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

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Интерпретируемое машинное обучение»

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 №7, от 29 сентября 2021 года)

Москва 2021

Рабочая программа дисциплины (модуля) разработана в соответствии с Федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки 01.04.02 "Прикладная математика и информатика" утвержденного Приказом Министерства образования и науки РФ от 10 января 2018 г. N 13.

1. Место дисциплины (модуля) в структуре ОПОП ВО:
Дисциплина (модуль) относится к обязательной части основной профессиональной обра-зовательной программы.

2. Входные требования для освоения дисциплины (модуля), предварительные условия:
Изучение дисциплины базируется на освоении дисциплины «Робастные модели в машинном обучении».

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**
ОПК-3. Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	ОПК-3.1. Знает возможности современных инструментальных средств и систем программирования в области профессиональной деятельности. ОПК-3.2. Умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач в области профессиональной деятельности. ОПК-3.3. Имеет практический опыт разработки математических моделей и их анализа при решении задач в области профессиональной деятельности.	ОПК-3.1. З-3 ЗНАТЬ: актуальные методы построения и анализа математических моделей в области естественных наук, экономики, социологии и информационно-коммуникационных технологий. ОПК-3.2.У-2 УМЕТЬ: применять современные методы построения математических моделей, а также разрабатывать новые аналитические и численные методы их анализа. ОПК-3.3. В-2 ВЛАДЕТЬ: навыками анализа, в том числе с применением информационно-коммуникационных технологий, модельных расчетов с учетом границ применимости модели, навыками интерпретации полученных результатов для выявления новых данных о моделируемом процессе или построения нового алгоритма управления этим процессом.

ОПК-6. Способен адаптировать и применять на практике классические и новые научные принципы и методы исследований для решения задач в области создания и применения технологий и систем искусственного интеллекта и методы исследований	ОПК-6.1. Адаптирует известные научные принципы и методы исследований с целью их практического применения ОПК-6.2. Решает профессиональные задачи на основе применения новых научных принципов и методов исследования	ОПК-6.1. З-1. Знает фундаментальные научные принципы и методы исследований ОПК-6.1. У-1. Умеет адаптировать с целью практического применения фундаментальные и новые научные принципы и методы исследований ОПК-6.2. З-1. Знает особенности решения профессиональные задачи на основе применения новых научных принципов и методов исследования ОПК-6.2. У-1. Умеет разрабатывать, контролировать, оценивать и исследовать компоненты профессиональной деятельности; планировать самостоятельную деятельность в решении профессиональных задач

4. Объем дисциплины (модуля) составляет 3 з.е., в том числе 36 академических часа, отведенных на контактную работу обучающихся с преподавателем, 72 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 рассматриваются способы интерпретации моделей машинного обучения. Рассматриваются алгоритмы для заранее известных моделей машинного обучения и для случаев, когда устройство модели представляет собой «черный ящик». Изучаются способы интерпретации прогнозов построенной модели и интерпретация отдельных примеров. Затрагиваются задачи выбора наилучших метрик качества для задачи интерпретации, визуализации полученных результатов, отбора и предобработки признаков. Приводятся способы модификации как алгоритмов построения модели машинного обучения, так и алгоритмов интерпретации данной модели, с целью увеличения показателя интерпретируемости.

		В том числе

Наименование и краткое содержание разделов и тем дисциплины (модуля), **форма промежуточной аттестации по дисциплине (модулю)**	**Всего (часы)**	**Контактная работа (работа во взаимодействии с преподавателем), часы из них**						**Самостоятельная работа обучающегося, часы из них**		
		Занятия лекционного типа	Практические занятия	Групповые консультации	Индивидуальные консультации	Учебные занятия, направленные на проведение текущего контроля успеваемости (коллоквиумы, практические контрольные занятия и др)	**Всего**	Выполнение домашних заданий	Подготовка рефератов и т.п.	**Всего**
Тема 1 Введение в интерпретируемое машинное обучение. Обзор существующих результатов в области машинного обучения и кибербезопасности	7	1					1	6		6
Тема 2 Основная терминология интерпретируемого машинного обучения. Ключевые аспекты. Открытые проблемы	7	1					1	6		6
Тема 3 Свойства и метрики качества интерпретируемости	7	1					1	6		6
Тема 4 Интерпретируемые модели - линейная регрессия - логистическая регрессия	12	2	4				6	6		6

- GLM, GAM - деревья решений - ассоциативные правила - RuleFit - наивный байесовский классификатор - knn										
Тема 5 Методы интерпретации независящие от модели - PDP - ICE - ALE - SHAP - глобальные методы - локальные методы, LIME	12	2	4				6	6		6
Тема 6 Интерпретация с помощью примеров	7	1					1	6		6
Тема 7 Правдоподоные и противоречащие объяснения	9	1	2				3	6		6
Тема 8 Визуализация реккурентных нейронных сетей	7	1					1	6		6
Тема 9 Методы интерпретации для многомерного прогнозирования и анализа чувствительности	11	1	4				5	6		6
Тема 10 Отбор признаков для задачи интерпретации	7	1					1	6		6
Тема 11 Методы устранения предвзятости. Методы определения причинно-следственных связей	12	2	4				6	6		6
Тема 12 Способы модификации модели для лучшей интерпретируемости	8	2					2	6		6
Тема 13 Состязательная робастность	2	2					2			0

Промежуточная аттестация – зачет										
Итого	108	18	18				54			18

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

Примеры заданий для практических занятий

Задача 1. Интерпретация для задачи регрессии. Задан датасет, содержащий в себе информацию о весе и росте 25000 человек. Обучить модель, построить уравнение линейной регрессии вида weight = $\beta_0+\beta_1$ height. Визуализировать обучающую выборку вместе с полученным уравнением. Посчитать MAE, MSE, коэффициент корреляции Пирсона. Объяснить полученные значения.

Задача 2. Поиск ассоциативных правил с помощью ПП SPMF:

(a) Для массива данных о контекстной рекламе 2000 компаний × 3000 словосочетаний найти ассоциативные правила для минимальной поддержки minsupp = 35 и minconf = 1. Необходимо указать число таких правил.

(b) Для исходного массива данных найти замкнутые ассоциативные правила для минимальной поддержки minsupp=35 и minconf=1. Необходимо указать число таких правил.

(c) Для исходного массива данных найти 5 самых частых правил при минимальной достоверности minconf = 0,8. Необходимо указать эти правила и дать интерпретацию.

Задача 3. Анализ посещаемости сайтов на основе решеток формальных понятий:

Для трех контекстов о посещаемости некоторого сайта в терминах посещений сайтов новостной, образовательной и финансовой тематики необходимо выполнить:

(a) Удалением некоторого числа сайтов (признаков) или пользователей (объектов) добиться числа формальных понятий не менее 100, но не сильно превышающего это значение.

(b) Для контекстов, полученных удалением объектов или признаков в пункте a), построить диаграммы решеток понятий.

(c) Привести 3–5 примеров понятий в виде пары (размер объема понятия, содержание понятия) для размера содержания 2 и более сайта. Дать содержательную интерпретацию найденных понятий.

(d) Привести пример импликации вида A → B, найденной по диаграмме решетки понятий с указанием ее поддержки.

Задача 4. Интерпретация для задачи классификации. Задан датасет о сердечно-сосудистых заболеваниях, 11 признаков, 70000 объектов. Обучить модель логистической регрессии. Определить, какие из признаков наиболее пагубно влияют на развитие сердечно-сосудистых заболеваний (глобальная интерпретация). Для конкретного человека определить, склонен ли он к сердечно-сосудистым заболеваниям

(локальная интерпретация). Определить значения для каждого из признаков, для которых вероятность иметь сердечно-сосудистые заболевания превышает 0.5. Визуализировать полученные граничные значения вместе с обучающей выборкой.

Задача 5. Сравнение метрик качества для задач регрессии. Задан датасет данных об опозданиях самолетов некоторой авиакомпании. Построить модель, позволяющую узнать, на сколько минут опоздает самолет. Обучить регрессоры:

- линейная регрессия
- полиномиальная регрессия
- полиномиальная регрессия без квадратов
- Ridge регрессия
- дерево решений
- регуляризированная линейная регрессия kNN
- случайный лес
- MLP регрессия

Сравнить полученные модели по метрикам RMSE, R^2-score.

Задача 6. Сравнение метрик качества для задач классификации. Задан датасет данных об опозданиях самолетов некоторой авиакомпании. Построить модель, позволяющую узнать, опоздает ли самолет. Обучить классификаторы:

- логистическая регрессия
- RidgeClassifierCV
- DecisionTreeClassifier
- KNeighborsClassifier
- GaussianNB
- GradientBoostingClassifier
- RandomForestClassifier
- MLPClassifier

Сравнить полученные модели по метрикам: accuracy, recall, ROC-AUC, F1-score, коэффициент корреляции Мэтьюса.

Задача 7. Сравнение метрик качества для методов сокращения размерности. Задан некоторый датасет. Необходимо сократить его размерность до 3, используя методы: PCA, t-SNE, VAE. Визуализировать и сравнить полученные результаты.

Задача 8. Изучение feature importance. Задан датасет, описывающий свойства характера людей в зависимости от того, каким ребенком был испытуемый (старший, младший, средний), определить на какие из свойств характера наиболее влияет очередность рождения. Использовать классификаторы:

- DecisionTreeClassifier
- GradientBoostingClassifier
- RandomForestClassifier
- LogisticRegression

- LinearDiscriminantAnalysis
- MLPClassifier

Для каждой из моделей определить feature importance. Вычислить PFI. Визуализировать графики PDP и ICE, провести их сравнительный анализ с методами поиска feature importance для известной модели.

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

Список вопросов для зачета.

1. Ключевые аспекты интерпретируемости. Открытые задачи в области интерпретируемого машинного обучения.
2. Оценка качества интерпретации модели обучения и способы ее получения.
3. Интерпретируемые модели. Линейная регрессия. Логистическая регрессия.
4. Интерпретируемые модели. GLM, GAM. Деревья решений.
5. Интерпретируемые модели. ассоциативные правила. Алгоритм RuleFit.
6. Интерпретируемые модели. Наивный байесовский классификатор. Knn.
7. Методы интерпретации независящие от модели. PDP. ICE. ALE.
8. Методы интерпретации независящие от модели. SHAP. Глобальные методы.
9. Методы интерпретации независящие от модели. Локальные методы. LIME.
10. Интерпретация с помощью примеров.
11. Правдоподоные и противоречащие объяснения.
12. Методы интерпретации для многомерного прогнозирования и анализа чувствительности.
13. Feature selection и feature engineering для задачи интерпретации.
14. Методы устранения предвзятости. Методы определения причинно-следственных связей.
15. Способы модификации модели для лучшей интерпретируемости.
16. Обеспечение надежности. Состязательная робастность.

Методические материалы для проведения процедур оценивания результатов обучения

Особенности организации процесса обучения

Для эффективного освоения курса рекомендуется перед каждым занятие привести в порядок конспекты лекций. После каждого занятия рекомендуется найти и прочитать дополнительную литературу по теме лекции и прочитать свои конспекты.

Система контроля и оценивания

За каждую домашнюю выставляются баллы (максимум 40 баллов). Пусть М – максимальное число баллов, которое может набрать студент. В конце семестра баллы конвертируются в оценку О1 следующим образом:

меньше М/2 баллов: О1=2;

больше или равно М/2 баллов, но меньше 2М/3: О1=3;

больше или равно 2М/3 баллов, но меньше 5М/6: О1=4;

больше или равно 5М/6 баллов: О1=5.

На зачете оценка О1 является стартовой. Окончательная оценка определяется исходя из оценки устного ответа студента, при этом она не может отличаться от стартовой оценки более чем на 1 балл.

Структура и график контрольных мероприятий

Сдача домашних заданий, устный зачет в конце семестра.

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 *Зачет*	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	Сформированные навыки (владения), применяемые при решении задач

Зачет, практические занятия			используемые не в активной форме	

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

Основная литература:

1. Шакла,Нишант Машинное обучение & TensorFlow : [пер. с англ.]. / Нишант Шакла при участии Кена Фрикласа. - СПб. [и др.] : Питер, 2019. - 331, [1] с.; 24 см - (Библиотека программиста).
2. Шолле,Франсуа Глубокое обучение на Python / Франсуа Шолле ; [пер. с англ. А. Киселева]. - СПб. [и др.] : Питер, 2020. - 397, [1] с.; 24 см - (Библиотека программиста).

Дополнительная литература:

1. Bishop C. M. Pattern recognition and machine learning. – Springer, 2006
2. .Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. – М: ДМК Пресс. – 2016. (Coelho L. P., Richert W. Building machine learning systems with Python. — 2nd ed. — Packt Publishing Ltd, 2015.)
3. Max Kuhn, Kjell Johnson. Applied Predictive Modeling. — Springer, 2013.
4. Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag, 2009. — 746 p. — ISBN 978-0-387-84857-0.
5. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «Распознавание». Математические методы. Программная система. Практические применения. — М.: Фазис, 2006. ISBN 5-7036-0108-8.
6. I.H. Witten, E. Frank Data Mining: Practical Machine Learning Tools and Techniques. — 2nd ed. — Morgan Kaufmann, 2005 ISBN 0-12-088407-0
7. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004. ISBN 966-00-0341-2.
8. Bishop C. M. Pattern recognition and machine learning. – Springer, 2006

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

При реализации дисциплины может быть использовано следующее программное обеспечение:

1. Программное обеспечение для подготовки слайдов лекций MS PowerPoint
2. Программное обеспечение для создания и просмотра pdf-документов Adobe Reader
3. Издательская система LaTeX.

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, 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.

Терёхина Ирина Юрьевна, Гамаюнов Денис Юрьевич.

10. Язык преподавания - русский.