

Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«Оренбургский государственный университет»

Кафедра прикладной математики

РАБОЧАЯ ПРОГРАММА

ДИСЦИПЛИНЫ

«Б1.Д.В.9 Математическая статистика машинного обучения»

Уровень высшего образования

БАКАЛАВРИАТ

Направление подготовки

01.03.05 Статистика

(код и наименование направления подготовки)

Статистика и управление данными

(наименование направленности (профиля) образовательной программы)

Квалификация

Бакалавр

Форма обучения

Очная

Год набора 2023

Рабочая программа дисциплины «Б1.Д.В.9 Математическая статистика машинного обучения» рассмотрена и утверждена на заседании кафедры

Кафедра прикладной математики

наименование кафедры

протокол № 6 от "18" января 2023г.

Заведующий кафедрой

Кафедра прикладной математики

наименование кафедры


подпись

И.П. Болодурина

расшифровка подписи

Исполнители:

профессор

должность


подпись

И.П. Болодурина

расшифровка подписи

ст. преподаватель

должность


подпись

С.В. Колесник

расшифровка подписи

СОГЛАСОВАНО:

Председатель методической комиссии по направлению подготовки

01.03.05 Статистика

наименование


личная подпись

В.Н. Афанасьев

расшифровка подписи

Заведующий отделом комплектования научной библиотеки


личная подпись

Н.Н. Бигалиева

расшифровка подписи

Уполномоченный по качеству факультета


личная подпись

И.В. Крючкова

расшифровка подписи

№ регистрации _____

1 Цели и задачи освоения дисциплины

Цель (цели) освоения дисциплины: изучение основ теории и методов машинного обучения (machine learning) в вероятностной постановке.

Задачи:

- формирование научного представления о современных методах машинного обучения;
- умения анализировать задачи машинного обучения и осуществлять взвешенный выбор того или иного решения;
- приобретение знаний о принципах и алгоритмах, лежащих в основе методов машинного обучения;
- овладение технологиями и методиками сбора, предварительной подготовки и анализа экспериментальных данных;
- приобретение практических навыков работы с конкретными программными средствами интеллектуального анализа данных;
- использование изученных языков программирования в качестве основы для разработки различных алгоритмов машинного обучения.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательным дисциплинам (модулям) вариативной части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.12.1 Математический анализ, Б1.Д.Б.13.2 Математическая статистика*

Постреквизиты дисциплины: *Б1.Д.В.18 Информационные и коммуникационные технологии в системах защиты статистических данных*

3 Требования к результатам обучения по дисциплине

Процесс изучения дисциплины направлен на формирование следующих результатов обучения

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1-В-4 Применяет методы сбора, хранения, обработки, передачи, анализа и синтеза информации с использованием компьютерных технологий для решения поставленных задач	<u>Знать:</u> - о методах, приемах, алгоритмах и способах сбора, обработки и интерпретации данных; - основные методы и алгоритмы машинного обучения для решения поставленных задач. <u>Уметь:</u> - собирать и обрабатывать статический, экспериментальный, теоретический, графический и т.п. материал; - применять основные модели машинного обучения для анализа и синтеза информации с использованием компьютерных технологий. <u>Владеть:</u> - проведением качественного и количественного анализа собранных данных; - методами разработки алгоритмов и

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
		программного обеспечения работы основных типов интеллектуальных моделей и их обучения.
ПК*-3 Способен собирать данные в ходе непосредственного наблюдения и измерения с применением технических средств в соответствии с утвержденными программами, правилами и методиками	ПК*-3-В-3 Применяет технические средства для измерения параметров различных естественных, гуманитарных, технических и медицинских процессов и явлений, являющихся объектами статистических наблюдений	Знать: - возможности инструментальных средств для анализа данных. Уметь: - применять технические средства для оценки параметров различных процессов и явлений, являющихся объектами статистических наблюдений; - разрабатывать прикладные программы анализа данных. Владеть: - инструментальными средствами анализа данных.
ПК*-5 Способен проводить первоначальную обработку первичных статистических данных в соответствии с утвержденными правилами и методиками	ПК*-5-В-1 Применяет технику перенесения информации на электронные носители и по каналам связи	Знать: - основные типы задач машинного обучения для последующего выбора методов обработки информации. Уметь: - формализовать знания экспертов с применением различных методов представления знаний. Владеть: - техникой предварительной обработки данных, в том числе перенесения информации на электронные носители и по каналам связи, для проведения последующего анализа данных.

4 Структура и содержание дисциплины

4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 3 зачетные единицы (108 академических часов).

Вид работы	Трудоемкость, академических часов	
	6 семестр	всего
Общая трудоёмкость	108	108
Контактная работа:	34,25	34,25
Лекции (Л)	18	18
Лабораторные работы (ЛР)	16	16
Промежуточная аттестация (зачет, экзамен)	0,25	0,25
Самостоятельная работа: - выполнение индивидуального творческого задания (ИТЗ); - самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий); - подготовка к лабораторным занятиям; - подготовка к рубежному контролю.	73,75	73,75
Вид итогового контроля (зачет, экзамен, дифференцированный зачет)	зачет	

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
1	Основные понятия и примеры прикладных задач	8	2		2	4
2	Основы Python для анализа данных	40	6		4	30
3	Линейные методы классификации и регрессии	12	2		2	8
4	Метрические методы классификации и регрессии	12	2		2	8
5	Метод опорных векторов	12	2		2	8
6	Многомерная линейная регрессия	12	2		2	8
7	Критерии выбора моделей и методы отбора признаков	12	2		2	8
	Итого:	108	18		16	74
	Всего:	108	18		16	74

4.2 Содержание разделов дисциплины

1 Основные понятия и примеры прикладных задач

Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Вероятностная постановка задачи обучения. Принцип максимума правдоподобия. Проблема переобучения и понятие обобщающей способности.

Типы задач: классификация, регрессия, ранжирование, прогнозирование.

Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль.

Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия.

Примеры прикладных задач. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных.

2 Основы Python для анализа данных

Основы программирования в Python. Ввод и вывод данных. Типы данных. Условный оператор, циклы. Символьные строки, списки. Процедуры и функции. Массивы. Множества. Словари. Работа с файлами в Python. Функции Python. Библиотеки Python для анализа данных и визуализации.

3 Линейные методы классификации и регрессии

Линейный классификатор, модель МакКаллока-Питтса, непрерывные аппроксимации пороговой функции потерь.

Метод стохастического градиента SG.

Метод стохастического среднего градиента SAG.

Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, «выбивание» из локальных минимумов.

Проблема мультиколлинеарности и переобучения, регуляризация или редукция весов (weight decay).

Вероятностная постановка задачи классификации. Принцип максимума правдоподобия.

Вероятностная интерпретация регуляризации, совместное правдоподобие данных и модели. Принцип максимума апостериорной вероятности.

Гауссовский и лапласовский регуляризаторы.

Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Метод стохастического градиента для логарифмической функции потерь. Многоклассовая логистическая регрессия. Регуляризованная логистическая регрессия. Калибровка Платта.

4 Метрические методы классификации и регрессии

Гипотезы компактности и непрерывности.

Обобщённый метрический классификатор.

Метод ближайших соседей kNN и его обобщения. Подбор числа k по критерию скользящего контроля.

Метод окна Парзена с постоянной и переменной шириной окна.

Метод потенциальных функций и его связь с линейной моделью классификации.

Задача отбора эталонов. Полный скользящий контроль (CCV), формула быстрого вычисления для метода 1NN. Профиль компактности.

Отбор эталонных объектов на основе минимизации функционала полного скользящего контроля.

Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Ядерное сглаживание.

Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра и ширины окна сглаживания.

Задача отсева выбросов. Робастная непараметрическая регрессия. Алгоритм LOWESS.

5 Метод опорных векторов

Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin).

Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция потерь.

Задача квадратичного программирования и двойственная задача. Понятие опорных векторов.

Рекомендации по выбору константы C.

Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера.

Способы конструктивного построения ядер. Примеры ядер.

SVM-регрессия.

Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

Метод релевантных векторов RVM.

6 Многомерная линейная регрессия

Задача регрессии, многомерная линейная регрессия.

Метод наименьших квадратов, его вероятностный смысл и геометрический смысл.

Сингулярное разложение.

Проблемы мультиколлинеарности и переобучения.

Регуляризация. Гребневая регрессия через сингулярное разложение.

Методы отбора признаков: Лассо Тибширани, Elastic Net, сравнение с гребневой регрессией.

Метод главных компонент и декоррелирующее преобразование Карунена-Лозва, его связь с сингулярным разложением.

Спектральный подход к решению задачи наименьших квадратов.

Задачи и методы низкоранговых матричных разложений.

7 Критерии выбора моделей и методы отбора признаков

Критерии качества классификации: чувствительность и специфичность, ROC-кривая и AUC, точность и полнота, AUC-PR.

Внутренние и внешние критерии. Эмпирические и аналитические критерии.

Скользящий контроль, разновидности эмпирических оценок скользящего контроля. Критерий непротиворечивости.

Разновидности аналитических оценок. Регуляризация. Критерий Акаике (AIC). Байесовский информационный критерий (BIC). Оценка Вапника-Червоненкиса.

Сложность задачи отбора признаков. Полный перебор.

Метод добавления и удаления, шаговая регрессия.

Поиск в глубину, метод ветвей и границ.

Усечённый поиск в ширину, многорядный итерационный алгоритм МГУА.

Генетический алгоритм, его сходство с МГУА.

Случайный поиск и Случайный поиск с адаптацией (СПА).

4.3 Лабораторные работы

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	1	Анализ данных на примере набора Ирисов Фишера.	2
2	2	Основы Python для анализа данных	4
3	2	Линейные методы классификации и регрессии: метод стохастического градиента	2
4	3	Метрические методы классификации и регрессии: метод ближайших соседей, метод Парзенковского окна и метод потенциальных функций (реализация)	2
5	4	Линейные методы классификации: метод опорных векторов	2
6	5	Многомерная линейная регрессия. Метод главных компонент	2
7	6	Критерии выбора моделей и методы отбора признаков	2
		Итого:	16

5 Учебно-методическое обеспечение дисциплины

5.1 Основная литература

1. Болодурина, И. П. Основы систем искусственного интеллекта [Электронный ресурс] : учебное пособие для обучающихся по образовательным программам высшего образования по направлениям подготовки 01.03.02 Прикладная математика и информатика, 02.03.02 Фундаментальная информатика и информационные технологии и 09.03.02 Информационные системы и технологии / И. П. Болодурина, Л. С. Гришина, А. Ю. Жигалов; М-во науки и высш. образования Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. образования "Оренбург. гос. ун-т". - Электрон. дан. - Оренбург : ОГУ, 2022. - ISBN 978-5-7410-2781-3.

2. Лимановская, О. В. Основы машинного обучения : учебное пособие / О. В. Лимановская, Т. И. Алферьева ; науч. ред. И. . Обабков ; Уральский федеральный университет им. первого Президента России Б. Н. Ельцина. – Екатеринбург : Издательство Уральского университета, 2020. – 91 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=699059> (дата обращения: 26.06.2023). – ISBN 978-5-7996-3015-7. – Текст : электронный.

3. Бураков, М. В. Системы искусственного интеллекта [Текст] : учебное пособие / М. В. Бураков. - Москва : Проспект, 2021. - 432 с. : ил. - Библиогр.: с. 406-411. - Прил.: с. 412-431. - ISBN 978-5-392-33563-3.

5.2 Дополнительная литература

1. Келлехер, Д. Наука о данных : базовый курс : учебное пособие : [16+] / Д. Келлехер, Б. Тирни ; науч. ред. З. Мамедьяров ; пер. с англ. М. Белоголовского. – Москва : Альпина Паблишер, 2020. – 224 с. : схем., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=598235> (дата обращения: 26.06.2023). – ISBN 978-5-9614-3170-4. – Текст : электронный.

2. Чубукова, И. А. Data Mining : учебное пособие : [16+] / И. А. Чубукова. – 2-е изд., испр. – Москва : Интернет-Университет Информационных Технологий (ИНТУИТ) : Бином. Лаборатория знаний, 2008. – 383 с. – (Основы информационных технологий). – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=233055> (дата обращения: 26.06.2023). – ISBN 978-5-94774-819-2. – Текст : электронный.

3. Воробьев, Г. А. Основы программирования на Python : учебно-методическое пособие : [16+] / Г. А. Воробьев ; Липецкий государственный педагогический университет им. П. П. Семенова-Тян-Шанского. – Липецк : Липецкий государственный педагогический университет им. П.П. Семенова-

Тян-Шанского, 2022. – 89 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=700515> (дата обращения: 26.06.2023). – Библиогр. в кн. – ISBN 978-5-907461-84-0. – Текст : электронный.

4. Титов, А. Н. Python. Обработка данных : учебно-методическое пособие : [16+] / А. Н. Титов, Р. Ф. Тазиева ; Казанский национальный исследовательский технологический университет. – Казань : Казанский национальный исследовательский технологический университет (КНИТУ), 2022. – 104 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=702252> (дата обращения: 26.06.2023). – Библиогр. в кн. – ISBN 978-5-7882-3171-6. – Текст : электронный.

5. Титов, А. Н. Обработка данных в Python : основы работы с библиотекой Pandas : учебно-методическое пособие : [16+] / А. Н. Титов, Р. Ф. Тазиева ; Казанский национальный исследовательский технологический университет. – Казань : Казанский национальный исследовательский технологический университет (КНИТУ), 2022. – 116 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=702231> (дата обращения: 26.06.2023). – Библиогр. в кн. – ISBN 978-5-7882-3164-8. – Текст : электронный.

6. Титов, А. Н. Визуализация данных в Python. Работа с библиотекой Matplotlib : учебно-методическое пособие : [16+] / А. Н. Титов, Р. Ф. Тазиева ; Казанский национальный исследовательский технологический университет. – Казань : Казанский национальный исследовательский технологический университет (КНИТУ), 2022. – 92 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=702253> (дата обращения: 26.06.2023). – Библиогр. в кн. – ISBN 978-5-7882-3176-1. – Текст : электронный.

7. Программные системы статистического анализа : обнаружение закономерностей в данных с использованием системы R и языка Python : учебное пособие : [16+] / В. М. Волкова, М. А. Семенова, Е. С. Четвертакова, С. С. Вожов. – Новосибирск : Новосибирский государственный технический университет, 2017. – 74 с. : ил., табл. – Режим доступа: по подписке. – URL: <https://biblioclub.ru/index.php?page=book&id=576496> (дата обращения: 26.06.2023). – Библиогр.: с. 48. – ISBN 978-5-7782-3183-2. – Текст : электронный.

5.3 Периодические издания

1. Вестник компьютерных и информационных технологий: журнал. - М.: Изд. "Спектр"
2. Информационные технологии: журнал. - М.: Изд. "Новые технологии".
3. Информатика и системы управления : журнал. - Москва : Агентство "Роспечать".
4. Программные продукты и системы : журнал. - Москва : Агентство "Роспечать".

5.4 Интернет-ресурсы

1. <http://machinelearning.ru> - Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.
2. <https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie/> «Coursera»; MOOK: «Введение в машинное обучение».
3. <https://stepik.org/course/116618/promo> - бесплатный курс «Основы анализа данных в Python».
4. <https://youtube.com/playlist?list=PLrCZzMib1e9p6lpNv-yt6uvHGyBxQncEh> – курс «Введение в анализ данных».
5. <https://www.kaggle.com/> - система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению.
6. pythontutor.ru – интерактивный учебник языка Питон.

5.5 Программное обеспечение, профессиональные базы данных и информационные справочные системы

1. Операционная система РЕД ОС.
2. Пакет офисных приложений LibreOffice.
3. Программная система для организации видео-конференц-связи Webinar.ru.

4. Система программирования Python, свободно распространяемая по лицензии GPL <https://www.python.org/>.

5. ГАРАНТ Платформа F1 [Электронный ресурс]: справочно-правовая система. / Разработчик ООО НПП «ГАРАНТ-Сервис», 119992, Москва, Воробьевы горы, МГУ, [1990–2023]. – Режим доступа в сети ОГУ для установки системы: <\\fileserv1\GarantClient\garant.exe>

6. КонсультантПлюс [Электронный ресурс]: электронное периодическое издание справочная правовая система. / Разработчик ЗАО «Консультант Плюс», [1992–2023]. – Режим доступа к системе в сети ОГУ для установки системы: <\\fileserv1\CONSULT\cons.exe>

7. <http://edu.garant.ru/garant/study/> - Интернет-версия ГАРАНТ-Образование, Система ГАРАНТ для студентов, аспирантов и преподавателей

8. Автоматизированная интерактивная система сетевого тестирования - АИССТ (зарегистрирована в РОСПАТЕНТ, Свидетельство о государственной регистрации программы для ЭВМ №2011610456, правообладатель – Оренбургский государственный университет), режим доступа - <http://aist.osu.ru>.

6 Материально-техническое обеспечение дисциплины

Учебные аудитории для проведения занятий лекционного типа, семинарского типа, для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации.

Аудитории оснащены комплектами ученической мебели, техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Для проведения лабораторных занятий используется компьютерный класс, оснащенный компьютерной техникой, подключенной к сети "Интернет".

Помещение для самостоятельной работы обучающихся оснащено компьютерной техникой, подключенной к сети "Интернет", и обеспечено доступом в электронную информационно-образовательную среду ОГУ.