МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Астраханский государственный университет имени В. Н. Татищева»
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО Руководитель ОПОП А.В. Григорьев «5» мая 2025 г.	УТВЕРЖДАЮ И.о. заведующего кафедрой информационных технологий О.Н. Выборнова «5» мая 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

«Машинное обучение»

Составитель(и)	**Выборнова О.Н., доцент, к.т.н., доцент кафедры ИТ;** **Соболевский В.В., ст. преподаватель кафедры ИТ;**
Направление подготовки / специальность	**09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА**
Направленность (профиль) / специализация ОПОП	**Прикладная информатика в социальных науках**
Квалификация (степень)	**бакалавр**
Форма обучения	**очная**
Год приёма	**2023**
Курс	**3**
Семестр(ы)	**6**

Астрахань – 2025

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

1.1. Целями освоения дисциплины (модуля) «Машинное обучение» являются формирование у студентов практических навыков применения инструментальных средств при решении профессиональных задач в области машинного обучения

1.2. Задачи освоения дисциплины (модуля):

– сформировать теоретические знания по основам машинного обучения для построения формальных математических моделей и интерпретации результатов моделирования;

– выработать умения по практическому применению методов машинного обучения для построения формальных математических моделей и интерпретации результатов моделирования при решении прикладных задач в различных прикладных областях;

– выработать умения и навыки использования различных программных инструментов анализа баз данных и систем машинного обучения.

2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

2.1. Учебная дисциплина (модуль) «Машинное обучение» относится к части, формируемой участниками образовательных отношений, и осваивается в 6 семестре.

2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения, навыки, формируемые предшествующими учебными дисциплинами (модулями):

– математические основы информационных технологий и вычислительной техники

– системы искусственного интеллекта

– основы программирования

Знания: линейной алгебры, дифференциального исчисления и теории вероятностей, язык программирования Python, численные методы, принципы векторизации

Умения: применять математический аппарат, применять знания для подготовки обучающего набора данных, составлять программы на языке программирования Python

Навыки: использования технологии векторизации для обработки векторных и матричных данных, математического мышления, определения алгоритмов и структур данных при изучении существующих моделей машинного обучения

2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения, навыки, формируемые данной учебной дисциплиной (модулем):

– практикум по обработке данных

– бакалаврская работа

3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс освоения дисциплины (модуля) направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки / специальности:

б) профессиональной (ПК): ПК-6 Способен обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций

Таблица 1. Декомпозиция результатов обучения

Код и наименование компетенции	Планируемые результаты обучения по дисциплине (модулю)		
	Знать (1)	Уметь (2)	Владеть (3)
ПК-6 Способен обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций	ПК-6.1. Знает методы обработки данных для подготовки аналитических решений	ПК-6.2. Умеет использовать соответствующие методы обработки данных для подготовки аналитических решений	ПК-6.3. Владеет обработкой данных для подготовки аналитических решений

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Объем дисциплины (модуля) составляет 3 зачетные единицы, в том числе 68 час, выделенных на контактную работу обучающихся с преподавателем (из них 17 часов – лекции, 51 часов – лабораторные работы), и 40 часов – на самостоятельную работу обучающихся:

Таблица 2.2. Структура и содержание дисциплины (модуля)

Раздел, тема дисциплины (модуля)	Семестр	Контактная работа (в часах)			Самост. работа		Формы текущего контроля успеваемости, форма промежуточной аттестации
		Л	ПЗ	ЛР	КР	СР	
Тема 1. Введение в машинное обучение.	6	2		6		5	Тест, отчет по лабораторной работе
Тема 2. Методы на основе символьного представления информации		2		6		5	отчет по лабораторной работе
Тема 3. Деревья принятия решений		2		6		5	Тест, отчет по лабораторной работе
Тема 4. Поиск в пространстве версий		2		6		5	отчет по лабораторной работе
Тема 5. Обучение без учителя		2		8		5	Тест, отчет по лабораторной работе
Тема 6. Нейронные сети: основы		2		8		5	Тест, отчет по лабораторной работе
Тема 7. Обучение нейронных сетей		3		7		5	отчет по лабораторной работе
Тема 8. Самообучение и социальные принципы		2		6		5	отчет по лабораторной работе, устный опрос
		17		**51**		**40**	**Зачет**

Примечание: Л – лекция; ПЗ – практическое занятие, семинар; ЛР – лабораторная работа; КР – курсовая работа; СР – самостоятельная работа.

Таблица 3. Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых компетенций

Раздел, тема дисциплины (модуля)	Кол-во часов	Компетенции	Общее количество компетенций
		ПК-6	
Тема 1. Введение в машинное обучение.	13	+	1
Тема 2. Методы на основе символьного представления информации	13	+	1
Тема 3. Деревья принятия решений	13	+	1
Тема 4. Поиск в пространстве версий	13	+	1

Раздел, тема дисциплины (модуля)	Кол-во часов	Компетенции	Общее количество компетенций
		ПК-6	
Тема 5. Обучение без учителя	15	+	1
Тема 6. Нейронные сети: основы	15	+	1
Тема 7. Обучение нейронных сетей	15	+	1
Тема 8. Самообучение и социальные принципы	13	+	1
Итого	**108**		

Краткое содержание каждой темы дисциплины (модуля)

Тема 1. Введение в машинное обучение

Понятие машинного обучения, цели и задачи. Подходы к обучению на основе символьного представления информации. Типы обучения: контролируемое, неконтролируемое, с подкреплением. Примеры приложений машинного обучения.

Тема 2. Методы на основе символьного представления информации

Представление знаний в символической форме. Логическое представление данных. Применение логики предикатов первого порядка в машинном обучении. Использование формальных грамматик и конечных автоматов.

Тема 3. Деревья принятия решений

Принцип построения дерева решений. Критерии выбора признаков. Алгоритмы ID3, CART, CHAID. Проблемы переобучения и способы борьбы с ними. Оценка качества классификации.

Тема 4. Поиск в пространстве версий

Пространства гипотез. Принцип исключения кандидатов. Детали реализации алгоритма исключения кандидатов. Ограничения метода и области применения.

Тема 5. Обучение без учителя

Задача кластеризации. Методы k-means, иерархическая кластеризация. Анализ главных компонент (PCA). Приложения методов обучения без учителя.

Тема 6. Нейронные сети: основы

Анатомия биологического нейрона. Модель нейрона Мак-Каллок-Питтса. Преимущества и ограничения модели. Функции активации нейронов. Топологии искусственных нейронных сетей.

Тема 7. Обучение нейронных сетей

Алгоритм обучения перцептрона Розенблатта. Линейная разделимость. Обратное распространение ошибок. Параметры обучения: скорость обучения, регуляризация. Градиентный спуск.

Тема 8. Самообучение и социальные принципы

Концепция популяционного подхода. Генетический алгоритм: кроссовер, мутация, отбор. Практическое применение генетических алгоритмов в задачах оптимизации. Эволюционные стратегии и их отличия от генетических алгоритмов

5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

Учебная деятельность студента в процессе изучения строится из контактных форм работы с преподавателем (аудиторные занятия, зачет с оценкой) и самостоятельной работы.

Для успешного освоения дисциплины является обязательным посещение всех занятий, выполнение домашнего задания и иных форм самостоятельной работы, которые назначаются преподавателем.

Методическая поддержка дисциплины обеспечивается использованием дистанционных технологий. Студентам предлагается информационный ресурс «Электронное образование». Доступ студентов к учебным ресурсам осуществляется по учетной записи и паролю после регистрации на курс «Машинное обучение» на период обучения по данной дисциплине.

На сервере размещен методический материал по данной дисциплине, в содержание которого входит:

– теоретический материал;

– задания и указания по выполнению лабораторно-практических работ, требования к содержанию и их оформлению, рекомендации по их защите;

– тестовые вопросы, предназначенные всех видов контроля, включая самоконтроль освоения учебного материала.

Аудиторные занятия проводятся на основе теоретического материала, опубликованного на образовательном портале, это позволяет студентам изучить пропущенный материал или самостоятельно разобраться с темой, не освоенной на занятии.

Для исключения отрыва студентов от учебного процесса проводится учет посещаемости аудиторных занятий.

5.2. Указания для обучающихся по освоению дисциплины (модулю)

В рамках дисциплины «Машинное обучение» предполагается организация следующих видов

– самостоятельной работы студентов (таблица 4):

– работа с лекционным материалом, учебно-методическим информационным обеспечением;

– подготовка к лабораторно-практическим работам, подготовка отчетов к защите отчетов;

– подготовка к компьютерному тестированию.

В качестве форм и методов контроля внеаудиторной самостоятельной работы используются: электронные отчеты по выполнению лабораторных работ; устный опрос, протоколы компьютерного тестирования.

Задания к лабораторно-практическим занятиям размещены на образовательном портале «Электронное образование». Рекомендуется заранее ознакомиться с темой, основными вопросами, рекомендациями, требованиями к представлению отчета и критериями оценивания заданий.

В процессе подготовки к лабораторно-практическим занятиям, необходимо обратить особое внимание на самостоятельное изучение рекомендованной литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала.

Таблица 4. Содержание самостоятельной работы обучающихся

для очной формы обучения

Темы/вопросы, выносимые на самостоятельное изучение	Кол-во часов	Форма работы
Тема 1. Введение в машинное обучение. Подготовка к тесту, Подготовка отчета по лабораторной работе	5	Изучение теоретического материала
Тема 2. Методы на основе символьного представления информации Подготовка отчета по лабораторной работе	5	
Тема 3. Деревья принятия решений Подготовка отчета по лабораторной работе Подготовка к тесту	5	
Тема 4. Поиск в пространстве версий Подготовка отчета по лабораторной работе	5	
Тема 5. Обучение без учителя Подготовка к тесту Подготовка отчета по лабораторной работе	5	
Тема 6. Нейронные сети: основы Подготовка отчета по лабораторной работе Подготовка к тесту	5	
Тема 7. Обучение нейронных сетей Подготовка отчета по лабораторной работе	5	
Тема 8. Самообучение и социальные принципы Подготовка отчета по лабораторной работе Подготовка к опросу на зачете с оценкой	5	

5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины (модуля), выполняемые обучающимися самостоятельно

Отчет по лабораторным работам.

Результатом работы, выполняемой студентами, является электронный отчет по выполнению лабораторно-практической работы, тематика которых представлена в таблице 4,

Электронный отчет представляет собой файл формата Jupiter Notebook, содержащий программный код, результаты выполнения программы и текстовые пояснения. Код должен быть исполняемым и его запуск должен приводить к тем же результатам, что и в предоставленном отчете. Файл передается на проверку преподавателю путем загрузки на платформу «Электронное образование» в соответствующий заданию раздел.

6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

6.1. Образовательные технологии

Учебные занятия по дисциплине могут проводиться с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) взаимодействии обучающихся и преподавателя в режимах on-line в формах: видеолекций, лекций-презентаций, видеоконференции, собеседования в режиме чат, форума, чата, выполнения виртуальных практических и/или лабораторных работ и др.

Таблица 5. Образовательные технологии, используемые при реализации учебных занятий

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Тема 1. Введение в машинное обучение.	Обзорная лекция	Не предусмотрено	Тест, Выполнение лабораторной работы
Тема 2. Методы на основе символьного представления информации	Обзорная лекция	Не предусмотрено	Выполнение лабораторной работы
Тема 3. Деревья принятия решений	Обзорная лекция	Не предусмотрено	Тест, Выполнение лабораторной работы
Тема 4. Поиск в пространстве версий	Лекция-презентация	Не предусмотрено	Выполнение лабораторной работы
Тема 5. Обучение без учителя	Лекция-презентация	Не предусмотрено	Тест, Выполнение лабораторной работы
Тема 6. Нейронные сети: основы	Лекция-презентация	Не предусмотрено	Тест, Выполнение лабораторной работы
Тема 7. Обучение нейронных сетей	Лекция-презентация	Не предусмотрено	Выполнение лабораторной работы
Тема 8. Самообучение и социальные принципы	Лекция-презентация	Не предусмотрено	Выполнение лабораторной работы

6.2. Информационные технологии

При реализации учебной и внеучебной работы используются следующие информационные технологии:

- образовательный сайт http://moodle.asu-edu.ru (размещение учебно-методического материала, публикация заданий для предоставления студентами выполненных отчетов по всем видам работ, ознакомление учащихся с оценками и т.д., размещение объявлений, on-line консультации, обсуждение вопросов в форуме и т.д.), как элемента интерактивного взаимодействия участников образовательного процесса (технологии дистанционного обучения);
- среда разработки моделей машинного обучения https://colab.research.google.com
- веб-сервис для хостинга IT-проектов и их совместной разработки https://github.com
- онлайн-визуализатор n-мерных векторов https://projector.tensorflow.org
- ресурсы ЭБС и сети Internet, как источников информации.

6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
Adobe Reader	Программа для просмотра электронных документов
Платформа дистанционного обучения	Виртуальная обучающая среда

LMS Moodle	
Google Chrome	Браузер
Microsoft Office 2013, Microsoft Office Project 2013, Microsoft Office Visio 2013	Офисная программа
7-zip	Архиватор
Microsoft Windows 10 Professional	Операционная система
Kaspersky Endpoint Security	Средство антивирусной защиты
MATLAB R2014a	Пакет прикладных программ для решения задач технических вычислений

6.3.2. Современные профессиональные базы данных и информационные справочные системы

1. Универсальная справочно-информационная полнотекстовая база данных периодических изданий ООО «ИВИС» http://dlib.eastview.com

2. Электронные версии периодических изданий, размещенные на сайте информационных ресурсов www.polpred.com

3. Электронный каталог Научной библиотеки АГУ на базе MARK SQL НПО «Информ-систем»: https://library.asu-edu.ru/catalog/.

4. Электронный каталог «Научные журналы АГУ»: http://journal.asu.edu.ru/issledovaniya-i-innovacii/11745-nauchnye-jurnaly-agu.html.

5. Корпоративный проект Ассоциации региональных библиотечных консорциумов (АРБИКОН) «Межрегиональная аналитическая роспись статей» (МАРС) http://mars.arbicon.ru

7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

7.1. Паспорт фонда оценочных средств

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) «Машинное обучение» проверяется сформированность у обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

Таблица 6. Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств

Контролируемый раздел, тема дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Тема 1. Введение в машинное обучение.	ПК-6	Тест
Тема 2. Методы на основе символьного представления информации	ПК-6	отчет по лабораторной работе
Тема 3. Деревья принятия решений	ПК-6	Тест, отчет по лабораторной работе
Тема 4. Поиск в пространстве версий	ПК-6	отчет по лабораторной работе
Тема 5. Обучение без учителя	ПК-6	Тест
Тема 6. Нейронные сети: основы	ПК-6	Тест, отчет по лабораторной работе
Тема 7. Обучение нейронных сетей	ПК-6	отчет по лабораторной работе

Контролируемый раздел, тема дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Тема 8. Самообучение и социальные принципы	ПК-6	отчет по лабораторной работе, устный опрос

7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

Таблица 7. Показатели оценивания результатов обучения в виде знаний

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

Таблица 8. Показатели оценивания результатов обучения в виде умений и владений

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание по подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задания

7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)

Тема 1. Введение в машинное обучение

Тест

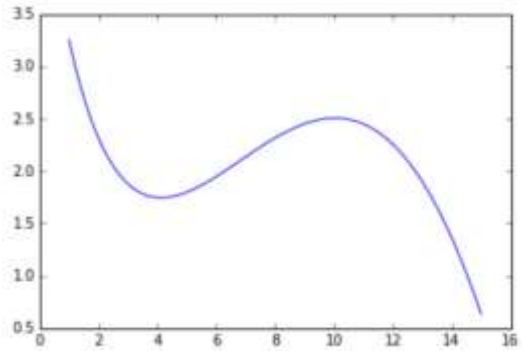
Вопрос 1

Диаграмма традиционного программирования на входе имеет Правила и Данные. А что у нее на выходе? Выберите один ответ:

1) Двоичные числа
2) Машинное обучение
3) Ошибки
4) Ответы

Лабораторная работа
Рассмотрим сложную математическую функцию на отрезке [1, 15]:

f(x) = sin(x / 5) * exp(x / 10) + 5 * exp(-x / 2)



Она может описывать, например, зависимость оценок, которые выставляют определенному сорту вина эксперты, в зависимости от возраста этого вина. По сути, задача машинного обучения состоит в том, чтобы приблизить сложную зависимость с помощью функции из определенного семейства. В этом задании мы будем приближать указанную функцию с помощью многочленов.

Как известно, многочлен степени n (то есть w_0 + w_1 x + w_2 x^2 + ... + w_n x^n) однозначно определяется любыми n + 1 различными точками, через которые он проходит. Это значит, что его коэффициенты w_0, ... w_n можно определить из следующей системы линейных уравнений:

$$\begin{cases} w_0 + w_1 x_1 + w_2 x_1^2 + \cdots + w_n x_1^n = f(x_1) \\ \cdots \\ w_0 + w_1 x_{n+1} + w_2 x_{n+1}^2 + \cdots + w_n x_{n+1}^n = f(x_{n+1}) \end{cases}$$

где через x_1, ..., x_n, x_{n+1} обозначены точки, через которые проходит многочлен, а через f(x_1), ..., f(x_n), f(x_{n+1}) — значения, которые он должен принимать в этих точках.

Воспользуемся описанным свойством, и будем находить приближение функции многочленом, решая систему линейных уравнений.

1. Сформируйте систему линейных уравнений (то есть задайте матрицу коэффициентов A и свободный вектор b) для многочлена первой степени, который должен совпадать с функцией f в точках 1 и 15. Решите данную систему с помощью функции scipy.linalg.solve. Нарисуйте функцию f и полученный многочлен. Хорошо ли он приближает исходную функцию?
2. Повторите те же шаги для многочлена второй степени, который совпадает с функцией f в точках 1, 8 и 15. Улучшилось ли качество аппроксимации?
3. Повторите те же шаги для многочлена третьей степени, который совпадает с функцией f в точках 1, 4, 10 и 15. Хорошо ли он аппроксимирует функцию? Коэффициенты данного многочлена (четыре числа в следующем порядке: w_0, w_1, w_2, w_3) являются ответом на задачу. Округлять коэффициенты не обязательно, но при желании можете произвести округление до второго знака (т.е. до числа вида 0.42)
4. Запишите полученные числа в файл, разделив пробелами. Обратите внимание, что файл должен состоять из одной строки, в конце которой не должно быть переноса.

Тема 2. Методы на основе символьного представления информации

Лабораторная работа

Скачать датасет статей, содержащий два класса («наука» и «культура»), используя библиотеки Python (pandas, numpy).

Провести предварительную обработку текста: удаление стоп-слов, знаков пунктуации, приведение к нижнему регистру, лемматизацию.

Применить метод CountVectorizer для преобразования текстов в числовое представление (символьное пространство признаков). Это позволит перевести символы в числа, подходящие для дальнейшего анализа.

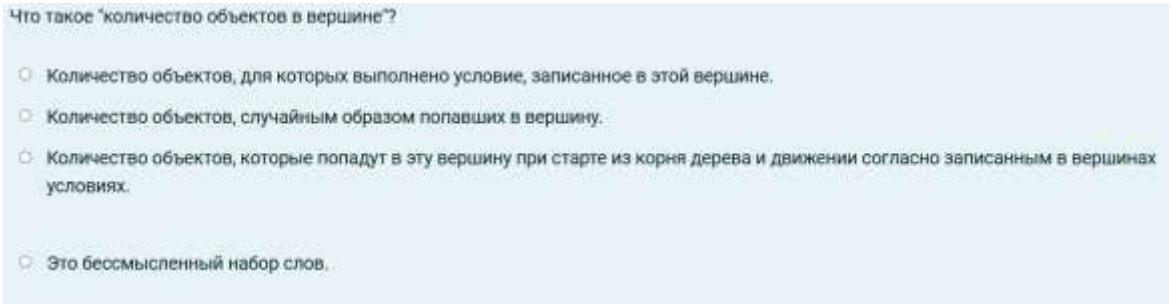
Разделите исходный корпус данных на тренировочную и тестовую части в соотношении 80% / 20%.

Используйте простой алгоритм логистической регрессии для построения классификатора.

Рассчитайте точность предсказания модели на тестовом множестве

Тема 3. Деревья принятия решений

Тест



…….

Лабораторная работа

Загрузите датасет digits с помощью функции load_digits из sklearn.datasets и подготовьте матрицу признаков X и ответы на обучающей выборке y (вам потребуются поля data и target в объекте, который возвращает load_digits).

Для оценки качества далее нужно будет использовать cross_val_score из sklearn.model_selection с параметром cv=10. Эта функция реализует k-fold cross validation с k равным значению параметра cv. Предлагаю использовать k=10, чтобы полученные оценки качества имели небольшой разброс, и было проще проверить полученные ответы. На практике же часто хватает и k=5. Функция cross_val_score будет возвращать numpy.ndarray, в котором будет k чисел - качество в каждом из k экспериментов k-fold cross validation. Для получения среднего значения (которое и будет оценкой качества работы) вызовите метод .mean() у массива, который возвращает cross_val_score.

С небольшой вероятностью вы можете натолкнуться на случай, когда полученное вами качество в каком-то из пунктов не попадет в диапазон, заданный для правильных ответов - в этом случае попробуйте перезапустить ячейку с cross_val_score несколько раз и выбрать наиболее «типичное» значение. Если это не помогает, то где-то была допущена ошибка.

Если вам захочется ускорить вычисление cross_val_score - можете попробовать использовать параметр n_jobs, но будьте осторожны: в одной из старых версий sklearn была ошибка, которая приводила к неверному результату работы cross_val_score при задании n_jobs отличным от 1. Сейчас такой проблемы возникнуть не должно, но проверить, что все в порядке, не будет лишним.

1.Создайте DecisionTreeClassifier с настройками по умолчанию и измерьте качество его работы с помощью cross_val_score.

2.Воспользуйтесь BaggingClassifier из sklearn.ensemble, чтобы обучить бэггинг над DecisionTreeClassifier. Используйте в BaggingClassifier параметры по умолчанию, задав только количество деревьев равным 100.

Качество классификации новой модели - ответ на пункт 2. Обратите внимание, как соотносится качество работы композиции решающих деревьев с качеством работы одного решающего дерева.

3.Теперь изучите параметры BaggingClassifier и выберите их такими, чтобы каждый базовый алгоритм обучался не на всех d признаках, а на d случайных признаков. Качество работы получившегося классификатора - ответ на пункт 3. Корень из числа признаков - часто используемая эвристика в задачах классификации, в задачах регрессии же часто берут число признаков, деленное на три. Но в общем случае ничто не мешает вам выбирать любое другое число случайных признаков.

4.Наконец, давайте попробуем выбирать случайные признаки не один раз на все дерево, а при построении каждой вершины дерева. Сделать это несложно: нужно убрать выбор случайного подмножества признаков в BaggingClassifier и добавить его в DecisionTreeClassifier. Какой параметр за это отвечает, можно понять из документации sklearn, либо просто попробовать угадать (скорее всего, у вас сразу получится). Попробуйте выбирать опять же d признаков. Качество полученного классификатора на контрольной выборке и будет ответом на пункт 4.

5.Полученный в пункте 4 классификатор - бэггинг на рандомизированных деревьях (в которых при построении каждой вершины выбирается случайное подмножество признаков и разбиение ищется только по ним). Это в точности соответствует алгоритму Random Forest, поэтому почему бы не сравнить качество работы классификатора с RandomForestClassifier из sklearn.ensemble. Сделайте это, а затем изучите, как качество классификации на данном датасете зависит от количества деревьев, количества признаков, выбираемых при построении каждой вершины дерева, а также ограничений на глубину дерева. Для наглядности лучше построить графики зависимости качества от значений параметров, но для сдачи задания это делать не обязательно.

На основе наблюдений выпишите через пробел номера правильных утверждений из приведенных ниже в порядке возрастания номера (это будет ответ на пункт 5)

1) Случайный лес сильно переобучается с ростом количества деревьев

2) При очень маленьком числе деревьев (5, 10, 15), случайный лес работает хуже, чем при большем числе деревьев

3) С ростом количества деревьев в случайном лесе, в какой-то момент деревьев становится достаточно для высокого качества классификации, а затем качество существенно не меняется.

4) При большом количестве признаков (для данного датасета - 40, 50) качество классификации становится хуже, чем при малом количестве признаков (5, 10). Это связано с тем, что чем меньше признаков выбирается в каждом узле, тем более различными получаются деревья (ведь деревья сильно неустойчивы к изменениям в обучающей выборке), и тем лучше работает их композиция.

5) При большом количестве признаков (40, 50, 60) качество классификации лучше, чем при малом количестве признаков (5, 10). Это связано с тем, что чем больше признаков - тем больше информации об объектах, а значит алгоритм может делать прогнозы более точно.

6) При небольшой максимальной глубине деревьев (5-6) качество работы случайного леса намного лучше, чем без ограничения глубины, т.к. деревья получаются не переобученными. С ростом глубины деревьев качество ухудшается.

7) При небольшой максимальной глубине деревьев (5-6) качество работы случайного леса заметно хуже, чем без ограничений, т.к. деревья получаются недообученными. С ростом глубины качество сначала улучшается, а затем не меняется существенно, т.к. из-за усреднения прогнозов и различий деревьев их переобученность в бэггинге не сказывается на

итоговом качестве (все деревья преобучены по-разному, и при усреднении они компенсируют переобученность друг-друга).

Тема 4. Поиск в пространстве версий

Лабораторная работа

В этом задании вам предстоит проверить работу центральной предельной теоремы, а также поработать с генерацией случайных чисел и построением графиков в Питоне.

Выберите непрерывное распределение (чем меньше оно будет похоже на нормальное, тем интереснее; попробуйте выбрать какое-нибудь распределение из тех, что мы не обсуждали в курсе). Сгенерируйте из него выборку объёма 1000, постройте гистограмму выборки и нарисуйте поверх неё теоретическую плотность распределения вашей случайной величины (чтобы величины были в одном масштабе, не забудьте выставить у гистограммы значение параметра normed=True).

Ваша задача — оценить распределение выборочного среднего вашей случайной величины при разных объёмах выборок. Для этого при трёх и более значениях n (например, 5, 10, 50) сгенерируйте 1000 выборок объёма n и постройте гистограммы распределений их выборочных средних. Используя информацию о среднем и дисперсии исходного распределения (её можно без труда найти в википедии), посчитайте значения параметров нормальных распределений, которыми, согласно центральной предельной теореме, приближается распределение выборочных средних.

Обратите внимание: для подсчёта значений этих параметров нужно использовать именно теоретические среднее и дисперсию вашей случайной величины, а не их выборочные оценки. Поверх каждой гистограммы нарисуйте плотность соответствующего нормального распределения (будьте внимательны с параметрами функции, она принимает на вход не дисперсию, а стандартное отклонение).

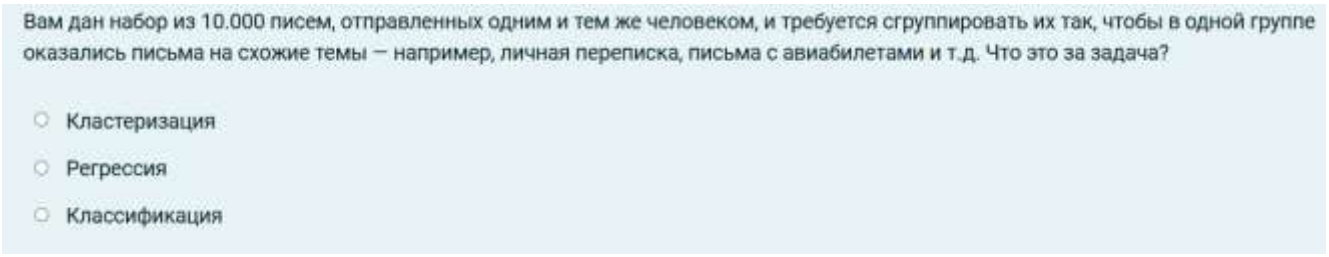
Опишите разницу между полученными распределениями при различных значениях n. Как меняется точность аппроксимации распределения выборочных средних нормальным с ростом n?

Решение должно представлять собой IPython-ноутбук, содержащий:

- код, генерирующий выборки и графики;
- краткие описания каждого блока кода, объясняющие, что он делает;
- необходимые графики (убедитесь, что на них подписаны оси);
- выкладки с вычислениями параметров нормальных распределений, аппроксимирующих выборочные средние при различных n;
- выводы по результатам выполнения задания.

Тема 5. Обучение без учителя

Тест



……

Лабораторная работа

Загрузите датасет с помощью функции **pandas.read_csv** в переменную **df**. Выведите первые 5 строчек, чтобы убедиться в корректном считывании данных:

In [ ]:
```
# (0 баллов)
# Считайте данные и выведите первые 5 строк
```

Для каждого дня проката известны следующие признаки (как они были указаны в источнике данных):

- *season*: 1 - весна, 2 - лето, 3 - осень, 4 - зима
- *yr*: 0 - 2011, 1 - 2012
- *mnth*: от 1 до 12
- *holiday*: 0 - нет праздника, 1 - есть праздник
- *weekday*: от 0 до 6
- *workingday*: 0 - нерабочий день, 1 - рабочий день
- *weathersit*: оценка благоприятности погоды от 1 (чистый, ясный день) до 4 (ливень, туман)
- *temp*: температура в Цельсиях
- *atemp*: температура по ощущениям в Цельсиях
- *hum*: влажность
- *windspeed(mph)*: скорость ветра в милях в час
- *windspeed(ms)*: скорость ветра в метрах в секунду
- *cnt*: количество арендованных велосипедов (это целевой признак, его мы будем предсказывать)

Итак, у нас есть вещественные, бинарные и номинальные (порядковые) признаки, и со всеми из них можно работать как с вещественными. С номинальныеми признаками тоже можно работать как с вещественными, потому что на них задан порядок. Давайте посмотрим на графиках, как целевой признак зависит от остальных

In [ ]:
```
fig, axes = plt.subplots(nrows=3, ncols=4, figsize=(15, 10))
for idx, feature in enumerate(df.columns[:-1]):
    df.plot(feature, "cnt", subplots=True, kind="scatter", ax=axes[idx / 4, idx % 4])
```

Блок 1. Ответьте на вопросы (каждый 0.5 балла):

1. Каков характер зависимости числа прокатов от месяца?
   - ответ:
2. Укажите один или два признака, от которых число прокатов скорее всего зависит линейно
   - ответ:

Давайте более строго оценим уровень линейной зависимости между признаками и целевой переменной. Хорошей мерой линейной зависимости между двумя векторами является корреляция Пирсона. В pandas ее можно посчитать с помощью двух методов датафрейма: corr и corrwith. Метод df.corr вычисляет матрицу корреляций всех признаков из датафрейма. Методу df.corrwith нужно подать еще один датафрейм в качестве аргумента, и тогда он посчитает попарные корреляции между признаками из df и этого датафрейма.

In [ ]:
```
# Код 1.1 (0.5 балла)
# Посчитайте корреляции всех признаков, кроме последнего, с последним с помощью метода corrwith:
```

Тема 6. Нейронные сети: основы

Тест

Выберите верные высказывания о качестве работы нейросетей.

- ☐ Двухслойная нейросеть может корректно разделить любую линейно разделимую выборку.
- ☐ Существует нейронная сеть, у которой качество классификации лучше, чем у всех других алгоритмов классификации на любой выборке.
- ☐ Любую непрерывную функцию нескольких аргументов можно представить как суперпозицию функций одного аргумента и функции суммирования.
- ☐ Существует такая нейронная сеть, которая аппроксимирует любую непрерывную разделяющую поверхность.

……

Лаборторная работа

1NN против RF

В этом задании будет использоваться датасет digits из sklearn.datasets. Оставьте последние 25% объектов для контроля качества, разделив X и y на X_train, y_train и X_test, y_test.

Целью задания будет реализовать самый простой метрический классификатор — метод ближайшего соседа, а также сравнить качество работы реализованного вами 1NN с RandomForestClassifier из sklearn на 1000 деревьях.

Задание

Реализуйте самостоятельно метод одного ближайшего соседа с евклидовой метрикой для задачи классификации. Можно не извлекать корень из суммы квадратов отклонений, т.к. корень — монотонное преобразование и не влияет на результат работы алгоритма.

Никакой дополнительной работы с признаками в этом задании делать не нужно — мы еще успеем этим заняться в других курсах. Ваша реализация может быть устроена следующим образом: можно для каждого классифицируемого объекта составлять список пар (расстояние до точки из обучающей выборки, метка класса в этой точке), затем сортировать этот список (по умолчанию сортировка будет сначала по первому элементу пары, затем по второму), а затем брать первый элемент (с наименьшим расстоянием).

Сортировка массива длиной N требует порядка N log N сравнений (строже говоря, она работает за O(N log N)). Подумайте, как можно легко улучшить получившееся время работы. Кроме простого способа найти ближайший объект всего за N сравнений, можно попробовать придумать, как разбить пространство признаков на части и сделать структуру данных, которая позволит быстро искать соседей каждой точки. За выбор метода поиска ближайших соседей в KNeighborsClassifier из sklearn отвечает параметр algorithm — если у вас уже есть некоторый бэкграунд в алгоритмах и структурах данных, вам может быть интересно познакомиться со структурами данных ball tree и kd tree.

Доля ошибок, допускаемых 1NN на тестовой выборке, — ответ

Тема 7. Обучение нейронных сетей

Лабораторная работа

Напишите классификатор MNIST, который обучается до точности 99% или выше и делает это без фиксированного числа эпох - то есть вы должны прекратить обучение, как только достигнете этого уровня точности.

1. Этого надо достигнуть менее чем за 10 эпох, поэтому можно установить epoches=10, но не более.
2. Когда точность станет 99% или больше, надо распечатать строку «Достигнута точность 99%, поэтому обучение закончено!»

```
import tensorflow as tf
from os import path, getcwd, chdir
# заберите mnist.npz из github
# поместите в локальную папку и пропишите на нее путь
path = f"{getcwd()}/../tmp2/mnist.npz"
def train_mnist():
    # Здесь ваш код
    import tensorflow as tf
    mnist = tf.keras.datasets.mnist
    (x_train, y_train),(x_test, y_test) = mnist.load_data()
    # Здесь ваш код
    model = tf.keras.models.Sequential([
    # Здесь ваш код
    ])
    model.compile(optimizer='adam',
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])
    # Здесь ваш код
    return history.epoch, history.history['acc'][-1]
train_mnist()
```

Тема 8. Самообучение и социальные принципы

Лабораторная работа

1. Исследовать функцию f(x1,x2)=100*(x2-x1 2) 2+(1-x2) 2 с помощью генетических алгоритмов. Определить глобальный минимум и значение функции в этой точке.

2. Провести эксперимент при различном размере начальной популяции: 10; 50; 300; 800. Для каждого из этих значений принять следующие операторы отбора родительских особей:

- Stochastic uniform;
- Uniform;
- Roulette.

3. Сделать вывод о том, как влияет размер исходной популяции и оператор отбора родительских особей на результаты.

4. *Найти с помощью генетического алгоритма точные значения глобального минимума функции (с помощью гибридных функций).

Перечень вопросов и заданий,
выносимых на зачёт

1. Что такое машинное обучение и какие задачи оно решает?
2. Какие существуют типы машинного обучения и чем они отличаются друг от друга?
3. Каково значение символьного представления информации в машинном обучении?
4. Приведите примеры использования логического представления данных в задачах машинного обучения.
5. Опишите принцип работы алгоритма построения дерева решений.
6. Чем различаются алгоритмы ID3, CART и CHAID?
7. Объясните проблему переобучения и предложите способы её устранения.
8. В чём заключается метод исключения кандидатов?
9. Дайте определение пространства гипотез и поясните его использование в машинном обучении.
10. Назовите преимущества и недостатки метода пространственного поиска версии.
11. В чём состоит задача кластеризации?
12. Расскажите о работе алгоритма k-means и приведите пример его применения.
13. Охарактеризуйте метод анализа главных компонент (PCA).
14. Какие преимущества имеет обучение без учителя перед классическим подходом?
15. Как устроены биологические нейроны и какую модель предложили Мак-Каллок и Питтс?
16. Почему важно правильно выбрать функцию активации нейрона?
17. Какие топологии используются в искусственных нейронных сетях?
18. Какой основной принцип лежит в основе алгоритма обратного распространения ошибок?
19. Для чего применяется градиентный спуск в процессе обучения нейронных сетей?
20. В чём заключаются особенности линейной разделимости в задаче классификации?
21. Как работает генетический алгоритм и зачем нужен кроссовер и мутация?
22. Объясните разницу между генетическими алгоритмами и эволюционными стратегиями.
23. Где находят своё применение генетические алгоритмы?
24. Предложите рекомендации по подготовке к экзамену по машинному обучению.
25. Какие перспективы развития машинного обучения вы видите в будущем?

Таблица 9. Примеры оценочных средств с ключами правильных ответов

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
ПК-6 Способен обрабатывать и анализировать данные для подготовки аналитических решений, экспертных заключений и рекомендаций				

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
1.	Задание закрытого типа	Блок-схема машинного обучения на входе получает Ответы и Данные. А что у нее на выходе? 1 Двоичные числа 2 Модели 3 Правила 4 Ошибки	2	1
2.		Что делает оптимизатор? 1 Генерирует новое улучшенное предположение 2 Выясняет, насколько эффективно скомпилирован ваш код 3 Принимает решение об остановке обучения нейронной сети 4 Измеряет, насколько хорошо текущее предположение	1	1
3.		С какой целью данные делят на обучающий и тестовый наборы? 1 чтобы обучить сеть с ранее неизвестными данными 2 чтобы тестировать сеть с ранее неизвестными данными 3 сделать тестирование быстрее 4 сделать обучение быстрее	2	1
4.		Какой метод вызывается, когда заканчивается эпоха обучения? 1 On_epoch_end 2 On_training_complete 3 On_end 4 On_epoch_finished	1	1
5.		Если размер изображения 150x150, и к нему применили свертку 3x3, какой размер получится в результате? Поясните ответ 1 450x450 2 148x148 3 153x153 4 150x150	2 W' = W - F + 1, где: - W' — новый размер ширины (или высоты), - W — исходный размер ширины (или высоты), - F — размер фильтра (стандартно - 1).	3
6.	Задание открытого типа	Что делает функция ReLU?	Для положительных значений аргумента (x>0) функция возвращает сам аргумент (x). Для отрицательных значений аргумента	3

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			($x \leq 0$) функция возвращает ноль.	
7.		Для каких задач применяется обучение с учителем?	Обучение с учителем (supervised learning) применяется для решения задач, в которых известны правильные ответы (метки классов или целевые переменные) для каждого примера в обучающем наборе. Основные типы задач, решаемых методом обучения с учителем: классификация, регрессия	3
8.		Для каких задач применяется обучение без учителя?	Обучение без учителя (unsupervised learning) применяется в ситуациях, когда отсутствуют размеченные данные, и целью является выявление скрытых структур или закономерностей в необработанных данных. Основные задачи, решаемые методами обучения без учителя: кластеризация, редукция размерности, ассоциативные правила	3
9.		Как называется библиотека Tensorflow, содержащая общедоступные датасеты, которые можно использовать для тренировки и тестирования нейронных сетей?	TensorFlow Datasets	2
10.		Что такое автокорреляция?	Автокорреляция — это мера корреляции временного ряда с самим собой, сдвинутого на некоторое количество временных интервалов	3

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			назад или вперед. Говоря простыми словами, это степень, в которой значения ряда зависят друг от друга через временные промежутки.	

Полный комплект оценочных материалов по дисциплине (модулю) (фонд оценочных средств) хранится в электронном виде на кафедре, утверждающей рабочую программу дисциплины (модуля).

7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)

Итоговая оценка по промежуточной аттестации выставляется в соответствии с Положением АГУ о балльно-рейтинговой системе (БАРС). Итоговая оценка складывается из баллов, полученных студентом за текущую успеваемость. Для получения положительной оценки студенту необходимо набрать в семестре минимально 60 баллов. В течение семестра студент может набрать максимально 90 баллов за выполнение аудиторной и самостоятельной работы.

Для текущего контроля знаний, умений, навыков и (или) опыта деятельности, необходимых для формирования компетенции дисциплины «Машинное обучение», используется инструментарий системы Moodle: *Тест, Задание*.

Для стимулирования развития творческого и научно-исследовательского потенциала студентов при промежуточном оценивании предусмотрена система дополнительных баллов, а именно: начисление до 10 поощрительных баллов за участие в конференциях, семинарах, выставках и т.п. в области машинного обучения, программировании с представлением индивидуальных проектов в области машинного обучения. Начисление баллов зависит от статуса мероприятия и статуса участия в нем студента. Начисление баллов происходит при предоставлении диплома, сертификата, грамоты, материалов конференции, опубликованной статьи, тезисов и т.п.

Таблица 10. Технологическая карта рейтинговых баллов по дисциплине (модулю)

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Основной блок				
1.	*Тест*	4/6	24	По расписанию
2.	*Лабораторная работа*	8/6	48	
3.	*Устный опрос*	1/18	18	
Всего			**90**	-
Блок бонусов				
4.	*Посещение занятий*		2	
5.	*Своевременное выполнение всех заданий*		2	
6.	*Участие в профильных мероприятиях*		6	

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
Всего			**10**	-
ИТОГО			**100**	-

Таблица 11. Система штрафов (для одного занятия)

Показатель	Балл
Нарушение учебной дисциплины	-1
Неготовность к занятию	-1
Пропуск занятия без уважительной причины	-1
Списывание	-5

Таблица 12. Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)

Сумма баллов	
90–100	Зачтено
85–89	
75–84	
70–74	
65–69	
60–64	
Ниже 60	Не зачтено

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

8.1. Основная литература

1. Мирзоев М.С., Основы математической обработки информации / М.С. Мирзоев - М. : Прометей, 2016. - 316 с. - ISBN 978-5-906879-01-1 - // ЭБС "Консультант студента": [сайт]. - URL : http://www.studentlibrary.ru/book/ISBN9785906879011.html.

2. Рыбина Г.В., Основы построения интеллектуальных систем : учеб. пособ./ Г.В. Рыбина. - М. : Финансы и статистика, 2014. - 432 с. - ISBN 978-5-279-03412-3 - // ЭБС "Консультант студента" : URL : http://www.studentlibrary.ru/book/ISBN9785279034123.html.

3. Системы искусственного интеллекта. Модуль "Модели и методы извлечения знаний" [Электронный ресурс] / Яковина И.Н. - Новосибирск : Изд-во НГТУ, 2014. - http://www.studentlibrary.ru/book/ISBN9785778225879.html.

8.2. Дополнительная литература

1. Флах П., Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / Флах П. - М. : ДМК Пресс, 2015. - 400 с. - ISBN 978-5-97060-273-7. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970602737.html.

2. Рашка, С. Python и машинное обучение / С. Рашка; пер. с англ. А. В. Логунова. - Москва : ДМК Пресс, 2017. - 418 с. - ISBN 978-5-97060-409-0. - Текст : электронный // ЭБС

"Консультант студента" : [сайт]. - URL : https://www.studentlibrary.ru/book/ISBN9785970604090.html

8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)

1. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». Многопрофильный образовательный ресурс «Консультант студента». www.studentlibrary.ru.
2. среда разработки моделей машинного обучения https://colab.research.google.com
3. веб-сервис для хостинга IT-проектов и их совместной разработки https://github.com
4. онлайн-визуализатор n-мерных векторов https://projector.tensorflow.org

9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Учебные аудитории, библиотеки АГУ, компьютерные классы, мультимедийные аудитории.

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. Для инвалидов содержание рабочей программы дисциплины (модуля) может определяться также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).