МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой
Математических методов исследования операций
Азарнова Т.В.
26.05.2020

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.01 Прикладное машинное обучение на языке Python

1. Код и наименование направления подготовки/специальности:

02.04.02 Фундаментальная информатика и информационные технологии

2. Профиль подготовки/специализация: Машинное обучение и интеллектуальные информационные технологии

3. Квалификация выпускника: магистр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: Математических методов исследования операций

6. Составители программы: Каширина Ирина Леонидовна, доктор техн. наук, профессор кафедры математических методов исследования операций

7. Рекомендована: НМС факультета Прикладной математики, информатики и механики, протокол №9 от 23.05.2020

8. Учебный год: 2020/2021 **Семестр(-ы):** 1

9.Цели и задачи учебной дисциплины

Цель изучения дисциплины:

ознакомление будущих специалистов в области Data Science с процессами, алгоритмами и инструментами, относящимися к основным принципам машинного обучения

Задачи учебной дисциплины:

−сформировать теоретические знания по основам машинного обучения для построения формальных математических моделей, анализа и обработки информации по тематике исследований;

−выработать умения по практическому применению методов машинного обучения при решении прикладных задач в различных областях, в том числе при работе с большими данными;

−выработать умения и навыки использования библиотек языка Python для разработки прикладного программного обеспечения на основе алгоритмов машинного обучения.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к обязательным дисциплинам вариативной части базового цикла. Для изучения курса необходимы базовые знания информатики, линейной алгебры, математического анализа, теории вероятностей, математической статистики, методов оптимизации.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения:

Код	Название компетенции	Код(ы)	Индикатор(ы)	Планируемые результаты обучения
ПК-1	Способен проводить работы по обработке и анализу научно-технической информации результатов исследований	ПК-1.2	Анализирует и обрабатывает информацию по тематике исследований.	знать: – методы предварительной обработки данных (кодирование, стандартизация и нормализация, устранение выбросов, заполнение пропусков); – методы отбора информативных признаков; – методы классификации; – методы регрессионного анализа – методы анализа текстовых данных. уметь: – анализировать многомерные данные и преодолевать вычислительные проблемы, связанные с высокой размерностью данных; владеть (иметь навык(и)): – построения и проверки качества моделей машинного обучения; интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов;
ПК-4;	Способен разрабатывать профессионально-ориентированные программные средства и приложения на	ПК-4.2	Использует знания в области искусственного интеллекта, инженерии знаний, машинного	знать: – возможности актуальных алгоритмов машинного обучения, которые широко используются на практике, основные сферы их применения; уметь:

	основе интеллектуальных информационных технологий.		обучения для разработки прикладного программного обеспечения.	– применять методы машинного обучения при решении задач в различных прикладных областях; – использовать библиотеки языка Python для построения моделей машинного обучения; владеть (иметь навык(и)): использования библиотек языка Python для построения систем, обучающихся по прецедентам.
ПК-5	Способен совершенствовать и разрабатывать новые методы, модели, алгоритмы, технологии работы с большими данными.	ПК-5.1	Совершенствует и разрабатывает модели и алгоритмы машинного обучения для работы с большими данными.	знать: – модели и алгоритмы машинного обучения для работы с большими данными; уметь: – применять методы машинного обучения для работы с большими данными;

12. Объем дисциплины в зачетных единицах/час.(*в соответствии с учебным планом)* — 5/180.

Форма промежуточной аттестации *экзамен*

13. Трудоемкость по видам учебной работы

Вид учебной работы		Трудоемкость	
		Всего	По семестрам
			№ семестра
Контактная работа			
в том числе:	лекции	32	32
	практические		
	лабораторные	32	32
Самостоятельная работа		80	80
Промежуточная аттестация (экзамен)		36	36
Итого:		180	180

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК *
		1. Лекции	
1.1	Введение в машинное обучение. Основные определения и постановки задач.	Основные этапы решения задачи анализа данных. Примеры прикладных задач. Виды обучения: с учителем, без учителя, с подкреплением. Основные типы задач: задача классификации, задача регрессии, задача кластеризации, задача прогнозирования, задача ранжирования. Основные проблемы машинного обучения: недостаточный объем обучающей выборки, пропуски в	Машинное обучение на языке Python

		данных,переобучение	
1.2	Решение задачи регрессии	Метод наименьших квадратов. Измерение ошибки в задачах регрессии (MSE, $RMSE$, MAE, R^2). Многомерная регрессия, проблема мультиколлинеарности. Регрессия, линейная по параметрам, полиномиальная регрессия. Решение проблемы переобучения: L1- регуляризация (Lasso), L2- Регуляризация (гребневая регрессия), эластичная сеть. Настройка гиперпараметров алгоритма с помощью n-кратной перекрестной проверки.	Машинное обучение на языке Python
1.3	Решение задачи классификации.	Линейная модель классификации. Логистическая регрессия как линейный классификатор. Функция потерь (ошибок классификации). Логистическая функция потерь с учетом L2-регуляризации. Использование полиномиальных признаков для нелинейного разделения. Confusion matrix (матрица ошибок классификации). Метрики качества классификации: accuracy (доля правильных ответов), precision (точность), recall (полнота), F1-мера. AUC-ROC –площадь под кривой ошибок. Метрическая классификация - метод ближайших соседей (kNN). Использование наивной байесовской модели для классификации	Машинное обучение на языке Python
1.4.	Древовидные модели: деревья решений, случайный лес	Этапы построения дерева решений, выбор критерия точности прогноза. типа ветвления. Метрики ветвления на основе прироста информации (алгоритм ID3), нормализованного прироста информации (алгоритм C4.5), индекса Джини (алгоритм CART). Правила разбиения. Механизм отсечения дерева. Критерии останова алгоритма (минимальное число объектов, при котором выполняется расщепление , минимальное число объектов в листьях, максимальная глубина деревьев. Переобучение решающих деревьев. Случайный лес. Обучение случайного леса. Достоинства и недостатки случайного леса	Машинное обучение на языке Python
1.5	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	Бэггинг, случайный лес как пример бэггинга. Бэггинг линейных классификаторов. Бустинг. Adaboost для ансамбля из простых деревьев (пней). Сравнение результатов бустинга для слабых и сильных моделей. Градиентный бустинг. Градиентный бустинг в задаче регрессии. Градиентный бустинг в задаче классификации. Градиентный бустинг над деревьями.	Машинное обучение на языке Python
1.6	Анализ текстовых данных	Представление текстовых данных в виде «мешка слов». Стоп-слова. Масштабирование данных с помощью tf-idf. Модель «мешка слов» для последовательностей из нескольких слов (n-грамм) Продвинутая токенизация, стемминг и лемматизация Моделирование тем и кластеризация документов. Латентное размещение Дирихле	Машинное обучение на языке Python
2. Лабораторные занятия			
2.1	Обзор основных необходимых библиотек языка Python	Библиотека NumPy для оптимизированных вычислений над массивами данных. Введение в массивы библиотеки NumPy. Выполнение вычислений над массивами библиотеки NumPy, универсальные функции Операции над данными в библиотеке Pandas. Обработка отсутствующих данных. Агрегирование и группировка. Визуализация с помощью библиотеки Matplotlib. Линейные графики, диаграммы рассеяния, гистограммы, трехмерные графики. Знакомство с	Машинное обучение на языке Python

		библиотекой машинного обучения Scikit-Learn. Гиперпараметры и проверка качества модели	
2.2	Построение и отбор признаков	Извлечение признаков (Feature Extraction). Преобразования признаков (Feature transformations): кодирование нечисловых данных, нормировка и калибровка, заполнение пропусков Выбор признаков (Feature selection): статистические подходы, визуализация, отбор с использованием моделей	Машинное обучение на языке Python
2.3.	Решение задачи регрессии	Разбор примера построения модели линейной регрессии для задачи предсказания велосипедного трафика Отбор и кодирование признаков. Визуальное сравнение общего и предсказанного моделью трафика. Проверка качества	Машинное обучение на языке Python
2.4	Решение задачи классификации.	Разбор примера построения модели логистической регрессии для задачи предсказания оттока клиентов мобильного оператора. Отбор и кодирование признаков. Проверка качества модели с помощью перекрёстной проверки.	Машинное обучение на языке Python
2.5	Древовидные модели: деревья решений, случайный лес	Разбор примера построения модели дерева решений для задачи предсказания исхода футбольного матча. Анализ деревьев, полученных при использовании различных метрик. Построение модели случайного леса на примере задачи кредитного скоринга. Кодирование признаков и заполнение пропущенных данных.	Машинное обучение на языке Python
2.6	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	Разбор примера построения модели градиентного бустинга для задачи распознавания рукописных цифр из библиотеки MNIST.	Машинное обучение на языке Python
2.7	Анализ текстовых данных	Разбор примера построения модели анализа текстовых данных для задачи определения тональности киноотзывов.	Машинное обучение на языке Python

.

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практические	Лабораторные	Самостоятельная работа	Всего
1	Введение в машинное обучение. Основные определения и постановки задач.	4		4	10	18
2	Обзор основных необходимых библиотек языка Python	4		4	10	18
3	Построение и отбор признаков	4		4	10	18
4	Решение задачи регрессии	4		4	10	18
5	Решение задачи классификации.	4		4	10	18
6	Древовидные модели: деревья решений, случайный лес	4		4	10	18
7	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	4		4	10	18
8	Анализ текстовых данных	4		4	10	18
	Итого	32		32	80	

14. Методические указания для обучающихся по освоению дисциплины

Для лучшего усвоения материала студентам рекомендуется домашняя работа с конспектами лекций, презентациями, выполнение практических заданий для самостоятельной работы, выполнение лабораторных работ, использование рекомендованной литературы и методических материалов, в том числе размещенных на странице курса «Машинное обучение» на портале «Электронный университет ВГУ» https://edu.vsu.ru/course/view.php?id=3579, автор Каширина И.Л.
В рамках общего объема часов, отведенных для изучения дисциплины, предусматривается выполнение следующих видов самостоятельных работ студентов (СРС): изучение теоретического материала, написание программ по темам, изученным на лекционных и практических занятия
При использовании дистанционных образовательных технологий и электронного обучения выполнять все указания преподавателей по работе на LMS-платформе, своевременно подключаться к online-занятиям, соблюдать рекомендации по организации самостоятельной работы

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины *(список литературы оформляется в соответствии с требованиями ГОСТ и используется общая сквозная нумерация для всех видов источников)*

а) основная литература:

№ п/п	Источник
1	Рашка, С. Python и машинное обучение: [Электронный ресурс] : руководство / С. Рашка ; пер. с англ. Логунова А.В.. — Электрон. дан. — Москва : ДМК Пресс, 2017. — 418 с. — Режим доступа: https://e.lanbook.com/book/100905
2	Коэльо, Л.П. Построение систем машинного обучения на языке Python [Электронный ресурс] / Л.П. Коэльо, В. Ричарт ; пер. с англ. Слинкин А. А.. — Электрон. дан. — Москва : ДМК Пресс, 2016. — 302 с. — Режим доступа: https://e.lanbook.com/book/82818
3	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных [Электронный ресурс] / П. Флах. — Электрон. дан. — Москва : ДМК Пресс, 2015. — 400 с. — Режим доступа: https://e.lanbook.com/book/69955

б) дополнительная литература:

№ п/п	Источник
4	Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с. Материалы к книге: https://github.com/jakevdp/PythonDataScienceHandbook
5	Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. -СПб.: Питер, 2017. -336 с.
6	Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. -СПб.: Питер, 2017. -336 с.:
7	Шарден, Б. Крупномасштабное машинное обучение вместе с Python [Электронный ресурс] : учебное пособие / Б. Шарден, Л. Массарон, А. Боскетти ; пер. с анг. А. В. Логунова. — Электрон. дан. — Москва : ДМК Пресс, 2018. — 358 с. — Режим доступа: https://e.lanbook.com/book/10583
8	Вьюгин, В.В. Математические основы машинного обучения и прогнозирования [Электронный ресурс] : учебное пособие / В.В. Вьюгин. — Электрон. дан. — Москва : МЦНМО, 2013. — 304 с. — Режим доступа: https://e.lanbook.com/book/56397
9	Кук, Д. Машинное обучение с использованием библиотеки H2O [Электронный ресурс] / Д. Кук ; пер. с англ. Огурцова А.Б.. — Электрон. дан. — Москва : ДМК Пресс, 2018. — 250 с. — Режим доступа: https://e.lanbook.com/book/97353

в)информационные электронно-образовательные ресурсы (официальные ресурсы интернет)**:**

№ п/п	Ресурс
10	А.Мюллер, С.Гвидо - Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными – 2017 электронный ресурс свободного доступа: https://owlweb.ru/wp-content/uploads/2017/06/a.myuller-s.gvido-vvedenie-v-mashinnoe-obuchenie-s-pomoshhyu-python.-rukovodstvo-dlya-specialistov-po-rabote-s-dannymi-2017.compressed-1.pdf материалы к книге: https://github.com/amueller/introduction_to_ml_with_python
11	Машинное обучение (курс лекций, К.В.Воронцов) http://www.machinelearning.ru/wiki/index.php?title=Машинное обучение (курс лекций, К.В.Воронцов)
12	https://www.kaggle.com/ - онлайн платформа для проектов в области науки о данных
13	UCI Machine Learning Repository — репозиторий наборов данных для машинного обучения

	- http://archive.ics.uci.edu/ml/
14	Ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. - http://machinelearning.ru
15	Открытый курс машинного обучения https://habr.com/company/ods/blog/322626/
16	Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. *https://edu.vsu.ru/course/view.php?id=3579*

16. Перечень учебно-методического обеспечения для самостоятельной работы *(учебно-методические рекомендации, пособия, задачники, методические указания по выполнению практических (контрольных), курсовых работ и др.)*

№ п/п	Источник
1	Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. https://edu.vsu.ru/course/view.php?id=3579
2	Бринк Х., Ричардс Д., Феверолф М. Машинное обучение. -СПб.: Питер, 2017. -336 с.: Материалы к книге: *https://github.com/brinkar/real-world-machine-learning*
3	Каширина И.Л. Искусственные нейронные сети, Воронеж, Из-во ВГУ, 2014-96 с.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий. Для организации самостоятельной работы обучающихся используется онлайн-курс электронный учебный Курс «Машинное обучение на языке Python» на портале «Электронный университет ВГУ», автор Каширина И.Л. https://edu.vsu.ru/course/view.php?id=3579 , размещенный на платформе Электронного университета ВГУ (LMS moodle), а также Интернет-ресурсы, приведенные в п.15в.

18. Материально-техническое обеспечение дисциплины:

Лекционная аудитория должна быть оснащенной современным компьютером с подключенным к нему проектором с видеотерминала на настенный экран. Практические и лабораторные занятия должны проводиться в специализированной аудитории, оснащенной современными персональными компьютерами и программным обеспечением в соответствии с тематикой изучаемого материала - Python 3 с подключенными библиотеками (дистрибутив Anaconda).

Перечень специализированных лабораторий:

Лаборатория машинного обучения (корпус 1, ауд. 407п)

Компьютер в составе (16 шт.): системный блок: процессор Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz, оператив-ная память 16 Гб, SSD 256 Гб, HDD 1Тб, ви-деокарта NVIDIA GeForce GTX 1080 Ti; мо-нитор DELL S2419HN; Компьютер в составе (1 шт.):
системный блок: процессор Intel(R) Core(TM) i7-7800X CPU @ 3.50GHz, опера-тивная память 96 Гб, SSD 1Тб, HDD 4Тб, ви-деокарта NVIDIA GeForce RTX 2080 Ti (2 шт.); монитор DELL S2419HN; Источник бесперебойного питания APC Back-UPS BV1000I-GR, line-interactive, мощ-ность:1000ВА, 600Вт (16 шт.); Источник бесперебойного питания Legrand KEOR LINE RT 1500ВА (1 шт.); Коммутатор HP 2530-24G Switch (Managed, 24*10/100/1000 + 4 SFP, 19"); Интерактивная доска SMART SBM685 (87 дюймов, ПО SMART SLS) с пассивным лот-ком; Проектор Vivitek DH758UST (ультракорот-кофокусный, DLP, Full HD 1080p (1920 x 1080) , 3500 ANS, 10000:1, полная поддержка 3D)

Лаборатория искусственного интеллекта (корпус 1, ауд. 124)

Компьютер в составе (17 шт.): системный блок: процессор AMD Ryzen 7 3800X 8-Core Processor, оперативная память 32Гб, HDD 1Тб, SSD 256Гб, видеокарта NVIDIA GeForce GTX 1050; монитор:

Dell S2419H; Интерактивная доска SMART SBM685 (87 дюймов); Мультимедиа-проектор Vivitek ультрако-роткофокусный; Источник бесперебойного питания Legrand Keor SPX 1000 BA IEC C13 (16 шт.); Источник бесперебойного питания Legrand Keor Line RT 1000 BA (1 шт.); Коммутатор HP 2530-48G Switch (1 шт.)

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименование раздела дисциплины (модуля)	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1.	Введение в машинное обучение. Основные определения и постановки задач.	ПК-1	ПК-1.2	Тест 1
2.	Обзор основных необходимых библиотек языка Python	ПК-1	ПК-1.2	Тест 2
3	Построение и отбор признаков	ПК-1	ПК-1.2	Тест 3
4	Решение задачи регрессии	ПК-4	ПК-4.2	Задание для лабораторной работы 1
5	Решение задачи классификации.	ПК-4	ПК-4.2	Задание для лабораторной работы 2
6	Древовидные модели: деревья решений, случайный лес	ПК-4	ПК-4.2	Задание для лабораторной работы 3
7	Ансамбли моделей Бэггинг, бустинг, градиентный бустинг	ПК-5	ПК-5.1	Тест 4
8	Анализ текстовых данных	ПК-5	ПК-5.1	Задание для лабораторной работы 4
Промежуточная аттестация форма контроля - экзамен				*Перечень вопросов* Практическое задание

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

Тестовые задания, Лабораторные работы, Устный опрос

Тестовые задания

Тест № 1

Задание 1:

Что такое линейная регрессия?

Вариант 1 это функциональная зависимость, которая позволяет прогнозировать изменения непрерывных числовых параметров;

Вариант 2 модель зависимости непрерывной переменной y от объясняющих ее факторов, в которой функция зависимости является линейной

Вариант 3 модель зависимости дискретной переменной y от объясняющих ее факторов, в которой функция зависимости является линейной

Задание 2:
Основная характеристика задачи бинарной классификации:

Вариант 1 классификация осуществляется по одному признаку

Вариант 2 зависимая переменная может принимать только два значения

Вариант 3 классификация осуществляется по двум признакам

Задание 3:
Классификация относится к стратегии:

Вариант 1 обучения без учителя

Вариант 2 обучения с учителем

Вариант 3 оба ответа неверны

Задание4:
Какое выражение прибавляется к функции ошибок про L_2 регуляризации?

Вариант 1 $\lambda \sum_{i=1}^{N} w_i^2$

Вариант 2 $\lambda \sum_{i=1}^{N} |w_i| \lambda \sum_{i=1}^{N} |w_i|$

Вариант 3 $\frac{\lambda}{2} \sum_{i=1}^{N} w_i^2 + \frac{\lambda}{2} \sum_{i=1}^{N} |w_i|$

Задание5:
Явление переобучения характеризуется …

Вариант 1 чрезмерно точным соответствием модели конкретному набору обучающих примеров, при котором модель теряет способность к обобщению

Вариант 2 возникновением, в случае слишком долгого обучения, недостаточного числа обучающих примеров или слишком сложной структуры модели

Вариант 3 возникновением, в случае слишком долгого обучения, слишком сложной структуры модели

Задание6:
Задача классификации сводится к …

Вариант 1 нахождению частых зависимостей между объектами или событиями;

Вариант 2 определению класса объекта по его характеристикам;

Вариант 3 определению по известным характеристиками объекта значение некоторого его параметра;

Вариант 4 поиску независимых групп и их характеристик в всем множестве анализируемых данных.

Задание7:
Множество примеров, используемое для конструирования модели, называется...

Вариант 1 обучающим множеством

Вариант 2 тестовым множеством

Вариант 3 валидационным множеством

Задание8:
Дана матрица ошибок алгоритма классификации. Вычислите precision

	y=1	y=0
$\hat{y} = 1$	15	10
$\hat{y} = 0$	5	20

Вариант 1 0.7

Вариант 2 0,6

Вариант 3 0,75

Задание9:

В задаче определения спама recall - это…

Вариант 1 доля спама среди писем, распознанных алгоритмом как спам

Вариант 2 доля не спама среди писем, распознанных алгоритмом как спам

Вариант 3 доля писем, распознанных алгоритмом как спам, среди спама

Вариант 4 доля писем, распознанных алгоритмом как не спам, среди спама

Задание10:

Какие функции ошибок используются в задачах регрессии?

Вариант 1 $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

Вариант 2 $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

Вариант 3 $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$

Вариант 4 $E = \frac{1}{N}\sum_{i=1}^{N}[y_i \neq \hat{y}_i]$

Перечень заданий для лабораторных работ

Лабораторная работа № 1 (по теме линейная регрессия)

1) Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2) Обучите, а затем провалидируйте на тестовых данных следующие модели, используя в качестве метрики качества R^2, предварительно отмасштабировав данные

- LinearRegression;
- Lasso с коэффициентом регуляризации, равным 0.01.

3) Проанализируйте качество получившихся моделей и сравните количество строго нулевых весов в них.

Лабораторная работа № 2 (по теме логистическая регрессия)

1) Разбейте предоставленный Вам преподавателем набор данных на обучающую и тестовую части в соотношении 8:2.
2) Проведите предобработку данных: заполнение пропусков, кодирование, масштабирование
3). Обучите, а затем провалидируйте на тестовых данных модель логистической регрессии
4) Вычислите значения метрик: recall, presision, F1-мера, AUC-ROC. Постройте ROC-кривую.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств:

Теоретические воросы, практические задания

Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний и практические) задания, позволяющие оценить степень сформированности умений и навыков.

Для оценивания результатов обучения на экзамене используются следующие показатели:
1) знание учебного материала и владение понятийным аппаратом теории машинного обучения;

2) умение анализировать многомерные данные и преодолевать вычислительные проблемы, связанные с высокой размерностью данных;
3) умение применять методы машинного обучения при решении задач в различных прикладных областях; ;
5) владение навыками использования библиотек языка Python для построения систем, обучающихся по прецедентам
6) владение навыками построения и проверки качества моделей машинного обучения;
7) владение навыками интерпретации полученных результатов в терминах прикладной области с целью получения новых знаний и выводов.

Перечень вопросов к экзамену

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
2. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
3. Метрики качества алгоритмов регрессии и классификации.
4. Линейная регрессия. Простая многомерная регрессия. Регрессия с полиномиальными признаками. Методы регуляризации: Ridge, Lasso, ElasticNet.
5. Логистическая регрессия.
6. Деревья решений. Методы построения деревьев. Их регуляризация.
7. Композиции алгоритмов. Разложение ошибки на смещение и разброс.
8. Случайный лес, его особенности.
9. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов.
10. Анализ текстов. Масштабирование данных с помощью tf-idf. Модель «мешка слов» для n-грамм.

Практическое задание № 1 (по теме Линейная регрессия)

Ответьте на вопросы о данных по авиарейсам в США за январь-апрель 2008 года.
По ссылке расположены Данные и их описание
1) Считайте выборку из файла при помощи функции pd.read_csv и ответьте на следующие вопросы:

- Имеются ли в данных пропущенные значения?
- Сколько всего пропущенных элементов в таблице "объект-признак"?
- Сколько объектов имеют хотя бы один пропуск?
- Сколько признаков имеют хотя бы одно пропущенное значение?

2) Преобразуйте каждый признак FeatureName из указанных в пару новых признаков FeatureName_Hour, FeatureName_Minute, разделив каждое из значений на часы и минуты. Не забудьте при этом исключить исходный признак из выборки. В случае, если значение признака отсутствует, значения двух новых признаков, его заменяющих, также должны отсутствовать.
3) Некоторые из признаков, отличных от целевой переменной, могут оказывать чересчур значимое влияние на прогноз, поскольку по своему смыслу содержат большую долю информации о значении целевой переменной. Изучите описание датасета и исключите признаки, сильно коррелирующие с ответами. Ваш выбор признаков для исключения из выборки обоснуйте.
4) Приведите данные к виду, пригодному для обучения линейных моделей. Для этого вещественные признаки надо отмасштабировать, а категориальные — привести к числовому виду. Также надо устранить пропуски в данных. Реализуйте функцию transform_data, которая принимает на вход DataFrame с признаками и выполняет следующие шаги:

- Замена пропущенных значений на нули для вещественных признаков и на строки 'nan' для категориальных.
- Масштабирование вещественных признаков с помощью StandardScaler.
- One-hot-кодирование категориальных признаков с помощью DictVectorizer или функции pd.get_dummies.

Метод должен возвращать преобразованный DataFrame, который должна состоять из масштабированных вещественных признаков и закодированных категориальных (исходные признаки должны быть исключены из выборки).
5) Разбейте выборку и вектор целевой переменной на обучение и контроль в отношении 70/30 (для этого можно использовать функцию train_test_split).
6) Обучите линейную регрессию на 1000 объектах из обучающей выборки и выведите значения MSE и R^2 и на этой подвыборке и контрольной выборке. Проинтерпретируйте полученный результат — насколько качественные прогнозы строит полученная модель? Какие проблемы наблюдаются в модели?

7) Обучите линейные регрессии с L1- и L2-регуляризатором, подобрав лучшее значение параметра регуляризации при помощи кросс-валидации. Выведите значения и на обучающей и контрольной выборках. Удалось ли решить указанные вами ранее проблемы?

Практическое задание № 3 (по теме Анализ текстов)

Практическое задание 2 посвящено работе с текстовыми данными и категориальными признаками и задачам бинарной классификации.
В рамках данного задания нужно решить задачу бинарной классификации для предсказания уровня заработной платы по тексту объявления о вакансии на примере набора данных с соревнования на Kaggle. Данные доступны по ссылке.
1) Разбейте получившуюся выборку на обучающую и контрольную в соотношении 70/30
2) Создайте текстовое описание объектов обучающей и контрольной выборок, объединив значения всех признаков каждого объекта выборки через символы пробела. После этого получите признаковое описание объектов, осуществив векторизацию получившихся текстов при помощи CountVectorizer, обучив его на обучающей выборке и применив на тестовой.
3) Обучите логистическую регрессию из модуля sklearn с параметрами по умолчанию на обучающей выборке:
4) Вычислите значения ROC-AUC, F-меры, а также постройте матрицу ошибок на тестовой выборке.
5) Отсортируйте веса признаков для модели. Какие слова из встречающихся в выборке имеют наибольшее/наименьшее влияние на значение целевой переменной? Проинтерпретируйте полученный результат.
6) Создайте текстовое описание объектов обучающей и контрольной выборок, объединив значения всех признаков каждого объекта выборки через символы пробела. После этого получите признаковое описание объектов, вычислив вектор tf-idf для каждого объекта помощи TfidfVectorizer, обучив его на обучающей выборке и применив на тестовой.
7) Заново обучите модель
8) Вычислите значения ROC-AUC, F-меры, а также постройте матрицу ошибок на контрольной выборке..
9). Сравните значения метрик из п. 8 со значениями, полученными в п. 4, и сравните соответствующие модели по качеству из работы.
11. Отсортируйте веса признаков для модели логистической регрессии из scikit-learn, полученной в п. 7. Какие слова из встречающихся в выборке имеют наибольшее/наименьшее влияние на значение целевой переменной? Проинтерпретируйте полученный результат.

Для оценивания результатов обучения на экзамене используется 4-балльная шала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».
Соотношение показателей, критериев и шкалы оценивания результатов обучения.

Критерии оценивания	Шкала оценок
Обучающийся в полной мере владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), сдал все практические и лабораторные работы, среднее количество правильных ответов на вопросы тестов превышает 80%.	*Отлично*
Обучающийся владеет понятийным аппаратом данной области науки (теоретическими основами дисциплины), но не сдал одну практическую или лабораторную работу, среднее количество правильных ответов на вопросы тестов находится в диапазоне70-80%.	*Хорошо*
Обучающийся демострирует неуверенное владение понятийным аппаратом данной области науки (теоретическими основами дисциплины), не сдал две практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов находится в диапазоне60-70%.	*Удовлетвори-тельно*
Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки, не сдал более двух практических или лабораторных работы, среднее количество правильных ответов на вопросы тестов менее 70%.	*Неудовлетвори-тельно*