РАБОЧАЯ ПРОГРАММА

модуля(курса)

«Основы технологии машинного обучения и искусственного интеллекта»

1. АННОТАЦИЯ

Дисциплина «Основы технологии машинного обучения и искусственного интеллекта» дает общее представление о задачах и методах искусственного интеллекта, а также истории его развития. Рассказывается в теории и на практике об основных методах решения задач обучения с учителем и без учителя.

Дисциплина рассматривается, как один из основных курсов программы профессиональной переподготовки «Искусственный интеллект и глубокое обучение».

2. СОДЕРЖАНИЕ

Учебная программа курса

№ п/п	Наименование модуля, разделов и тем	Содержание обучения (по темам в дидактических единицах), наименование и тематика лабораторных работ, практических занятий (семинаров), самостоятельной работы с указанием кол-ва часов, используемых образовательных технологий и рекомендуемой литературы
1.	*2.*	*3.*
1	Тема 1. Введение в искусственный интеллект и машинное обучение.	Основная цель раздела - дать общее представление о задачах и методах искусственного интеллекта, а также истории его развития. Рассматриваются следующие концепции: сильный (общий) и слабый (прикладной) искусственный интеллект; ИИ, основанный на правилах (дедуктивный подход) и ИИ, основанный на машинном обучении (дедуктивный подход); обучение с учителем и без учителя; задачи классификации, регрессии, предсказания временного ряда, кластеризации. Даются краткие сведения об истории и перспективах развития ИИ; обзор основных достижений ИИ, в том числе в решении задач обработки естественного языка. Рассматриваются примеры задач машинного обучения; связь между машинным обучением обработкой данных. Вводятся концепции недообучения и переобучения и обсуждаются возможные их причины. Рассматриваются некоторые специфические вопросы, относящиеся к методам машинного обучения для решения задач обработки естественного языка (классификация задач, метод “мешок слов”, векторные представления и др.) (2 часа)

2	Тема 2. Задача обучения с учителем.	Основная цель раздела – дать представление об основных методах решения задач обучения с учителем. Рассматриваются линейная регрессия и метод наименьших квадратов; метод ближайшего соседа. Дается представление о методах, основанных на построении деревьев решений и их ансамблей. Рассматривается логистическая регрессия и нейронные сети. Дается понятие о методе обратного распространения ошибки (backpropagation). Введение в библиотеку scikit-learn и решение с помощью нее задач обучения с учителем. Основная цель - привить студентам навыки работы в библиотеке scikit-learn для решения задач обучения с учителем. Дается обзор библиотеки. Рассматриваются основные методы. На примере решения задач классификации и регрессии (в том числе из области обработки естественного языка и лингвистики) рассматривается, как обучать (настраивать) модель, тестировать и верифицировать. (2 часа)
3	Тема 3. Задача обучения без учителя.	Основная цель раздела - дать представление об основных методах решения задач обучения без учителя. Дается представление о методе k-средних и DBSCAN для решения задач кластеризации. Ставится задача понижения размерности и дается представление о методе главных компонент. Решения задач обучения без учителя с помощью библиотеки scikit-learn. Основная цель раздела - привить студентам навыки работы в библиотеке scikit-learn для решения задач кластеризации и понижения размерности. (2 часа)
4	Тема 4. Глубокое обучение	Глубокое обучение. Рассматриваются основные принципы глубокого обучения, в том числе для решения задач обработки естественного языка и обработки неструктурированных данных. Методы глубокого обучения в библиотеке TensorFlow. Основная цель раздела - дать представление и привить некоторые практические навыки, как в библиотеке TensorFlow (или другой эквивалентной) можно решать задачи ИИ, включая задачи компьютерной лингвистики и обработки естественного языка, методами глубокого обучения. (2 часа)
5	Тема 5. Обеспечение информационной безопасности в сфере искусственного	Риски информационной безопасности для искусственного интеллекта. Атака на алгоритм (модель). Внесение изменений в алгоритм с целью принятия неверных решений.

	интеллекта	Подстройка под алгоритм с целью изучения принципов принятия решения и последующего обхода модели. Состязательные атаки с целью генерации данных, которые обманывают модель. Атака на датасет. Внесение посторонних данных с целью принятия моделью неверных решений. Изменение существующих данных с целью принятия моделью неверных решений. (2 часа)
5	Практические занятия (семинары)	Библиотеки данных для анализа. Начало работы с Python. Визуализация эмпирических распределений средствами Python Анализ многомерных данных: выявление зависимости Инструменты Python для построения регрессии Средства Python кластеризации данных Оценивание неизвестных параметров закона распределения данных Решение кейсовых заданий (24 часа)
	Самостоятельная работа	Выполнение домашних заданий по теме занятия (24 часа)
	Зачет	Лабораторная работа (2 час)

3. ОЦЕНКА КАЧЕСТВА ОСВОЕНИЯ ПРОГРАММЫ

(формы аттестации, оценочные и методические материалы)

Промежуточная аттестация представляет собой дифференцированный зачет, который проводится по результатам выполнения лабораторной работы.

Перечень вопросов для подготовки к лабораторной работе

1. Способы визуализации научных данных с использованием Matplotlib. Иерархическая структура рисунка в Matplotlib. Основные элементы графика. Основные типы графиков.
2. Основы статистики в Python. Описательные статистики. Построение диаграмм размаха, гистограмм, функций плотности вероятности, планок погрешностей.
3. Основы статистики в Python. Корреляционный анализ. Проверка статистических гипотез. FDR-коррекция.
4. Численное решение дифференциальных уравнений, решение систем линейных уравнений с помощью библиотек SciPy и NumPy. Методы Монте-Карло.
5. Постановка задачи машинного обучения. Основные классы задач в машинном обучении. Примеры практических задач.
6. Вероятностная постановка задачи обучения с учителем. Функция потерь. Средний риск. Эмпирический риск.
7. Экспериментальная оценка качества обучения и выбор параметров модели. Обучающая, проверочная и тестовая выборки. Метод перекрестного (скользящего) контроля. Переобучение и недообучение.
8. Наивный байесовский классификатор. Сглаживание Лапласа. Ошибки 1-го и 2-го рода. Чувствительность, специфичность, точность, полнота. ROC-кривая. Использование наивного байесовского классификатора для количественных признаков.
9. Метод k ближайших соседей в задачах классификации и восстановления регрессии.
10. Методы предобработки данных. Методы понижения размерности. Метод главных

компонент. Сингулярное разложение.
11. Линейная регрессия. Метод наименьших квадратов. Проверка статистической значимости модели. Коэффициент детерминации.
12. Методы борьбы с переобучением в задаче восстановления регрессии. Отбор признаков. Регуляризация.
13. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ.
14. Логистическая регрессия. Логистическая функция и softmax.
15. Машина опорных векторов. Оптимальная разделяющая гиперплоскость. Опорные векторы. Случаи линейно-разделимых и неразделимых классов. Ядра и спрямляющие пространства.
16. Деревья решений. Алгоритм CART.
17. Ансамбли решающих правил. Баггинг. Случайный лес. Экстремально случайные деревья.
18. Ансамбли решающих правил. Бустинг. AdaBoost. Градиентный бустинг деревьев решений.
19. Нейронные сети. Персептрон Розенблатта. Сети прямого распространения. Алгоритм обратного распространения ошибки.
20. Глубокое обучение. Сверточные нейронные сети.
21. Обучение без учителя. Задача кластеризации. Метод центров тяжести. Метод медоидов.
22. Алгоритмы кластеризации: EM-алгоритм, алгоритм DBSCAN. Алгоритмы иерархической кластеризации.

Примеры наборов данных для теории и практики:

1. FLATS – база данных квартир г. Нижнего Новгорода
2. https://raw.githubusercontent.com/NikolaiZolotykh/MachineLearningCourse/master/flats.csv
3. MNIST – база данных образцов рукописного написания цифр https://www.kaggle.com/datasets/hojjatk/mnist-dataset

Примеры наборов данных для самостоятельной работы:
https://www.kaggle.com/purumalgi/music-genre-classification

https://www.kaggle.com/sagnik1511/car-insurance-data

https://www.kaggle.com/fedesoriano/heart-failure-prediction/version/1

https://www.kaggle.com/teertha/personal-loan-modeling

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

https://www.kaggle.com/rashikrahmanpritom/177k-english-song-data-from-20082017

https://www.kaggle.com/shivan118/hranalysis?select=train.csv

https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction?select=fake_job_postings.csv

https://www.kaggle.com/naveengowda16/logistic-regression-heart-disease-prediction

https://www.kaggle.com/kaushiksuresh147/customer-segmentation?select=Train.csv

https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

https://www.kaggle.com/sobhanmoosavi/us-accidents/version/10

https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

https://www.kaggle.com/crowdflower/twitter-user-gender-classification

https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey

https://www.kaggle.com/mssmartypants/water-quality

https://www.kaggle.com/code/gauravduttakiit/covid-19-sentiment-analysis-on-train-data/

https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website

Задания для самостоятельной работы

(можно выбрать любой набор данных из приведенных выше)

Задания должны быть выполнены в JupiterNotebook и содержать соответствующие текстовые пояснения, программный код и результаты его работы.

1. Загрузите данные
2. Опишите задачу словами. В том числе напишите, что значит каждый признак
3. Разбейте данные на обучающую и тестовую выборки
4. Визуализируйте данные из обучающей выборки. В частности, имеет смысл построить диаграммы рассеивания для количественных признаков. Построить гистограммы распределений и т.п. Вычислить основные характеристики (среднее, разброс, корреляционную матрицу и т.д.). Интерпретируйте результаты
5. Обработать пропущенные значения (или убедиться, что их нет)
6. Исключить нерелевантные признаки (объяснить, как вы их нашли)
7. Если необходимо, то обработать коррелированные признаки
8. Обработать категориальные признаки
9. Провести масштабирование (или объяснить, почему в вашем случае она не нужна)
10. Вам может понадобиться другая предобработка. Например, если в вашем датасете есть текстовые признаки с уникальными значениями (например, аннотации товаров, отзывы пользователей, другие тексты), как в двух последних датасетах из перечисленных, то вам понадобится этап извлечения признаков, т.е. простые методы NLP, как, например, bag-of-words. Воспользуйтесь библиотеками re, nltk
11. После шагов 5–10 разумно вернутся к шагу 4 (а может, возвращаться к нему после каждого из этапов 5–10.
12. Попробуйте как минимум 3 метода классификации (регрессии). Объясните ваш выбор. Найдите значения метрик на обучающей и тестовой выборке. Сделайте вывод.
13. На одном из методов (объясните выбор) найдите оптимальное значение параметров. Постройте график зависимости ошибок (на обучающей выборке и валидационной/CV) от значения гиперпарамтра. Для найденного оптимального значения параметра (параметров) снова обучите модель. Сделайте вывод.
14. Довольны ли вы результатами? В частности, если классы не сбалансированы, то результат может оказаться неприемлемым. В этом случае можете применить методы балансировки из библиотеки imbalanced-learn.
15. Для ваших данных сформулируйте задачу кластеризации. Обучите несколько методов кластеризации (не менее двух, например, k-means, DBSCAN). Объясните ваш выбор.

Сравните результаты работы алгоритмов, а также полученные результаты с результатами работы методов обучения с учителем.

16. Сделать общие выводы

Методы контроля и оценки результатов освоения модуля

№ п/п	Наименование процедуры	Основные показатели оценки	Формы и методы контроля и оценки
1	Промежуточная аттестация. Модуль 3. Основы технологии машинного обучения и искусственного интеллекта	Владеет навыками проведения полного цикла работ по анализу данных от сбора данных до интерпретации полученных результатов и подготовки соответствующих отчетов с помощью искусственного интеллекта	Дифференцированный зачет / Лабораторная работа

Критерии оценки

№ п/п	Наименование процедуры	Основные показатели оценки		Формы и методы контроля и оценки
	Промежуточная аттестация. Основы технологии машинного обучения и искусственного интеллекта	Зачтено	При выполнении задания выполнены все этапы задачи.	Дифференцированный зачет / Лабораторная работа
			При выполнении задания выполнены все этапы алгоритма, но отдельные части и аргументация не уточнены или частично не были представлены.	
			Не выполнены все этапы алгоритма, допущены логические ошибки и полученный результат не обоснован.	
		Незачтено	Слушатель не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки и не умеет применять базовые термины и знания при решении типовых практических задач.	

4. УСЛОВИЯ РЕАЛИЗАЦИИ ПРОГРАММЫ МОДУЛЯ

4.1 Учебно-методическое и информационное обеспечение программы:

Для эффективного освоения компетенций, формируемых учебной дисциплиной важно использование в учебном процессе активных и интерактивных форм проведения занятий.

Изучение учебной дисциплины предполагает наличие аудиторной и самостоятельной видов работ слушателей. В ходе практических занятий рассматриваются практические задачи из практики с целью наиболее полного овладения умениями и навыками.

Лекции по учебной дисциплине призваны формировать знания, предусмотренные учебной программой, и включают теоретическую базу обработки данных, на базе которой строятся прикладные аспекты.

Наряду с проработкой основной литературы (глав базового учебника) предусмотрено самостоятельное чтение дополнительной литературы (статей и других научных публикаций).

Практические занятия в малых группах и самостоятельная внеаудиторная работа направлены на выработку навыков анализа данных.

Для достижения поставленных целей преподавания дисциплины реализуются следующие средства, способы и организационные мероприятия:

- изучение теоретического материала дисциплины на лекции с использованием компьютерных технологий;
- самостоятельное изучение теоретического материала дисциплины с использованием Internet-ресурсов, информационных баз, электронных библиотек, методических разработок, специальной и научной литературы;
- закрепление теоретического материала при проведении практических занятий с использованием учебного и научного оборудования, выполнения проблемно-ориентированных, поисковых, творческих заданий.

Самостоятельная работа слушателей включает:

1. Изучение учебной литературы по курсу.
2. Решение практических ситуаций и задач
3. Изучение источников управленческой информации
4. Работу с ресурсами Интернет
5. Решение практических ситуаций в виде творческих заданий
6. Изучение практических материалов деятельности конкретных предприятий
7. Изучение статистической информации
8. Подготовку к экзамену по курсу «Анализ данных и элементы машинного обучения».

Цель самостоятельной работы - подготовка современного компетентного специалиста и формирование способностей и навыков к непрерывному самообразованию и профессиональному совершенствованию.

4.2. Содержание комплекта учебно-методических материалов.

1. Федоткин М.А. Основы прикладной теории вероятностей и статистики. — М.: Высшая школа. 2006. - 168 с.

2. Теория вероятностей и математическая статистика. Авторы: Федоткин М.А., Пройдакова Е.В.: Электронный управляемый курс. – Нижний Новгород: Нижегородский госуниверситет, 2014. (Идентификационный номер в электронном каталоге фонда электронных образовательных ресурсов ННГУ: 789Е.14.08)

3. Золотых Н.Ю.Машинное обучение. Курс лекций. Нижний Новгород: ННГУ, 2007. http://www.uic.nnov.ru/~zny/m

4. Курс «Машинное обучение» – https://www.intuit.ru/studies/courses/13844/1241/info

5. Практическая статистика для специалистов Data Science: Пер. с англ./ П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018..

б) дополнительная литература:

6. Bhasin H. (2019). Python Basics : A Self-Teaching Introduction. Dulles, Virginia: Mercury Learning & Information. http://search.ebscohost.com/login.aspx?direct=true&site=eds-live&db=edsebk&AN=1991381

7. G. James, D. Witten, T. Hastie, R. Tibshirani An Introduction to Statistical Learning with Applications in R. Springer, 2013. https://faculty.marshall.usc.edu/gareth-james/ISL/

Рус. пер.: Г. Джеймс, Д. Уиттон, Т. Хасти, Р. Тибширани Введение в статистическое обучение с примерами на языке R. ДМК Пресс, 2016.

8. I. Goodfellow, Y. Bengio, A. Courville Deep Learning The MIT Press 2016. Рус. пер.: И. Бенджио, Я. Гудфеллоу, А. Курвилль Глубокое обучение. ДМК Пресс, 2017

9. Материалы курса лекций «Основы программирования»: НОУ ИНТУИТ:

10. http://www.intuit.ru/studies/courses/2193/67/info, режим доступа – свободный.

в) программное обеспечение и Интернет-ресурсы

1. Open-source среда Spyder.

2. Веб-интерактивная вычислительна среда Jupyter Notebook (для поддержки языка Pyt Лекции и практические занятия проводятся с использованием возможностей мультимедийного класса. Использование в учебном процессе активных и интерактивных форм проведения занятий (компьютерных симуляций, деловых и ролевых игр, разбор конкретных ситуаций).

3. Anaconda: the Most World's Most Popular Data Science Platform https://www.anaconda.com/

4. scikit-learn: Machine Learning in Python https://scikit-learn.org/

5. TensorFlow: Комплексная платформа машинного обучения с открытым исходным кодом https://www.tensorflow.org/

6. pyTorch: An open source machine learning framework that accelerates the path from research prototyping to production deployment. https://pytorch.org/

4.3. Материально-технические условия реализации программы:

Материально-техническая база

№ п.п.	Наименование модуля (тем, разделов)	Материально-технические условия для реализации программ (наличие лабораторий, производственных участков и т.п. по профилю программы профессиональной переподготовки)
1.	Тема 1. 1. Введение в искусственный интеллект и машинное обучение.	Реализация дисциплины предполагает наличие: - The R Project for Statistical Computing https://www.r-project - Welcome to Python.org https://www.python.org/ - scikit-learn: machine learning in Python scikit-learn.org/. В ходе проведения занятий рекомендуется использовать компьютерные иллюстрации для поддержки различных видов занятий или других средств визуализации материала.
2.	Тема 2. Задача обучения с учителем.	
	Тема 3. Задача обучения без учителя.	
3.	Тема 4. Глубокое обучение	
4.	Практические занятия (семинары)	