

МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
Ярославский государственный университет им. П.Г.Демидова

УЧЕБНО-МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ
ПО ДИСЦИПЛИНЕ
МАШИННОЕ ОБУЧЕНИЕ

Направление подготовки (специальность):
02.04.02 Фундаментальная информатика и информационные технологии

Образовательная программа
Искусственный интеллект и компьютерные науки

очная форма обучения

Составитель:
ЛЕВАНОВА О. А., К.Ф.-М.Н.,
ДОЦЕНТ Ф-ТА ИВТ ЯРГУ ИМ. П.Г. ДЕМИДОВА

г. Ярославль

Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

Основная литература:

1. Тюгашев, А. А. Компьютерные средства искусственного интеллекта : учебное пособие / А. А. Тюгашев. — Самара : Самарский государственный технический университет, ЭБС АСВ, 2020. — 270 с. — ISBN 978-5-7964-2293-9. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/105021.html> (дата обращения: 08.10.2021). — Режим доступа: для авторизир. Пользователей

Дополнительная литература:

2. Маккинли, Уэс Python и анализ данных / Уэс Маккинли ; перевод А. Слинкина. — 2-е изд. — Саратов : Профобразование, 2019. — 482 с. — ISBN 978-5-4488-0046-7. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/88752.html> (дата обращения: 07.10.2021). — Режим доступа: для авторизир. Пользователей
3. Рафаэл, Гонсалес Цифровая обработка изображений / Гонсалес Рафаэл, Вудс Ричард ; перевод Л. И. Рубанов, П. А. Чочиа ; под редакцией П. А. Чочиа. — Москва : Техносфера, 2012. — 1104 с. — ISBN 978-5-94836-331-8. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: <https://www.iprbookshop.ru/26905.html> (дата обращения: 08.10.2021). — Режим доступа: для авторизир. пользователей
4. Шапиро, Л. Компьютерное зрение : учебное пособие / Л. Шапиро, Д. Стокман ; под редакцией С. М. Соколова ; перевод с английского А. А. Богуславского. — 4-е изд. — Москва : Лаборатория знаний, 2020. — 763 с. — ISBN 978-5-00101-696-0. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/135496> (дата обращения: 07.10.2021). — Режим доступа: для авториз. пользователей.
5. Ян, Э. С. Программирование компьютерного зрения на языке Python / Э. С. Ян ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2016. — 312 с. — ISBN 978-5-97060-200-3. — Текст : электронный // Лань : электронно-библиотечная система. —

URL: <https://e.lanbook.com/book/93569> (дата обращения: 07.10.2021). — Режим доступа: для авториз. пользователей.

6. Клетте, Р. Компьютерное зрение. Теория и алгоритмы : учебник / Р. Клетте ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 506 с. — ISBN 978-5-97060-702-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/131691> (дата обращения: 07.10.2021). — Режим доступа: для авториз. пользователей.
7. Кэлэр, А. Изучаем OpenCV 3. Разработка программ компьютерного зрения на C++ с применением библиотеки OpenCV / А. Кэлэр, Г. Брэдски ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2017. — 826 с. — ISBN 978-5-97060-471-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: <https://e.lanbook.com/book/108126> (дата обращения: 07.10.2021). — Режим доступа: для авториз. пользователей.

Учебно-методическое обеспечение самостоятельной работы обучающихся по дисциплине (модулю)

- 1 Электронный университет Moodle ЯрГУ URL: <https://moodle.uniyar.ac.ru/>
- 2 Единое окно доступа к образовательным ресурсам URL: <http://window.edu.ru/>

Перечень информационных технологий, используемых при изучении дисциплины, включая программное обеспечение

- В процессе осуществления образовательного процесса используются:
- для формирования текстов материалов для промежуточной и текущей аттестации – программы Microsoft Office, издательская система LaTeX;
 - для поиска учебной литературы библиотеки ЯрГУ – Автоматизированная библиотечная информационная система "БУКИ-NEXT" (АБИС "Буки-Next");

**Учебно-методические указания и рекомендации
к изучению тем лекционных и практических занятий,
самостоятельной работе студентов**

Содержание дисциплины

Наименование дисциплины (модуля) с указанием разделов (элементов) /наименование раздела дисциплины	Название темы с кратким содержанием
1 Машинное обучение	<p>Введение в машинное обучение. Классификация задач машинного обучения: обучение с учителем, обучение без учителя, рекомендательные системы. Основные термины: объекты, признаки, метки (ответы), решающее правило, обобщающая способность, переобучение. Постановки задач классификации и регрессии. Виды признаков (бинарные, числовые, порядковые, номинальные, количественные). Примеры задач. Современное состояние области, успешные проекты. Связь с другими областями, в частности компьютерным зрением.</p>
2	<p>Язык Python введение. Основные характеристики языка, синтаксис. Базовые типы данных и операции с ними. Преобразования типов. Условные операторы (ветвления). Циклы в Python (for и while). Функция range. Структуры данных: списки, кортежи, словари и множества. Функции, модули и библиотеки, импорт.</p>
3	<p>Знакомство с библиотекой Numpy. Сравнение numpy.array и list. Срезы. Логическая индексация. Особенности операций +, *, **. Сравнение range, arrange, xrange. Генерация случайных чисел из разных распределений. Знакомство с библиотекой Matplotlib. Рисование графиков plot, subplot, подписи к осям, название, легенда, задание цвета и вида графика. Диаграммы: bar, hist, bxpplot, pie. Рисование тепловой карты imshow.</p>
4	<p>Знакомство с библиотекой Pandas. Тип данных Series: объявление, индексация, обращение к элементам, отбор по условию. Пропуски и их заполнению (fillna). Тип данных DataFrame. Индексация, названия столбцов. Обращение к элементам (at, loc, iloc). Добавление/удаление элементов. Проверка на пропуски, удаление, заполнение. Нахождение и удаление дубликатов. Переиндексация. Методы apply, sort, groupby.</p>

Наименование дисциплины	Название темы с кратким содержанием
5	<p>Первичный анализ данных (библиотека sklearn). Этапы анализа данных.</p> <p>Чтение csv файла, информация о наборе данных (info, describe, head/tail). Подходы к заполнению пропусков. Преобразование номинальных признаков (Label Encoding, One-Hot Encoding, Hashing). Нормировка числовых признаков (StandardScaling, MinMax Scaling). Работа с текстовыми признаками (map, мешок слов, N-граммы, TF-IDF).</p> <p>Как правильно оценивать качество модели (отложенная выборка, кросс-валидация, LOO). Метрики качества (accuracy, точность, полнота, F1-мера).</p>
6	<p>Метрические методы. Гипотеза компактности и непрерывности. Меры близости, обобщенная метрика Минковского. Пример задачи классификации цветков ириса. Обобщенный метрический классификатор. Метод kNN. Окно Парзена. Задача регрессии, метод наименьших квадратов. Непараметрическая регрессия, ядерное сглаживание Надарая-Ватсона. Виды функции ядра. Алгоритм LOWESS.</p> <p>Применение методов библиотеки sklearn: train_test_split, cross_val_score, accuracy_score.</p> <p>Применение метрических методов с использованием библиотеки sklearn. Методы KNeighborsClassifier, GridSearchCV, Pipeline (нормировка признаков и kNN).</p>
7	<p>Линейные методы. Линейная регрессия. Метрики качества регрессии. Градиентный спуск и способы оценивания градиента. Переобучение и регуляризация. Свойства L1, L2 регуляризации. Методы библиотеки sklearn: LinearRegression, Lasso, Ridge, ElasticNet.</p> <p>Линейная классификация. Логистическая регрессия и оценки вероятности классов.</p> <p>Многоклассовая классификация, сведение к бинарным задачам. Многоклассовая логистическая регрессия.</p> <p>Подходы к вычислению метрик качества macro micro усреднение. Понятие PR, ROC-кривая, метрики AUC-PR, AUC-ROC.</p>
8	<p>Решающие деревья. Жадный алгоритм построения. Выбор лучшего разбиения с помощью критерия информативности. Критерии информативности для регрессии и классификации. Учёт пропусков в деревьях. Решающие деревья и категориальные признаки. Усечение дерева. Деревья для задачи регрессии. Небрежные решающие деревья.</p>
9	<p>Методы понижения размерности. Линейные методы, метод главных компонент (PCA). Нелинейные методы MDS (многомерное шкалирование), SNE, Tsne.</p>

Наименование дисциплины	Название темы с кратким содержанием
10	Обучение без учителя. Кластеризация. Метрики близости. Алгоритмы: k средних, fuzzy k-means, иерархические методы. Нелинейные методы кластеризации DB-SCAN. Обсуждение метрик качества кластеризации (бизнес-метрики).
11	Рекомендательные системы. Подходы content-base, collaborative filtering, user-based, item-base, на основе SVD разложения. Использование регуляризации. Оценка качества рекомендаций
Итого, 1й семестр:	
11	Введение в обработку изображений. Цветовые пространства: RGB, CMY, CIE Lab, HSB/HSL/HSI. Понятие гистограммы яркости. Фильтрация изображений. Понятие свертка, подходы к вычислению. Фильтр Гаусса, медианный фильтр (нелинейный). Повышение резкости. Знакомство с библиотекой OpenCV. Общее описание что содержит библиотека, какие языки и платформы поддерживает. Модули. Пример простого приложения с загрузкой и отображением картинки. Практическое использование изученных понятий.
12	Нахождение границ. Понятие градиента изображения. Методы вычисления, операторы Робертса, Превитта, Собеля, сила и направление градиента. Детектор Кэнни выделение границ. Гистограмма градиентов (HOG).
13	Простые методы детекции объектов. Бинаризация изображений. Морфологические операции сужение, расширение. Морфологическая фильтрация, нахождение границы, скелета. Алгоритм вычисления компонент связности (объектов).
14	Вычисление признаков из изображения. Цветовые признаки, гистограмма яркости HOG как признаки. Геометрические признаки объектов (площадь, периметр, вытянутость, компактность, моменты). Сопоставление контуров, Distance Transform для быстрого вычисления расстояния между контурами. Текстурный анализ, идея банка фильтров (фильтры разных направлений).

Наименование дисциплины	Название темы с кратким содержанием
16	<p>Метод опорных векторов (SVM). Линейно разделяемая выборка. Постановка задачи регрессии, аппроксимация функции для задачи классификации. Случай линейной отделимости, постановка задачи оптимизации и ее решение. Понятия опорного вектора, опорные нарушители. Нелинейное обобщение, ядерный трюк. Виды ядер. SVM-регрессия.</p> <p>Практическое применение метода SVM для классификации и регрессии (библиотека sklearn). Подбор параметра C. Как влияет выбор ядра на форму разделяющей поверхности, тестирование разных ядер.</p> <p>Детекция пешеходов. Идея скользящего окна. Вычисление признаков для фрагмента изображения (HOG). Применение метода SVM для классификации пешеход/НЕ пешеход. Улучшения: использование блоков, улучшение выборки при обучении за счет использования сложных примеров с предыдущего шага. Проблема дисбаланса классов, выбор правильной метрики качества.</p>
18	<p>Композиция алгоритмов. Разные подходы: взвешенное голосование, Bagging, Boosting, Staking. Случайный лес. Бустинг над деревьями. XGBoost.</p> <p>Детектор лиц Виоло-Джонса. Идея слабых классификаторов. Понятие интегральной матрицы. Алгоритм AdaBoost. Каскад классификаторов, требования к их качеству.</p>
20	<p>Знакомство в нейронными сетями.</p> <p>Математическая модель нейрона. Идея суммирования входов и применение функции активации. Виды функций активации. Многослойный перцептрон. Проблема исключаящего или. Возможности двухслойного перцептрона. Понятие функции потерь, сведение к задаче оптимизации, обратное распространение ошибки, градиентный спуск.</p> <p>Сверточные НС. Сверточный слой, слой прореживания, функция softmax. Современные архитектуры СНС.</p> <p>Примеры успешно решенных задач: распознавание рукописных цифр, набор ImageNet, распознавание лиц.</p>

Пример теста

- Чему будет равен корень из среднеквадратичной ошибки для набора из 3 наблюдений, где отклонение предсказания линейной регрессии от реальных значений равны: -1, 2, -2?
 - 2
 - 3
 - 0
 - 1.
- Рассмотрим признак “Образовательная программа” при анализе данных по студентам университета. Этот признак может принимать три значения: “Экономика”, “Математика”, “Философия”. Воспользуемся one-hot кодированием и заменим этот признак на три бинарных, которые будут соответствовать категориям

в том порядке, в котором они перечислены выше. Как будет закодирован признак со значением “Философия”?

- а) (1, 0, 0)
 - б) (1, 1, 0)
 - в) (1, 1, 1)
 - г) (0, 1, 0)
 - д) (1, 0, 1)
 - е) (0, 0, 1)
 - ж) (0, 1, 1).
3. Модель линейной регрессии выглядит так: $a(x) = w_0 + w_1 x_1 + \dots + w_d x_d$. Сколько у неё параметров?
- а) d ;
 - б) $d+1$;
 - в) 1;
 - г) $d-1$.
4. Предположим, что мы строим модель предсказания роста по возрасту и весу человека. Модель с какими коэффициентами вероятнее всего переобучилась?
- а) $0.001 * (\text{возраст}) + 0.5 * (\text{вес})$;
 - б) $0.1 * (\text{возраст}) + 0.33 (\text{вес})$;
 - в) $1402325.3 * (\text{возраст}) + -1404370.5 (\text{вес})$.
5. Предположим, что мы строим модель предсказания стоимости дома по количеству комнат и средней цене дома в районе. Перед количеством комнат коэффициент равен 1400230, а перед средней ценой дома в районе 0.8. Можно ли утверждать, что количество комнат — более важный признак для качества предсказания, чем средняя цена в районе и почему?
- а) Нет, так как коэффициенты несравнимы, поскольку признаки имеют разный масштаб;
 - б) Нет, так как средняя цена дома в районе — это признак с большим разбросом, а именно разброс характеризует ценность признака;
 - в) Да, так как количество комнат — это признак, который может принимать небольшое количество значений, а значит, каждое значение содержит в себе больше информации;
 - г) Да, так как коэффициент перед количеством комнат больше.
6. Какая из моделей приводит к отбору признаков?
- а) Линейная регрессия без регуляризации
 - б) Ridge-регрессия;
 - в) Lasso-регрессия;
 - г) ElasticNet.

Правильные ответы

Вопрос №	Правильный ответ	Вопрос №	Правильный ответ
1	б	4	в
2	е	5	а
3	б	6	в

Критерии оценки

- «Отлично» – 6 правильных ответов;
- «Хорошо» – 5 правильных ответов;
- «Удовлетворительно» – 4 правильных ответов;

- «Неудовлетворительно» – 3 и менее правильных ответов.

Примеры лабораторных работ

Написать программу с использованием библиотек, которая решает следующую задачу:

1. Бинаризируйте изображение robo3.jpg, чтобы выделить красные кружки, выбор порога бинаризации должен быть автоматизирован, выбор цветового пространства остается за автором. Проверьте работу Вашего алгоритма на изображениях robo2.jpg, robo1.jpg.
2. Выделите внутреннюю и внешнюю границы на изображении binar1.jpg binar2.jpg с помощью морфологической обработки. Поясните результаты.
3. Выделите объекты на изображении binar1.jpg, вычислите площадь каждого объекта.
4. Продемонстрируйте промежуточные результаты работы алгоритма.
5. Выделите объекты на изображении Clusters.jpg. На выходе должно получиться изображение, на котором разные объекты помечены разными цветами, а фон черным.
6. Для изображения circles.jpg реализуйте морфологический алгоритм для построения трех изображений, которые бы содержали соответственно: 1) только частицы, касающиеся краев изображения; 2) только группы перекрывающихся частиц; 3) только одиночные круглые частицы.

Задания для группы из нескольких человек

7. Выделите в изображениях table1.jpg, table2.bmp, table3.jpg границы таблицы с использованием морфологических операций. Результатом обработки должно быть изображение, в котором удален весь текст и оставлены только границы линий. Какие были проблемы, какие есть пути их решения, какие выбрали вы.
8. Предложите и реализуйте алгоритм нахождения не до конца заполненных бутылок на изображении FigP1126.tif. Бутылка считается заполненной до конца, если уровень выше середины между началом сужения и низом горлышка.
9. На изображениях eye_.bmp выделите (постройте бинарное изображение) зрачек/радужку. Алгоритм должен работать приемлемо для всех (4 изображений). Какие были проблемы, какие есть пути их решения, какие выбрали вы. Вычислите характеристики выделенного объекта (площадь, периметр, вытянутость, компактность).
10. Реализовать алгоритм выделения на изображении объемлющего прямоугольника, в котором располагается штрих-код.

Студент защищает лабораторную работу, при этом готовит презентацию с результатами работы программы на разных этапах (подходах к решению). Студент должен быть готов ответить на вопросы по коду программы и пояснить выбор значений параметров в функциях, которые использует. А также понимать алгоритм, который используется в той или иной, сторонней функции.

Показатели	Критерии
Содержание программы	Анализирует изученный материал, Правильно выбирает путь решения задачи и использует подходящие алгоритмы, Программа работает корректно.

Аргументированно отвечает на вопросы	Знает изложенный материал, Проявляет критическое мышление
Представление лабораторной	Использует иллюстративные, наглядные материалы, Владеет культурой речи.

Критерии оценки

- «Отлично» – программа правильно решает поставленную задачу, лабораторная работа полностью соответствует описанным критериям;
- «Хорошо» – программа правильно решает поставленную задачу, лабораторная работа соответствует описанным критериям за исключением некоторых замечаний не более чем по нескольким пунктам критериев;
- «Удовлетворительно» – программа правильно решает поставленную задачу, лабораторная работа соответствует более чем половине описанных критериев;
- «Неудовлетворительно» – программа не предоставлена, или не верно решает поставленную задачу, лабораторная работа не соответствует большей части описанных критериев.

Вариант билета на зачете

1. Опишите метод «Решающие деревья», перечислите его достоинства и недостатки..
2. Что такое переобучение/недообучение? Как проводить анализ того переобучена или недообучена модель?
3. Опишите метод построения рекомендательной системы на основе SVD разложения.
4. При решении задачи бинарной классификации для 1000 тестовых примеров получены результаты, представленные в таблице ниже. Чему равны точность (Precision), полнота (Recall) и F-мера классификатора? Что показывают эти величины, когда какую лучше стоит использовать?

	Действительный класс	
	1	0
1	85	890
0	15	10

Критерии оценки

- «Отлично» – даны верные ответы на 4 вопросов из билета;
- «Хорошо» – даны верные ответы на 4 вопроса из билета, в некоторых вопросах были допущены существенные недочеты;
- «Удовлетворительно» – даны верные ответы на 3 вопроса из билета;
- «Неудовлетворительно» – даны верные ответы на 2 и менее вопросов из билета.

Вариант билета на экзамене

1. Опишите основные шаги детектора границ Canny.
2. Какие подходы к композиции алгоритмов Вы знаете?

3. Опишите алгоритм SVM. Перечислите плюсы минусы метода, за что отвечают основные параметры алгоритма?
4. Понятие свертки, свойства, что может вычислять свертка. Понятие сверточного слоя, карта свертки, зачем используют? Сколько параметров содержит сверточный слой со сверткой 5×5 для изображения $128 \times 128 \times 3$?

Практическая часть:

Задан набор данных в виде csv-файла. Напишите программу, которая обучает модель (алгоритм) задачу классификации. Оцените качество работы модели. Набор данных содержит информацию о клиентах сотовой связи (тариф, число минут разговора в месяц, звонков в кол-центр оператора, регион...), необходимо предсказать уйдет ли клиент в течении месяца.

Критерии оценки

- «Отлично» – практическая часть полностью реализована, даны верные ответы на 4 вопроса из билета, возможно допущены небольшие огрехи;
- «Хорошо» – практическая часть реализована почти полностью, даны верные ответы на 4 вопроса из билета, в некоторых вопросах были допущены существенные недочеты;
- «Удовлетворительно» – практическая часть реализована частично, даны верные ответы на 3 вопроса из билета;
- «Неудовлетворительно» – практическая часть не реализована или даны верные ответы лишь на 2 и менее вопросов из билета.