Федеральное государственное бюджетное образовательное учреждение
высшего образования
Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

УТВЕРЖДАЮ

декан факультета вычислительной
математики и кибернетики

_____/И.А. Соколов /

«14» октября 2021г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины (модуля):

Прикладные задачи анализа данных

Уровень высшего образования:

магистратура

Направление подготовки / специальность:

01.04.02 "Прикладная математика и информатика" (3++)

Направленность (профиль) ОПОП:

Искусственный интеллект в кибербезопасности

Форма обучения:

очная

Рабочая программа рассмотрена и утверждена
на заседании Ученого совета факультета ВМК
(протокол №4, от 29 сентября 2021 года)

Москва 2021

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки 01.04.02 "Прикладная математика и информатика" программы магистратуры в редакции приказа МГУ от 21 декабря 2021 года No 1404.

1. Место дисциплины (модуля) в структуре ОПОП ВО:
Дисциплина (модуль) относится к части дисциплин основной профессиональной образовательной программы, формируемых участниками образовательных отношений.

2. Входные требования для освоения дисциплины (модуля), предварительные условия:
Учащиеся должны владеть знаниями по теории вероятностей, математической статистике, машинному обучению в объеме, соответствующем основным образовательным программам бакалавриата по укрупненным группам направлений и специальностей 01.00.00 «Математика и механика», 02.00.00 «Компьютерные и информационные науки» и другим направлениям подготовки бакалавриата.

3. Результаты обучения по дисциплине (модулю), соотнесенные с требуемыми компетенциями выпускников.

Планируемые результаты обучения по дисциплине (модулю)		
Содержание и код компетенции.	**Индикатор (показатель) достижения компетенции**	**Планируемые результаты обучения по дисциплине, сопряженные с индикаторами достижения компетенций**
ПК-8. Способен разрабатывать и модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта с учетом требований информационной безопасности в различных предметных областях	ПК-8.2. Модернизирует программное и аппаратное обеспечение технологий и систем искусственного интеллекта для решения профессиональных задач с учетом требований информационной безопасности в различных предметных областях	ПК-8.2. З-1. Знает особенности модернизации программного и аппаратного обеспечения технологий и систем искусственного интеллекта для решения профессиональных задач в различных предметных областях ПК-8.2. У-1. Умеет модернизировать программное и аппаратное обеспечение технологий и систем искусственного интеллекта с учетом требований информационной безопасности для решения профессиональных задач в различных предметных областях

4. Объем дисциплины (модуля) составляет 3 з.е., в том числе 72 академических часа, отведенных на контактную работу обучающихся с преподавателем, 36 академических часов на самостоятельную работу обучающихся.

5. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий:

В курсе дается обзор современных задач анализа данных и методов их решения, включая анализ соцсетей, текстов, построение ансамблей алгоритмов, в том числе с помощью алгебраического подхода к решению задач классификации.

Наименование и краткое содержание разделов и тем дисциплины, форма промежуточной аттестации по дисциплине	**Всего (часы)**	В том числе								
		Контактная работа (работа во взаимодействии с преподавателем), часы из них						**Самостоятельная работа учащегося, часы** из них		
		Занятия лекционного типа	Практические занятия	Групповые консультации	Индивидуальные консультации	Учебные занятия, направленные на проведение текущего контроля успеваемости (практические контрольные занятия) и итогового контроля (экзамен).	**Всего**	Выполнение домашних заданий	Подготовка рефератов и т.п..	**Всего**
Тема 1. Прикладные программные системы для анализа данных Система для анализа данных Matlab Язык программирования Python, библиотеки numpy, scipy, scikit-learn, pandas Язык программирования R	**22**	6	6	-	-	2	**14**	6	2	**8**
Тема 2. Математические основы анализа	**30**	8	8	-	-	4	**20**	6	4	**10**

данных. Оценка среднего и вероятности Функционалы качества и ошибки, их оптимизация Теория нечётких множеств Пост-троечные последовательности Спектральная теория графов										
Тема 3. Прикладные задачи анализа данных и методы их решения. Исследование социальных сетей Анализ текстов, Случайные леса Линейные модели алгоритмов Категориальные признаки k ближайших соседей, настройка комбинаций алгоритмов	**30**	8	8	-	-	4	**20**	6	4	**10**
Тема 4. Алгебраический подход к анализу данных. Модели алгоритмов классификации (распознава-	**22**	6	6	-	-	2	**14**	6	2	**8**

ния образов) Операции над алгоритмами, алгебра над алгоритмами Описание алгебраических замыканий. Критерии разрешимости и корректности.										
Итоговая аттестация – экзамен	**40**	-	-	2	-	2	**4**	36	-	
Итого	**108**						**72**			**36**

6. Фонд оценочных средств (ФОС, оценочные и методические материалы) для оценивания результатов обучения по дисциплине (модулю).

6.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости, критерии и шкалы оценивания

Примерные практические контрольные задания для текущего контроля успеваемости.

ПКЗ ТК1 (по подтеме «нечёткие множества») . При каких $k \in [0,+\infty)$ нечёткое множество A с функцией принадлежности $\mu_A(x) = (1-x)^k$ на $[0,1]$ будет выпуклым?

ПКЗ ТК2 (по подтеме «нечёткие множества») . Всегда ли заострение выпуклого нечёткого множества выпуклое? Если да – доказать.

ПКЗ ТК3 (по подтеме «функционалы качества») . Пусть в задаче с двумя непересекающимися классами требуется построить алгоритм, который выдаёт значения на отрезке $[0,1]$. Ошибка измеряется по формуле $(y_i - a_i)^2 \cdot \begin{cases} 0.9, & y_i = 1, \\ 0.1, & y_i = 0, \end{cases}$

Какой константный алгоритм оптимален, если вероятность появления класса 1 – p.

ПКЗ ТК4 (по подтеме «функционалы качества») . Выписать формулу для коррекции весов в SGD, если функция ошибки LOG LOSS, а метод решения – логистическая регрессия.

ПКЗ ТК5 (проект – решение реальной задачи). Дана статистика посещений клиентов магазинов сети супермаркетов (даётся ссылка на интернет-ресурс): каждый клиент задаётся номером (id), для каждого клиента известны даны визитов до 31.03.2015. Предсказать для каждого клиента дату следующего визита. Функционал качества – доля верных прогнозов. Алгоритм написать на языке Python, подготовив notebook и отчёт по выполнению задания. Задание проводится на платформе для решения задач анализа данных https://inclass.kaggle.com/.

6.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации по дисциплине, критерии и шкалы оценивания

Список тем вопросов для экзамена.

1. Оценка среднего, оценка вероятности, оценка плотности. Весовые схемы
Проблема оценки среднего, выбросы, разные целевые функционалы, оценка минимального контраста, среднее по Колмогорову, SMAPE-минимизация, двухэтапные алгоритмы и их настройка, пересчёт вероятности и прямая оценка, введение весовых схем, устойчивость весовых схем, ансамблирование, непараметрическое восстановление плотности, весовые схемы при оценке плотности.
2. Система для анализа данных Matlab
3. Искусство визуализации
Признаки в задачах классификации/регрессии, выделение групп признаков, что можно увидеть в данных, оценка признаков и фолдов, деформация ответов, устойчивость закономерностей, профили лет (в прогнозировании вр.рядов), плотности, оценка качества признаков с помощью RF и удалений, результаты алгоритмов и их линейные комбинации, ручная деформация пространств, визуализация и сглаживание плотностей, построение профилей. Что надо знать о признаках. Визуализация по-вертикали и по-горизонтали. Шумы и шумовые признаки. Определение свойств признака (категориальность, группы значений и т.п.).
4. Функционалы качества и ошибки
MAE, RMSE, SMAPE, MAP, MRAE, REL_MAE, PB, нормированные ошибки, несимметричные ошибки, ошибки с точностью до порога, MCE, точность (Precision), полнота, специфичность, False Positive Rate, F1-мера, AUROC, GINI, Log Loss, Hamming Loss, MAP, Discounted Cumulative Gain (DCG), Quadratic Weighted Kappa, редакторское расстояние. Матожидание ошибок. Генерация признаков с помощью функций ошибок. Confusion matrix.
5. Минимизация ошибок
Построение дерева, максимизирующего ROC AUC, получение интервальных значений целевого признака, деформация для Root Mean Square Percentage Error, оптимизация log_loss для логистической регрессии, линейной рересии, оптимизация СКО для логистической регрессии, линейной регрессии.
6. Линейные алгоритмы
Персептронный алгоритм, режимы обучения, концепция поощрение-наказание, концепция минимизации функционала, линейная регрессия, SGD, delta-bar-delta, хэширование признаков, регуляризация, обобщения регрессии, прогноз раскупаемости, прогноз методом kNN, прогноз

линейным оператором, линейный алгоритм над SVD, признаковое прогнозирование спроса, профили товаров, сезонность, LibSVM, LibLinear.

7. Анализ текстов: классификация и регрессия

Этапы работы с текстом, токенизация, стоп-слова, векторное представление документа, n-граммы, стемминг, алгоритм Портера, TF*IDF, оценки качества (точность, полнота, F-мера), классификация спама, Local and Global Consistency, этапные алгоритмы, устойчивые признаки, иерархическая классификация текстов, основные методы (Роше, kNN, SVM), приведение к шаблону, обнаружение оскорблений, распределение по топикам (задача со многими классами), блендинг алгоритмов, фонетические алгоритмы.

8. Случайные леса

Универсальные методы анализа данных, бэггинг и бустинг, построение одного дерева, OOB(out of bag)-проверка, параметры случайного леса (random forest: mtry, nodesize, samplesize) и их настройка, рейтинг признаков (importance, %IncMSE, IncNodePurity, Boruta, ACE). Программирование случайного леса. Области устойчивости функционалов. Искусство генерации признаков: географические и временные признаки. Концепция чёрного ящика на примере GBM. Настройка параметров GBM, суммирование. Нестандартные функционалы и настройка на них. Калибровка ответов алгоритмов. Сведение задачи рекомендации к регрессии. Критерии расщепления.

9. Категориальные признаки

Терминология, задачи, one-hot-кодировки, конъюнкции признаков, хранение в sparse-матрицах, линейные методы, байесовские алгоритмы, сингулярные разложения, методы, основанные на близости (kNN+ABO), тензорные разложения, случайные кодировки, кодировки относительно вещественных признаков, SVD-кодировки, ансамбли алгоритмов, факторизационные машины.

10. k ближайших соседей, настройка комбинаций алгоритмов

Сглаживание функционалов качества при использовании весовых схем. Ограничение методов типа kNN (тренд, некорректность метрики). Примитивная настройка линейных комбинаций алгоритмов и метрик. Технология LENKOR (синтез близостей, составление комбинации близостей, настройка коэффициентов, добавление нелинейностей). Подробный разбор задачи детектирования оскорблений. Задачи «определение качества фотографии по метаданным», «предсказывание успешности выполнения гранта», «рекомендация видеолекций для просмотра». Деформация ответов, теоремы Колмогорова, Горбаня и Пинкуса, алгебраический подход к коррекции и его правильное применение на практике.

11. Пост-троечные последовательности

Проблема обезличивания информации, построение рекомендательной системы для холодного старта в задаче рекомендации видеолекций для просмотра.

12. Теория нечётких множеств

Характеристическая функция, нечёткое множество, Т-нормы, Т-конормы, декомпозиция множеств, расстояния между нечёткими множествами, оценка нечёткости (энтропийный, метрический, аксиоматический подходы), нечёткие отношения, транзитивные замыкания, принцип обобщения, нечёткие числа, модификаторы, проблемы формализации, полное ортогональное семантического пространство (ПОСП), степень нечёткости ПОСП.

13. Анализ социальных сетей

Примеры соцсетей, динамические графы, приложения анализа социальных сетей, понятие сложной сети (complex network), безмасштабные сети (scale-free), модель малого мира (small world), коэффициент кластеризации (clustering coefficient), признаковые пространства для графов, формализация сходства и важности вершин, разные виды центральности (centrality).
Прогнозирование появления ребра в динамическом графе (Link Prediction Problem), коэффициенты Жаккара, Адамик/Адара, Katz, PageRank, признаки для рёбер.
Выделение сообществ в графах (Community detection): переборные методы, Edge betweenness, модулярность, Multilevel, распространение меток, Walktrap, спектральная теория графов, разложения матриц графа. Определение кругов в эго-подграфах графа социальной сети (задача, данные, редакторское расстояние),

14. Спектральная теория графов

15. Алгебраический подход

Универсальные модели для классификации (распознавания образов), операции над алгоритмами, алгебра над алгоритмами, алгебраические замыкания, корректность модели алгоритмов, критерии корректности и разрешимости.

Пример экзаменационного билета

1. Что такое "оценка минимального контраста"?

2. Напишите функцию (на выбор: Python/R/Matlab) построковой нормировки матрицы данных: каждый элемент делится на сумму элементов в строке.

3. В задаче поиска на запрос выведено 6 документов, из которых половина релевантных, чему равна минимальное значение ap@n?

4. Пусть в задаче с двумя непересекающимися классами требуется построить алгоритм, который выдаёт значения на отрезке [0,1]. Ошибка измеряется по формуле

$$\begin{cases} (y_i - a_i)^2, & y_i = 1, \\ |\, y_i - a_i \,|, & y_i = 0, \end{cases}$$

Какой константный алгоритм оптимален, если вероятность появления класса 1 – p?

5. Выписать формулу для коррекции весов наискорейшим градиентным спуском, если функция ошибки – средний квадрат отклонения (СКО), а метод решения – линейная регрессия.

6. Что такое алгоритм дробной коррекции?

7. Что такое метод Роше (Роккио)?

8. Выписать известные критерии расщепления (для построения деревьев).

9. Покажите, что композиция транзитивных нечётких отношений не всегда транзитивна.

10. Что такое модулярность? (в анализе социальных сетей)

11. Чему равен спектр Лапласа полного графа?

Методические материалы для проведения процедур оценивания результатов обучения

В течение семестра даётся несколько контрольных заданий: как на решение реальных прикладных задач анализа данных, так и на умение применять математический аппарат. По результатам их выполнения студенты могут получать штрафные баллы от 0 до 10. Невыполнение задания – 10 штрафных баллов, идеальное выполнение – 0 (ноль). В случае решения реальной задачи, подготавливается код программы, отчёт по решению и презентация для доклада по решению. Оценивается все составляющие задания. В конце семестра вычисляется ориентировочная оценка: до 10 штрафных баллов (здесь и далее – включительно) – отлично, 11–20 штрафных баллов – хорошо, 21–30 штрафных баллов – удовлетворительно, больше 30 – неудовлетворительно. В случае неудовлетворительной оценки все невыполненные задание пересдаются (возможно, в новой постановке). Экзамен проводится письменно. Экзаменационный билет состоит из нескольких задач. Каждая задача оценивается по системе: решена / не решена. Задаются два порога (на число решённых задач): если число верно решённых задач превышает первый порог – ставится оценка на 2 балла выше ориентировочной, если превышает второй – на 1 балл.

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине (модулю)				
Оценка РО и соответствующие виды оценочных средств	2 (не зачтено)	3 (зачтено)	4 (зачтено)	5 (зачтено)
Знания *Экзамен*	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения *Практические задания*	Отсутствие умений	В целом успешное, но не систематическое умение	В целом успешное, но содержащее отдельные пробелы умение (допускает неточности непринципиального характера)	Успешное и систематическое умение
Навыки (владения, опыт деятельности)	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной	Сформированные навыки (владения), применяемые при решении задач

Экзамен, практические занятия			форме	

7. Ресурсное обеспечение:
7.1. Перечень основной и дополнительной литературы

Основная литература

1) К.Д. Маннинг, П. Рагхаван, Х. Шютце «Введение в информационный поиск» // . — Вильямс, 2011.
2) Дьяконов А.Г. Практикум на ЭВМ кафедры математических методов прогнозирования (системы WEKA, RapidMiner и MatLab): Учебное пособие. – М.: Издательский отдел факультета ВМиК МГУ им. М.В. Ломоносова; МАКС Пресс, 2010. – 133с.: ил. (ISBN 978-5-89407-432-0)
3) Дьяконов А.Г. Практикум на ЭВМ кафедры математических методов прогнозирования (логические игры, обучение по прецедентам): Учебное пособие. – М.: Издательский отдел факультета ВМиК МГУ им. М.В. Ломоносова; МАКС Пресс, 2010. – 164с.: ил. (ISBN 978-5-89407-431-3)

Дополнительная литература

1) Шурыгин А.М. Математические методы прогнозирования // М., Горячая линия — Телеком, 2009, 180 с.
2) К.Д. Маннинг, П. Рагхаван, Х. Шютце «Введение в информационный поиск» // . — Вильямс, 2011.
3) Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: Учебное пособие.– М.: Издательский отдел ф-та ВМиК МГУ им. М.В. Ломоносова, 2006. – 72с. (ISBN 5-89407-252-2)
4) Ту Дж., Гонсалес Р. Принципы распознавания образов // Издательство Мир , Мо-сква, 1978 - 412 стр.
5) Рыжов А.П. Элементы теории нечетких множеств и измерения нечеткости. Москва, Диалог-МГУ, 1998, 116 с.

7.2. Перечень лицензионного программного обеспечения, в том числе отечественного производства

При реализации дисциплины может быть использовано следующее программное обеспечение:

Программное обеспечение для подготовки слайдов лекций MS PowerPoint, MS Word
Программное обеспечение для создания и просмотра pdf-документов Adobe Reader
Издательская система LaTeX
Язык программирования Python и среда разработки Jupiter Notebook (вместе с библиотеками numpy, scikit-learn, pandas)
Язык программирования R и среда разработки R Studio
Среда разработки MATLAB.

7.3. Перечень профессиональных баз данных и информационных справочных систем

1. http://www.edu.ru – портал Министерства образования и науки РФ
2. http://www.ict.edu.ru – система федеральных образовательных порталов «ИКТ в образовании»
3. http://www.openet.ru - Российский портал открытого образования
4. http://www.mon.gov.ru - Министерство образования и науки Российской Федерации
5. http://www.fasi.gov.ru - Федеральное агентство по науке и инновациям

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. Math-Net.Ru [Электронный ресурс] : общероссийский математический портал / Математический институт им. В. А. Стеклова РАН ; Российская академия наук, Отделение математических наук. - М. : [б. и.], 2010. - Загл. с титул. экрана. - Б. ц.
URL: http://www.mathnet.ru
2. Университетская библиотека Online [Электронный ресурс] : электронная библиотечная система / ООО "Директ-Медиа" . - М. : [б. и.], 2001. - Загл. с титул. экрана. - Б. ц. URL: www.biblioclub.ru
3. Универсальные базы данных East View [Электронный ресурс] : информационный ресурс / East View Information Services. - М. : [б. и.], 2012. - Загл. с титул. экрана. - Б. ц.
URL: www.ebiblioteka.ru
4. Научная электронная библиотека eLIBRARY.RU [Электронный ресурс] : информационный портал / ООО "РУНЭБ" ; Санкт-Петербургский государственный университет. - М. : [б. и.], 2005. - Загл. с титул. экрана. - Б. ц.
URL: www.eLibrary.ru

7.5. Описание материально-технического обеспечения.

Факультет ВМК, ответственный за реализацию данной Программы, располагает соответствующей материально-технической базой, включая современную вычислительную технику, объединенную в локальную вычислительную сеть, имеющую выход в Интернет. Используются специализированные компьютерные классы, оснащенные современным оборудованием. Материальная база факультета соответствует действующим санитарно-техническим нормам и обеспечивает проведение всех видов занятий (лабораторной, практической, дисциплинарной и междисциплинарной подготовки) и научно-исследовательской работы обучающихся, предусмотренных учебным планом.

8. Соответствие результатов обучения по данному элементу ОПОП результатам освоения ОПОП указано в Общей характеристике ОПОП.

9. Разработчик (разработчики) программы.

д.ф.- м.н., профессор Дьяконов Александр Геннадьевич (djakonov@mail.ru)

10. Язык преподавания - русский.