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:

Методы машинного обучения

Уровень высшего образования:

магистратура

Направление подготовки / специальность:

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

Направленность (профиль) ОПОП:

Перспективные методы искусственного интеллекта
в сетях передачи и обработки данных

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 относится к базовой части ОПОП ВО.
Дисциплина входит в обязательную часть магистерской образовательной программы «Перспективные методы искусственного интеллекта в сетях передачи и обработки данных», изучается в 2-м семестре.

2. Входные требования для освоения дисциплины (модуля), предварительные условия (если есть):
Изучение дисциплины базируется на знаниях по математическому анализу, теории вероятностей, математической статистике, оптимизации в объеме, соответствующих 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):

Планируемые результаты обучения по дисциплине (модулю)		
Формируемые компетенции (*код и наименование компетенции*)	**Индикаторы достижения компетенций (*код и наименование индикатора*)**	**Результаты обучения *(знания, умения)***
ОПК-3. Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности	ОПК-3.1 Применяет современные методы построения математических моделей и их анализа при решении задач в области профессиональной деятельности	ОПК-3.1.З-1 Знает актуальные методы построения и анализаматематических моделей в области естественных наук, экономики, социологии и информационно-коммуникационных технологий. ОПК-3.1.У-1 Умеет применять современные методы построенияматематических моделей, а также разрабатывать новые аналитические и численные методы их анализа. ОПК-3.1.В-1 Владеет: навыками анализа, в том числе спримененим 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.

ПК-3. Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач	ПК-3.1. Ставит задачи по разработке или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	ПК-3.1. З-1. Знает классы методов и алгоритмов машинного обучения ПК-3.1. У-1. Умеет ставить задачи и разрабатывать новые методы и алгоритмы машинного обучения
	ПК-3.2. Руководит исследовательской группой по разработке или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	ПК 3.2. З-1. Знает методы и критерии оценки качества моделей машинного обучения ПК 3.2. У-1. Умеет определять критерии и метрики оценки результатов моделирования при построении систем искусственного интеллекта в исследуемой области
	ПК-3.3. Разрабатывает унифицированные и обновляемые методологии описания, сбора и разметки данных, а также механизмы контроля за соблюдением указанных методологий	ПК-3.3. З-1. Знает унифицированные и обновляемые методологии описания, сбора и разметки данных, а также механизмы контроля за соблюдением указанных методологий ПК-3.3. У-1. Умеет разрабатывать унифицированные и обновляемые методологии описания, сбора и разметки данных, а также механизмы контроля за соблюдением указанных методологий

4. Объем дисциплины (модуля) составляет 4 з.е., в том числе 72 академических часа контактная работа с преподавателем - 36 академических часа занятий лекционного типа, 36 академических часов занятий практического типа, 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:
5.1. Структура дисциплины (модуля) по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий (в строгом соответствии с учебным планом)

Наименование разделов и тем дисциплины (модуля), **Форма промежуточной аттестации по дисциплине (модулю)**	**Номинальные трудозатраты обучающегося**			**Всего академических часов**	**Форма текущего контроля успеваемости* (наименование)**
	Контактная работа (работа во взаимодействии с преподавателем) Виды контактной работы, академические часы		**Самостоятельная работа обучающегося, академические часы**		
	Занятия лекционного типа	Практические занятяия			
Тема 1. Систематизация задач и методов машинного обучения	6	6	12	24	контрольная работа
Тема 2. Базовые технологии	6	6	12	24	контрольная работа
Тема 3. Современные технологии	6	6	12	24	контрольная работа
Тема 4. Задачи с целевым признаком	6	6	12	24	контрольная работа
Тема 5. Задачи без целевого признака	6	6	12	24	контрольная работа
Тема 6. Слабоструктурированная информация	6	6	12	24	реферат
Другие виды самостоятельной работы (отсутствуют)	—	—			—
Промежуточная аттестация (экзамен)					
Итого	*36*	*36*	*72*	144	—

5.2. Содержание разделов (тем) дисциплины

№ п/п	Наименование разделов (тем) дисциплины	Содержание разделов (тем) дисциплин
1.	Тема 1. Систематизация задач и методов машинного обучения	Основания систематизации: по целевому признаку, по выходному признаку, по доступности распределений, по структуре данных, по разметке обучения виды моделей: дискриминативная функция, дискриминативная модель, генеративная модель
2.	Тема 2. Базовые технологии	Библиотека numpy, модель данных ndarray, механизм broadcasting, внутреннее представление моделей данных Библиотека pandas, индексы на строках и столбцах, модель данных dataframe Библиотека matplotlib, программный интерфейс pyplot Библиотека scikit-learn: раздел classification, раздел regression Библиотека scikit-learn: раздел model selection Библиотека scikit-learn: раздел clustering, раздел dimensionality reduction Библиотека scikit-learn: раздел preprocessing, раздел feature extraction
3.	Тема 3. Современные технологии	Библиотека tensorflow Библиотека и программный интерфейс keras
4.	Тема 4. Задачи с целевым признаком	бинарная классификация многоклассовая классификация функция потерь: 0-1-loss, meanabsoluteerror, meansquarederror, cross-entropy средний риск байесовский классификатор эмпирический риск одноклассовая классификация, P-classification классификация с пересекающимися классами (multi-label, multi-output) детекция идентификация верификация локализация
		SVM как дискриминативная функция kNN как дискриминативная модель линейные модели, их вероятностная интерпретация

		проблема смещения-дисперсии ансамбли классификаторов
5.	Тема 5. Задачи без целевого признака	Восстановление плотности KDE как генеративная модель метод главных компонент автоэнкодеры кластеризация сегментация
6.	Тема 6. Слабоструктурированная информация	Представление изображений Сверточные нейронные сети Представление последовательностей Текст как последовательность Представление текстов в виде мешка слов. tf-idf Модель языка. N-граммы. Рекуррентные нейронные сети Представление элементов описания как точек в линейном пространстве (embeddings)

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 (в отсутствие утвержденных соответствующих локальных нормативных актов на факультете)

Вопросы к контрольным работам

1. Показатель X в классах К1 и К2 распределён нормально с параметрами: в К1 математическое ожидание 2, стандартное отклонение 4; в К2 математическое ожидание 3, стандартное отклонение 1. Выделить на числовой оси значений показателя X области отнесения байесовским классификатором к классам К1 и К2. Априорные вероятности классов К1 и К2 равны 0.6 и 0.4 соответственно.
2. Каждый год варан подрастает на A% от своего веса в начале года. A – случайная величина с известными матожиданием 5 и дисперсией 1 (одна и та же для всех варанов во все годы). В начале жизни каждый варан имеет вес 1. Построить байесовский классификатор для определения возраста варана (полных лет) по его весу, минимизирующий частоту ошибки. Предположить, что распознаваться будут «достаточно» взрослые вараны.
3. Выборка объектов из класса 1 и класса 2 определяется таблицами ниже. Указать тупиковые тесты.

	X1	X2	X3	X4							X1	X2	X3	X4
Об.1	0	1	1	0						Об.1	0	1	0	0
Об.2	0	0	1	1						Об.2	1	0	1	0
Об.3	1	0	0	1						Об.3	0	1	0	0
Об.4	1	0	1	1						Об.4	1	1	0	0

4. Тестирование в банке системы распознавания для определения недобросовестных заёмщиков выявило связь между чувствительностью и ложной тревогой, показанную в таблице. Определить, приведёт ли эксплуатация системы к увеличению доходов банка. Определить возможный прирост дохода в расчёте на одну поданную заявку. Известно, что доход банка на одного заёмщика составляет 40000 денежных единиц, потери в результате отказа заёмщика от платежей составляют 120000 единиц. Доля недобросовестных заёмщиков составляет 7%.

Чувст.	Лож. Тр.
0.02	0.0001
0.12	0.003
0.23	0.05
0.38	0.12
0.47	0.16
0.58	0.19
0.67	0.23

0.78	0.34
0.89	0.52
0.97	0.72
1	0.87

5. В таблице даны значения переменных X и Y для четырёх экспериментов. Найти коэффициент корреляции и значения коэффициентов a и b для оптимальной по методу наименьших квадратов линейной модели Y=a +b*X.

X	Y
0.12	52
0.23	37
0.35	17
0.46	2

6. Рассматривается задача классификации на два класса: положительный и отрицательный. В ходе тестирования классификатора получены следующие результаты: полнота составляет 75%, общая точность составляет 80%. Какие значения может принимать точность?
7. Магазин собрал сведения о покупках (транзакции в файле). Были построены ассоциативные правила. Какое правило, содержащее в условии 2 элемента, имеет наибольшую поддержку?
8. Государственная избирательная комиссия зафиксировала результаты выборов по партиям и по регионам (таблица в файле). Требуется кластеризовать регионы по правилу k-средних для числа кластеров K от 1 до 12. Для каждого числа кластеров K найти максимальный радиус кластера. Построить график этой величины от K. На основании графика предположить, сколько групп регионов разумно выделить по итогам выборов.
9. В алгоритме вычисления оценок написать формулу для числа голосов, если система опорных множеств состоит из всех непустых подмножеств, а функция близости определяется только порогами e_1, …, e_n.
10. Обоснуйте способ построения всех тупиковых тестов через приведение системы тестовых уравнений к неупрощаемой ДНФ.

Темы для рефератов

1. Распознавание рукописных цифр, написанных разными людьми.
2. Выделение на изображении сегментов кожи.
3. Периоды покоя и извержения гейзеров.
4. Анализ зависимости времени вылета рейса от пункта назначения.
5. Анализ результатов ЕГЭ по регионам.
6. Связь стоимости просмотра фильма с его характеристиками.
7. Связь стоимости монитора с его характеристиками.

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

Вопросы к экзамену

1. Место и роль ИАД в современной структуре человеческой деятельности. Место ИАД в передаче научного знания. Уровни технологий анализа данных, их назначение, место ИАД в технологиях АД. Понятие о моделировании реального мира в науке. Физическая модель. Модель “решателя”. Информационная модель. Эвристическая модель. Основная особенность ИАД (обучение и эксплуатация эвристической информационной модели). Понятие о машинном обучении.
2. Основные моделиданных (dataframe, multidimensional, similaritytensor, transactional). Их назначение научное, технологическое. Гомогенные и гетерогенные модели.
3. Фундаментальные задачи ИАД и основные инструменты статистики. Прикладная жизнь ИАД: декомпозиция содержательных задач предметной области. Научная жизнь ИАД: сведение к задачам фундаментальной математики. Обучение и эксплуатация в фундаментальных задачах. Основания таксономии (способы группирования) фундаментальных задач. Таксономия по наличию в исходных данных целевого признака. Таксономия по моделям данных: в разрезе исходных данных, в разрезе результатов.
4. Модель данных «признаковое описание объектов». Понятие о шкалах значений атрибутов. Представление реляционными технологиями. Схемы «звезда» и «снежинка».
5. Многомерная модель данных. Группирование объектов как переход к многомерной модели данных. Аналитические пространства. Измерения и категории. Показатели. Детализация. Функции агрегирования, типы показателей по агрегированию.
6. Транзакционная модель данных. Связанные с ней задачи.
7. Общая задача классификации. Понятие об обучении и использовании. Объект, модель, алгоритм-классификатор. Универсальные ограничения.
8. Локальные ограничения. Оптимизационный подход. Функционалы качества на размеченной выборке. Частотные функционалы качества. Случай бинарной классификации. Стоимостные функционалы качества. Несоответствие частотных и стоимостных функционалов качества человеческому поведению.
9. Подходы к многокритериальной оптимизации.
10. Понятие байесовского классификатора как оптимального алгоритма распознавания. Классификаторы, основанные на использовании формулы Байеса. Линейный дискриминант Фишера. Логистическая регрессия.
11. Метрические методы распознавания образов. Метод k ближайших соседей. Взвешенное голосование. Метод парзеновского окна.
12. Метрические методы распознавания образов. Ядровая оценка плотности (kernel density estimation), связь с методами k ближайших соседей и парзеновского окна.
13. Метрические методы распознавания образов. Проклятие размерности (curse of dimensionality).
14. Форматы представления информации. Текстовые файлы, их атрибуты, проблема определения атрибутов.
15. Текстовые форматы представления таблиц: separate dvalues, delimitedtext. Экранирование символов.
16. Форматы представления транзакционных данных.
17. Диаграммы для наборов точек из конечномерных евклидовых пространств.
18. Диаграммы для многомерной модели данных. Системы отчётности.

19. Задача восстановления регрессии: аппроксимационный подход, статистический подход. Простая регрессия. Множественная регрессия. Поиск коэффициентов по МНК. Недостатки МНК. Трёхкомпонентное разложение ошибки регрессионных моделей. Регуляризация по Тихонову. Гребневая регрессия, лассо, эластичные сети.
20. Модель данных «метрические тензоры», гомогенные и гетерогенные многомерные матрицы сходства. Группирование объектов как кластеризация по метрическим описаниям. Гомогенная кластеризация, бикластеризация, мультикластеризация. Основные типы результатов кластеризации (плоская, последовательная плоская, иерархическая, нечёткая, стохастическая, ранговая).
21. Плоская кластеризация. Задача и метод k-means.
22. Последовательная плоская кластеризация. Метод ФОРЕЛЬ.
23. Иерархическая кластеризация. Дивизивная. Агломеративная, функционалы связи (linkage).
24. ROC анализ.
25. Линейная модель. Линейная машина как метод обучения линейной модели.
26. Линейная модель. Метод опорных векторов.
27. Модель перцептрона Розенблатта. Метод его обучения. Теорема Новикова. Переход от сдвига к фиктивному признаку.
28. Многослойные перцептроны. Метод обратного распространения ошибки. Функции активации, удобные для распространения ошибки.
29. Многослойные перцептроны. Возможность разделения множеств для перцептронов разной глубины.
30. Решающие деревья.
31. Алгоритмы, основанные на голосования по наборам закономерностей.
32. Bagging. Boosting. Решающие леса (random forest).
33.
34. Линейные методы распознавания образов. Линейная регрессия методом наименьших квадратов. Логистическая регрессия. Вероятностная интерпретация методов.
35. Переобучение. Регуляризация. Основные методы регуляризации линейных моделей (L2, L1, elasticnet), их свойства. Механизм гребневой регрессии подробно. Вероятностная интерпретациярегуляризации.
36. SVM. Постановка задачи, вид решения, понятие опорного объекта. Нелинейное обобщение SVM (kernel trick), определение ядра, примеры.
37. Решающие деревья. Жадный алгоритм построения решающего дерева. Критерии ветвления (энтропийный, Джини), их обоснование.
38. Композиции решающих правил. Бустинг: постановка задачи, идея жадного алгоритма. AdaBoost, градиентный бустинг, AnyBoost.
39. Композиции решающих правил. Стохастические методы: bagging (+ bootstrap), RSM. Случайные леса. Обоснование стохастических методов: bias-variance decomposition.
40. Нейронные сети. Число слоёв и аппроксимационные возможности сети, причины глубокого обучения. Метод обратного распространения ошибки. Dropout, batch normalization.
41. Анализ изображений. Свёрточные нейронные сети, решение задачи классификации. Задача сегментации изображений: постановка, идея решения. Upsampling (подходы). U-net.
42. Анализ изображений. Свёрточные нейронные сети, решение задачи классификации. Задача локализации объекта на изображении: постановка, идея решения.
43. Анализ изображений. Свёрточные нейронные сети, решение задачи классификации. Задача обнаружения объектов: постановка, подходы к решению (R-CNN, FastR-CNN, FasterR-CNN).
44. Анализ текстов. Представление текстов в виде мешка слов. tf-idf. Модел ьязыка. N-граммы.

45. Анализ текстов. Скрытые марковские модели, алгоритм Витерби. Применение скрытых марковских моделей для выделения частей речи.
46. Анализ последовательностей. Рекуррентные нейросети, разные конфигурации входа и выхода. Обучение генератора текста. Задача описания изображений предложениями на естественном языке.
47. Рекуррентные нейросети. Проблема взрыва и затухания градиентов. LSTM.

ШКАЛА И КРИТЕРИИ ОЦЕНИВАНИЯ результатов обучения (РО) по дисциплине				
Оценка виды оценочных средств	2 (не зачтено)	3 (зачтено)	4 (зачтено)	5 (зачтено)
Знания *(виды оценочных средств: опрос, тесты)*	Отсутствие знаний	Фрагментарные знания	Общие, но не структурированные знания	Сформированные систематические знания
Умения *(виды оценочных средств: 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) *(виды оценочных средств: выполнение и защита курсовой работы, отчет по практике, отчет по НИР и т.п.)*	Отсутствие навыков (владений, опыта)	Наличие отдельных навыков (наличие фрагментарного опыта)	В целом, сформированные навыки (владения), но используемые не в активной форме	Сформированные навыки (владения), применяемые при решении задач

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

Основная литература

1. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – М: ДМК Пресс. – 2015. – 400 с. ISBN 978-5-97060-273-7 (Flach P.Machinelearning: theart and science of algorithms thatmakesense of data. – Cambridge University Press, 2012)

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.

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

При реализации дисциплины может быть использовано следующее программное обеспечение:

1. Операционная система ALT Linux MATE Starterkit 9 лицензия GPL
2. Операционная система Microsoft Windows 10 Education академическая лицензия
3. Программный продукт Python 3.5.1 (64-bit)Python Software Foundation
4. Статистический пакет MATLAB (или свободный аналог Octave)

7.3. Перечень профессиональных баз данных и информационных справочных систем

1. http://www.openet.ru - Российский портал открытого образования
2. http://www.mon.gov.ru - Министерство образования и науки Российской Федерации

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

1. mmp.cs.msu.ru
2. www.machinelearning.ru

7.5. Описание материально-технического обеспечения.

Образовательная организация, ответственная 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 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.

    к.ф.- м.н., доцент Майсурадзе Арчил Ивериевич