **5. Образовательные технологии, в том числе технологии электронного обучения и дистанционные образовательные технологии, используемые при осуществлении образовательного процесса по дисциплине**

В процессе обучения соответствующей дисциплине используются следующие образовательные технологии:

**Вводная лекция** – дает первое целостное представление о дисциплине и ориентирует студента в системе изучения данной дисциплины. Студенты знакомятся с назначением и задачами курса, его ролью и местом в системе учебных дисциплин и в системе подготовки в целом. Дается краткий обзор курса, история развития науки и практики, достижения в этой сфере, имена известных ученых, излагаются перспективные направления исследований. На этой лекции высказываются методические и организационные особенности работы в рамках данной дисциплины, а также дается анализ рекомендуемой учебно-методической литературы.

**Академическая лекция** (или лекция общего курса) – последовательное изложение материала, осуществляемое преимущественно в виде монолога преподавателя. Требования к академической лекции: современный научный уровень и насыщенная информативность, убедительная аргументация, доступная и понятная речь, четкая структура и логика, наличие ярких примеров, научных доказательств, обоснований, фактов.

**Лабораторная работа** – организация учебной работы с реальными материальными и информационными объектами, экспериментальная работа с аналоговыми моделями реальных объектов.

**Консультация** – занятие перед проведением экзамена, на котором проводится консультирование по изученному материалу, формам заданий итогового контроля, ответы на вопросы студентов по дисциплине.

## **6. Перечень лицензионного и (или) свободно распространяемого программного обеспечения, используемого при осуществлении образовательного процесса по дисциплине**

В процессе осуществления образовательного процесса по дисциплине используются:

для формирования материалов для текущего контроля успеваемости и проведения промежуточной аттестации, для формирования методических материалов по дисциплине:

- программы Microsoft Office;
- Adobe Acrobat Reader.

## **7. Перечень современных профессиональных баз данных и информационных справочных систем, используемых при осуществлении образовательного процесса по дисциплине (при необходимости)**

В процессе осуществления образовательного процесса по дисциплине используются:

Автоматизированная библиотечно-информационная система «БУКИ-NEXT»  
[http://www.lib.uniyar.ac.ru/opac/bk\\_cat\\_find.php](http://www.lib.uniyar.ac.ru/opac/bk_cat_find.php)

**8. Перечень основной и дополнительной учебной литературы, ресурсов информационно-телекоммуникационной сети «Интернет» (при необходимости), рекомендуемых для освоения дисциплины**

**а) основная литература**

1. Гелиг А.Х., Матвеев А.С. Введение в математическую теорию обучаемых распознающих систем и нейронных сетей: учебное пособие. – СПб.: Изд-во С.-Петерб. ун-та, 2014.

**б) дополнительная литература**

2. Местецкий Л.М. Математические методы распознавания образов: курс лекций. – М.: Национальный Открытый Университет «ИНТУИТ», 2008.

**в) ресурсы сети «Интернет»:**

Электронная библиотека учебных материалов ЯрГУ  
([http://www.lib.uniyar.ac.ru/opac/bk\\_cat\\_find.php](http://www.lib.uniyar.ac.ru/opac/bk_cat_find.php)).

**9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине**

Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине включает в свой состав специальные помещения:

- учебные аудитории для проведения занятий лекционного типа;
- учебные аудитории для проведения лабораторных работ;
- учебные аудитории для проведения групповых и индивидуальных консультаций,
- учебные аудитории для проведения текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы;
- помещения для хранения и профилактического обслуживания технических средств обучения.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду организации.

Число посадочных мест в аудитории для занятий лекционного типа больше либо равно списочному составу группы обучающихся.

Автор:

Доцент кафедры инфокоммуникаций  
и радиопизики, к.т.н.

\_\_\_\_\_ В.В. Хрящев

**Приложение №1 к рабочей программе дисциплины  
«Обработка больших данных и системы искусственного»**

**Фонд оценочных средств  
для проведения текущей и промежуточной аттестации студентов  
по дисциплине**

**1. Типовые контрольные задания или иные материалы,  
необходимые для оценки знаний, умений, навыков и (или) опыта деятельности,  
характеризующих этапы формирования компетенций**

**1.1. Контрольные задания и иные материалы,  
используемые в процессе текущей аттестации**

**Задания для самостоятельной работы**

1. Придумайте пример анализа статистических данных с использованием линейной регрессии с одной переменной.
2. Придумайте пример анализа статистических данных с использованием линейной регрессии со множеством переменных.
3. Придумайте пример анализа статистических данных с использованием логистической регрессии.
4. Придумайте пример анализа статистических данных с использованием подхода «один против всех» на основе логистической регрессии, а также нейронной сети прямого распространения.
5. Разработайте структурную схему алгоритма обучения нейронной сети прямого распространения.
6. Придумайте пример анализа статистических данных с использованием регуляризованной линейной регрессии при изучении проблемы недообучения моделей.
7. Придумайте пример анализа статистических данных с использованием регуляризованной линейной регрессии при изучении проблемы переобучения моделей.
8. Чему равно значение регуляризованной стоимостной функции для случая, когда все параметры модели, а также параметр регуляризации равны единице?
9. Чему равны значения параметров обученной модели регуляризованной линейной регрессии для случая, когда параметр сходимости равен 0.05, число итераций градиентного спуска равно 1500, а параметр регуляризации равен 0?
10. Чему равно значение наилучшего параметра регуляризации из набора чисел 0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10 для обученной модели полиномиальной регрессии с максимальной степенью полиномиального признака равной 8, когда параметр сходимости равен 0.05, а число итераций градиентного спуска равно 1500?
11. Чему равны значения параметров обученной модели полиномиальной регрессии из пункта 3 для наилучшего подобранного параметра регуляризации?
12. Придумайте пример анализа статистических данных с использованием алгоритма К-средних.
13. Придумайте пример обработки статистических данных с использованием анализа главных компонент.



## Тест по темам № 1 и № 2

### Введение и обзор материала курса. Линейная регрессия с одной переменной

1. Говорят, что компьютерная программа способна к обучению из опыта  $E$  по отношению к некоторой задаче  $T$  и критерию качества работы  $R$ , если ее производительность на  $T$ , как мера  $R$ , улучшается с опытом  $E$ . Допустим, что алгоритм машинного обучения обучен (на множестве исторических данных о погоде) предсказывать погоду. Чем в этом случае является  $E$ ?

- а) Среди ответов нет правильного.
- б) Вероятность правильно предсказывать погоду в будущем.
- в) Процедура исследования алгоритмом большого количества исторических данных о погоде.
- г) Задача предсказания погоды.

2. Допустим нам необходимо предсказать, будет или нет дождь завтра в 17.00. Для этой цели мы желаем использовать алгоритм машинного обучения. Как в данном случае будет трактоваться задача машинного обучения?

- а) Задача классификации.
- б) Задача регрессии.

3. Допустим, что мы решаем задачу прогнозирования на фондовой бирже. Нам необходимо предсказать будет ли доллар США выше по отношению к рублю завтра. Для этой цели мы желаем использовать алгоритм машинного обучения. Как в данном случае будет трактоваться задача машинного обучения?

- а) Задача классификации.
- б) Задача регрессии.

4. Какое из следующих определений является разумным определением машинного обучения?

- а) Машинное обучение является наукой программирования компьютеров.
- б) Машинное обучение – это область исследования, которая предоставляет компьютеру, не будучи явно запрограммированному, возможность к обучению.
- в) Машинное обучение означает извлечение смысла из размеченных данных.
- г) Машинное обучение является областью, позволяющей роботу действовать интеллектуально.

5. К какому классу алгоритмов машинного обучения может быть отнесен алгоритм К-средних?

- а) Обучение без учителя.
- б) Обучение с учителем.

6. Рассмотрим линейную регрессию с одной переменной применительно к задаче предсказания цены на недвижимость. Чему равно количество  $m$  тренировочных примеров в таблице, представленной ниже.

Площадь (фут <sup>2</sup> ) – $x$	Цена в 1000-х (\$) – $y$
2104	460
1416	232
1534	315
852	178
950	191

7. Вычислить значение стоимостной функции  $J(Q_0, Q_1)$  для случая линейной регрессии с одной переменной, если обучающее множество выглядит так, как представлено в таблице ниже. Положить  $Q_0 = 0$ ,  $Q_1 = 1$ .

х (свойство)	у (метка)
3	4
2	1
4	3
0	1

8. Допустим  $Q_0 = 0$ ,  $Q_1 = 1.5$ . Чему равняется  $h_Q(2)$  для случая линейной регрессии?

9. Является ли стоимостная функция для линейной регрессии с одной переменной выпуклой?

- а) Да.
- б) Нет.

10. Выберите правильные варианты, которые описывают особенности градиентного спуска.

- а) Если параметр сходимости маленький, то градиентный спуск может быть медленным.
- б) Шаг градиентного спуска на каждой итерации является фиксированным.
- в) Если параметр сходимости большой, то градиентный спуск может проскочить минимум. Алгоритм может не сходиться или даже расходиться.

### Тест по темам № 2 и № 3

Линейная регрессия со множеством переменных. Классификация. Логистическая регрессия.  
Искусственные нейронные сети (представление)

1. Алгоритм градиентного спуска запущен для 15 итераций и  $\alpha = 0.3$  в задаче поиска параметров линейной регрессии со множеством переменных. Значения  $J(Q)$  вычислены для каждой итерации. Обнаружено, что  $J(Q)$  медленно уменьшается и продолжает уменьшаться после 15 итераций. Какое из следующих заключений является правдоподобным?

- а) Для лучшей работы алгоритма необходимо выбрать  $\alpha = 0.1$ , то есть уменьшить.
- б) Для лучшей работы алгоритма необходимо выбрать  $\alpha = 1$ , то есть увеличить.
- в) Изначально выбранное  $\alpha = 0.3$  является эффективным выбором скорости обучения.

2. По какой из следующих причин используется масштабирование признаков в задаче линейной регрессии со множеством переменных?

- а) Масштабирование признаков позволяет ускорить градиентный спуск, уменьшив число итераций для получения хорошего решения.
- б) Масштабирование признаков позволяет избежать застревания в локальном минимуме при использовании нормальных уравнений.
- в) Масштабирование признаков позволяет ускорить градиентный спуск, уменьшая вычислительную стоимость каждой итерации.

3. Назовите основные отличия использования алгоритма градиентного спуска от нормальных уравнений в задаче линейной регрессии со множеством переменных.

- а) Необходим выбор  $\alpha$ .

- б) Нет необходимости выбирать  $\alpha$ .
- в) Необходимо много итераций.
- г) Нет необходимости в итерациях.
- д) Медленно работает, если число признаков большое.

4. Подходит ли алгоритм машинного обучения на основе линейной регрессии со множеством переменных для решения задачи классификации?

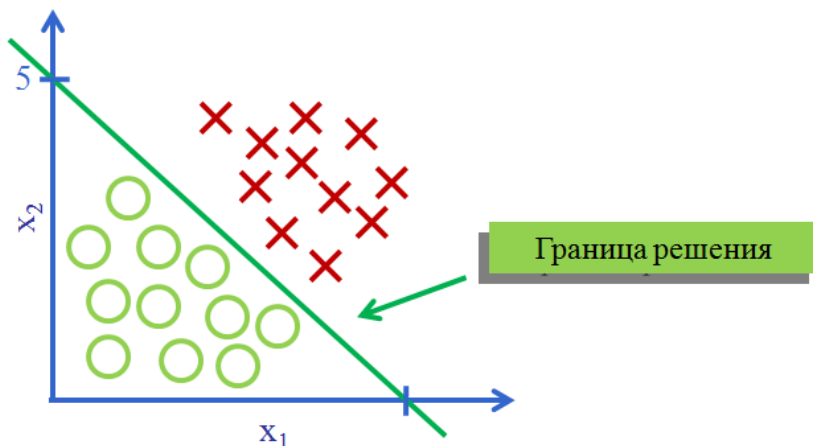
- а) Да.
- б) Нет.

5. Допустим, что обучен классификатор на основе логистической регрессии, который предсказывает значение  $h_Q(x) = 0.2$  для нового примера  $x$ . Что это означает?

- а) Вероятность  $P(y = 0|x; Q) = 0.2$ .
- б) Вероятность  $P(y = 0|x; Q) = 0.8$ .
- в) Вероятность  $P(y = 1|x; Q) = 0.8$ .
- г) Вероятность  $P(y = 1|x; Q) = 0.2$ .

6. Как выглядит математическое выражение, описывающее процедуру обновления параметров в алгоритме градиентного спуска для логистической регрессии?

7. Запишите уравнение, описывающее границу принятия решения для разделения двух классов на рисунке, представленном ниже.



8. Сколько классификаторов необходимо построить при решении задачи классификации для пяти классов с использованием логистической регрессии и подхода «один против всех»?

9. Изобразите нейронную сеть, которая позволит реализовать операцию логического И. Укажите конкретные значения коэффициентов.

10. Какие из следующих логических операций нельзя реализовать с использованием одного сигмоидного нейрона?

- а) И.
- б) ИЛИ.
- в) НЕ.
- г) Исключающее ИЛИ.

### Тест по темам № 4 и № 5

Искусственные нейронные сети (обучение). Рекомендации по применению алгоритмов машинного обучения. Построение систем машинного обучения. Оптическое распознавание символов. Формирование базы данных

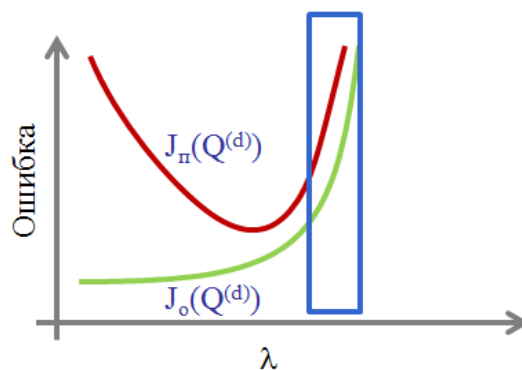
1. Как называется алгоритм, используемый для обучения нейронных сетей прямого распространения?

2. Для какой цели в машинном обучении используется процедура градиентной проверки?

- а) Борьба с переобучением нейронной сети.
- б) Обучение нейронной сети.
- в) Проверка правильности работы реализованного алгоритма обратного распространения ошибки.
- г) Борьба с недообучением нейронной сети.

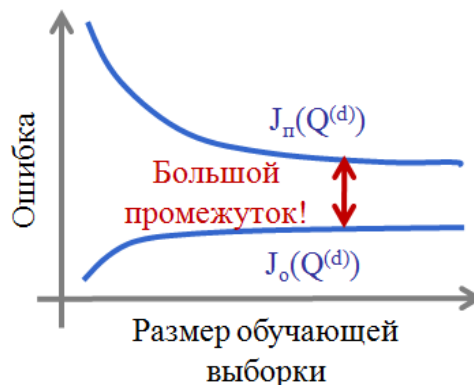
3. В чем проблема алгоритма машинного обучения при выбранных параметрах регуляризации в области, отмеченной на рисунке?

- а) Недообучение.
- б) Переобучение.



4. После обучения классификатора получены следующие кривые обучения для обучающего и проверочного множеств. В чем проблема обученного алгоритма?

- а) Недообучение.
- б) Переобучение.



5. Какие действия можно предпринять в случае переобучения алгоритма на базе логистической регрессии?

- а) Получить больше тренировочных примеров.

- б) Попытаться уменьшить число свойств.
- в) Попытаться получить дополнительные свойства.
- г) Попытаться уменьшить параметр регуляризации.
- д) Попытаться увеличить параметр регуляризации.

6. При решении задачи бинарной классификации для 1000 тестовых примеров получены результаты, представленные в таблице ниже. Чему равны точность (Precision) и полнота выборки (Recall) классификатора?

		Действительный класс	
		1	0
Предсказанный класс	1	85	890
	0	15	10

7. Допустим, обучен классификатор на базе логистической регрессии. При этом порог классификатора находится в точке 0.5. Что произойдет с классификатором, если этот порог будет увеличен до 0.7?

- а) Классификатор будет вероятно иметь более высокую точность (Precision).
- б) Классификатор будет вероятно иметь более высокую полноту выборки (Recall).
- в) Точность и полнота выборки вероятно не изменятся.

8. На основе какого подхода в лекционном материале выполнялось детектирование областей текста (пешеходов) на цифровых изображениях?

9. При анализе производительности конвейерной системы распознавания символов получены следующие результаты:



Компонент	Точность
Вся система	70%
Детектирование текста	72%
Сегментация символов	82%
Распознавание символов	100%

Какие из следующих утверждений верны?

а) Если мы приходим к выводу о том, что ошибки работы системы возникают из-за переобучения системы распознавания, то, возможно, стоит приложить значительные усилия при получении дополнительных примеров для обучения системы распознавания.

б) Не имеет смысла улучшать систему распознавания, так как она работает со 100% точностью.

в) Не имеет смысла прикладывать большие усилия для модернизации системы детектирования текста.

г) Требуется приложить значительные усилия при сборе доп. данных для обучения системы детектирования текста.

10. Отметьте плюсы, которыми обладает анализ производительности конвейерной системы?

а) Предоставление информации о качестве работы отдельных блоков всей системы и понимание того, какие подсистемы стоит улучшать, а какие – нет.

б) Понимание о недообучении и переобучении системы.

в) Понимание того, какой алгоритм машинного обучения является наиболее подходящим для обеспечения работоспособности отдельных блоков конвейера.

### Тест по темам № 6 и № 7

#### Кластеризация. Анализ главных компонент

1. Что такое кластеризация?

а) Операция, направленная на восстановление исходных данных из зашумленных.

б) Процесс разбиения множества векторов признаков на подмножества.

в) Операция, направленная на разбиение цифрового изображения на неперекрывающиеся области, покрывающие все изображение и однородные по некоторому признаку.

2. Для решения какой из следующих задач подходит алгоритм К-средних?

а) Даны записи о продаже продуктов в магазине. Необходимо принять решение о том, какие продукты часто покупаются одновременно и должны находиться на одной полке в магазине.

б) Дано множество новостных статей, полученных с разных веб-сайтов. Определить группы статей относящихся к одной общей теме.

в) Даны исторические записи о погоде. Предсказать объем осадков завтра (выход должен быть непрерывной величиной).

г) Даны исторические записи о погоде. Предсказать какой будет погода завтра (солнечной или дождливой).

3. Предположим, что существует три средних значения ( $m_1 = [1, 2]^T$ ,  $m_2 = [-3, 0]^T$ ,  $m_3 = [4, 2]^T$ ), вычисленных на одном из этапов алгоритма К-средних. Пусть тренировочный пример  $x^{(i)} = [3, 1]^T$ . К какому кластеру будет отнесен пример  $x^{(i)}$  после выполнения этапа определения принадлежности примера к некоторой группе?

4. Сколько основных параметров необходимо задать в алгоритме К-средних?

а) 1.

б) 2.

в) 3.

г) 4.

5. Как называется алгоритм кластеризации, в котором центры кластеров соответствуют пикам распределения данных.

а) Метод К-средних.

б) Метод графового разбиения Ши.

в) Рекурсивный гистограммный метод Оландера.

г) Метод сдвига среднего.

6. Для каких целей, как правило, используется анализ главных компонент?

- а) Сжатие данных.
- б) Визуализация данных.
- в) Борьба с переобучением алгоритмов машинного обучения.
- г) Ускорение работы классификатора, например, на базе логистической регрессии.

7. Анализ главных компонент и линейная регрессия – это одно и то же?

- а) Да.
- б) Нет.

8. Каким свойством удовлетворяет базис главных компонент?

- а) Ортогональность.
- б) Базисные вектора являются собственными векторами ковариационной матрицы данных.
- в) Представление данных в базисе главных компонент является оптимальным в смысле среднеквадратической ошибки.
- г) Все ответы неверны.

9. Как называются изображения лиц, формирующие базис главных компонент, рассчитанный для базы данных лиц?

10. Как называется алгоритм классификации данных, в котором неизвестный вектор признаков относится к тому классу, к отдельному эталонному образцу которого этот вектор наиболее близок.

- а) Логистическая регрессия.
- б) Нейронная сеть прямого распространения.
- в) Метод ближайшего соседа.
- г) Метод К-ближайших соседей.

## 1.2. Список вопросов и (или) заданий для проведения промежуточной аттестации

### Список вопросов к экзамену

1. Что такое искусственный интеллект, машинное обучение, глубокое обучение?
2. Примеры задач, решаемые в области машинного обучения (обучение с учителем и без учителя, стимулированное обучение, эволюционное обучение).
3. Общая постановка задачи регрессии.
4. Что такое гипотеза, параметры модели и стоимостная функция на примере задачи линейной регрессии?
5. Минимизация стоимостной функции (нормальные уравнения и численная оптимизация).
6. Метод градиентного спуска.
7. Масштабирование признаков и настройка скорости обучения алгоритма.
8. Общая постановка задачи классификации.
9. Логистическая регрессия и ее стоимостная функция.
10. Граница принятия решения.
11. Многоклассовая (мультиклассовая) классификация на основе логистической регрессии (подходы «один против всех» и «один против одного»).
12. Проблема нелинейной классификации.
13. Биологические нейроны и мозг.
14. Модель нейрона и искусственные нейронные сети.
15. Нейронные сети прямого распространения.

16. Примеры реализации логических операций на основе искусственных нейронных сетей.
17. Классификация рукописных цифр.
18. Классификация объектов в сложных сценах.
19. Многоклассовая классификация для ИНС.
20. Регуляризация и проблема переобучения.
21. Регуляризованная стоимостная функция для линейной и логистической регрессии, ИНС прямого распространения.
22. Алгоритм обратного распространения ошибки.
23. Градиентная проверка.
24. Проблема симметричности весов ИНС.
25. Отладка алгоритмов машинного обучения.
26. Что такое диагностика?
27. Оценка работоспособности гипотезы.
28. Обучающее, проверочное и тестовое множества данных.
29. Алгоритм выбора модели.
30. Подбор параметра регуляризации.
31. Кривые обучения.
32. Анализ ошибок.
33. Метрики ошибки для ассиметричных классов.
34. Машинное обучение в задаче оптического распознавания символов (детектирование текста, сегментация символов, классификация символов).
35. Что такое скользящее окно?
36. Формирование большого количества данных для решения задачи машинного обучения.
37. Анализ производительности конвейерной системы.
38. Что такое кластеризация?
39. Разновидности алгоритмов кластеризации данных.
40. Алгоритм К-средних (стоимостная функция, случайная инициализация, выбор числа кластеров).
41. Примеры использования алгоритма К-средних для обработки цифровых изображений.
42. Задача сокращения размерности данных.
43. Общая формулировка анализа главных компонент и его реализация на практике.
44. Выбор числа главных компонент.
45. Рекомендации по применению анализа главных компонент.
46. Распознавание лиц с использованием анализа главных компонент (пространство лиц, собственные лица, структура алгоритма распознавания лиц).
47. Общее описание проблемы детектирования лиц.
48. Признаки, используемые в алгоритме детектирования лиц Виола/Джонса.
49. Каскад классификаторов.
50. Интегральные изображения.
51. Слабые классификаторы.
52. Комитет классификаторов и бустинг.
53. Данные для обучения каскада классификаторов и тестирование алгоритма Виола/Джонса.
54. Общая постановка задачи обучения на больших базах данных.
55. Стохастический градиентный спуск и минигрупповой градиентный спуск.
56. Онлайн-обучение.
57. Технология MapReduce и распараллеливание данных.
58. Проблема глубокого обучения.



59. Архитектура свёрточной нейронной сети (свёрточный слой, слой прореживания, полносвязные слои).
60. Обучение свёрточной нейронной сети.
61. Предварительное обучение и автоэнкодеры.
62. Примеры использования свёрточных нейронных сетей для решения различных практических задач.

### Критерии оценивания ответов на вопросы билета

Критерий	Пороговый уровень (на «удовлетворительно»)	Продвинутый уровень (на «хорошо»)	Высокий уровень (на «отлично»)
<b>Соответствие ответа вопросу</b>	Хотя бы частичное (не относящееся к вопросу не подлежит проверке)	Полное	Полное
<b>Наличие примеров</b>	Имеются отдельные примеры	Много примеров	Есть практически ко всем утверждениям
<b>Содержание ответа</b>	Понятийные вопросы изложены с классификациями, проблемные с постановкой проблемы и изложением различных точек зрения. Имеются ошибки или пробелы.	Ответ почти полный, без ошибок, не хватает отдельных элементов и тонкостей	Исчерпывающий полный ответ

## 2. Описание процедуры выставления оценки

Изучение дисциплины заканчивается зачетом. Для подготовки ответа на вопрос билета отводится не менее 40 минут.

Оценка «зачтено» выставляется, если ответ на вопрос билета дан не ниже, чем на пороговом уровне.

Оценка «не зачтено» выставляется, если ответ на вопрос билета дан ниже, чем на пороговом уровне.

## **Приложение №2 к рабочей программе дисциплины «Обработка больших данных и системы искусственного»**

### **Методические указания для студентов по освоению дисциплины**

Одной из основных форм усвоения учебного материала по дисциплине **«Обработка больших данных и системы искусственного»** является самостоятельная работа студента, причем в достаточно большом объеме. По всем темам предусмотрены задания самостоятельной работы, на которых происходит закрепление изученного материала и отработка навыков анализа и синтеза систем на базе методов искусственного интеллекта.

Изучение дисциплины заканчивается зачетом. Оценка выставляется на основании уровня сформированности указанных компетенций, который оценивается как средняя оценка по совокупности параметров: оценки за самостоятельные задания и ответы на вопросы билета.

Освоить вопросы дисциплины **«Обработка больших данных и системы искусственного»** самостоятельно студенту достаточно сложно. Посещение всех предусмотренных лекционных занятий и занятий по выполнению лабораторных работ является совершенно необходимым. Без упорных и регулярных самостоятельных занятий в течение семестра сдать зачет практически невозможно.