ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

Рабочая программа дисциплины (модуля)
«Статистическая теория обучения»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика»*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2024

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 8
5. **Учебно-методическое обеспечение** 9
6. **Материально-техническое обеспечение** 9
7. **Методические и оценочные материалы** 11

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Статистическая теория обучения» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 02.03.01 «Математика и компьютерные науки», профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Дисциплина (модуль) «Статистическая теория обучения» позволяет понимать, почему и при каких условиях алгоритмы обучаются, как оценивать их обобщающую способность и избегать переобучения. Эта дисциплина критически важна для подготовки специалистов, способных не только применять, но и осмысливать, модифицировать и разрабатывать эффективные алгоритмы обучения.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 «Математика и компьютерные науки», профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) доступна для изучения на 3 или 4 курсе в 6, 7, 8 семестрах на выбор после успешного освоения дисциплин (модулей): «Математический анализ. Часть 1», «Математический анализ. Часть 2» и «Линейная алгебра и геометрия. Часть 1», «Линейная алгебра и геометрия. Часть 2» или «Математический анализ. Продвинутый уровень. Часть 1», «Математический анализ. Продвинутый уровень. Часть 2» и «Линейная алгебра и геометрия. Продвинутый уровень. Часть 1», «Линейная алгебра и геометрия. Продвинутый уровень. Часть 2».

Цель изучения дисциплины (модуля): формирование у студентов глубокого теоретического понимания принципов обучения по данным, включая математические основы обобщающей способности, оценку сложности моделей и анализ ошибок обучения, а также развитие способности критически оценивать и выбирать модели на основе статистических критериев.

Задачи изучения дисциплины (модуля):

— освоить теоретические основы обобщающей способности: неравенства концентрации (например, Хёфдинга), понятие равномерной сходимости;

— познакомиться с мерами сложности моделей: размерность Вапника–Червоненкиса (VC-размерность), радемахеровская сложность;

— изучить фундаментальные теоремы теории обучения, включая теоремы о сходимости эмпирического риска (ERM) и оценки обобщения;

— проанализировать компромисс между смещением и дисперсией и его влияние на выбор модели;

— рассмотреть регуляризацию как инструмент управления сложностью модели (L1, L2, структурная регуляризация);

— понять принципы кросс-валидации и оценки обобщающей способности на практике;

— научиться интерпретировать поведение алгоритмов обучения с теоретической точки зрения.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— основные понятия теории обучения: модель, функция потерь, эмпирический риск, обобщающая способность, класс гипотез;

— базовые модели обучения (pac-обучение, онлайн-обучение) и их теоретические гарантии;

— ключевые меры сложности класса гипотез: vc-размерность и радемахеровскую сложность;

— фундаментальные теоремы обучения (границы обобщения, теоремы о сходимости, роль регуляризации);

— основные алгоритмы (svm, adaboost, ядерные методы, нейронные сети) с точки зрения их теоретических свойств;

уметь:

— формализовать задачу обучения в рамках статистической теории (пространство объектов, гипотез, функция потерь);

— вычислять или оценивать vc-размерность и радемахеровскую сложность для простых классов гипотез;

— применять метод минимизации эмпирического риска (erm) и структурной минимизации риска (srm) для построения алгоритмов с гарантиями обобщения;

— использовать вероятностные неравенства (хёффдинга, чернова, ограниченных разностей) для получения границ обобщающей способности;

— реализовывать базовые алгоритмы статистического обучения (svm, adaboost) с помощью библиотек scikit-learn и интерпретировать результаты с точки зрения теории;

владеть:

— навыками построения и анализа математических моделей обучения с использованием аппарата теории вероятностей и математической статистики;

— методологией строгого анализа и оценки качества алгоритмов машинного обучения с точки зрения их теоретической обоснованности;

— методами получения оценок вычислительной сложности и требуемого объема выборки для различных моделей и алгоритмов;

— техникой применения вероятностных неравенств и методов концентрации для вывода теоретических гарантий обобщающей способности;

— пониманием текущих направлений исследований в области статистической теории обучения и умением интерпретировать результаты современных научных статей.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области искусственного интеллекта, основные принципы критической оценки источников информации и их релевантности.
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
УК-2.	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1.	Знает действующие правовые нормы, регулирующие деятельность в области решения задач, основные методы и подходы к определению круга задач
		УК-2.2.	Умеет определять круг задач в рамках поставленной цели, выбирать оптимальные способы решения задач, учитывая имеющиеся ресурсы и ограничения
		УК-2.3.	Имеет практический опыт применения знаний о правовых нормах и ресурсах в реальных ситуациях, разработки и реализации решений в соответствии с установленными ограничениями
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области
		ОПК-1.2.	Умеет анализировать и формулировать математические

			задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в понятной и доступной форме
		ОПК-1.3.	Имеет практический опыт работы над проектами или исследованиями в области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи
ПК-1.	Способен формулировать задачи с математической точностью, обосновывать утверждения строго и анализировать полученные результаты в области математики и компьютерных наук	ПК-1.1.	Знает методы и подходы к формулированию задач, а также основные принципы математического доказательства и анализа результатов.
		ПК-1.2.	Умеет корректно ставить и формулировать математические задачи, применять строгие методы доказательства и анализировать полученные результаты.
		ПК-1.3.	Имеет опыт работы с задачами в области математики и компьютерных наук, включая применение математических методов для решения практических задач
ПК-2.	Способен решать типовые задачи профессиональной деятельности в области искусственного интеллекта, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной безопасности	ПК-2.1.	Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности
		ПК-2.2.	Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования информационной безопасности
		ПК-2.3.	Имеет опыт работы с информационными ресурсами и технологиями в области искусственного интеллекта, включая соблюдение норм информационной безопасности

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары			
1	Основы и PAC-обучение	5	5		20	Домашнее задание Аудиторная работа
2	Неравенства и меры сложности	5	5	2	20	Домашнее задание Аудиторная работа Контрольная работа
3	SVM, ядра и ансамбли	5	5		20	Домашнее задание Аудиторная работа
4	Нейронные сети и регуляризация	5	5	2	20	Домашнее задание Аудиторная работа
5	Теория маржи и онлайн-обучение	5	5		20	Домашнее задание Аудиторная работа
6	Оценка качества и современные направления	2	2		10	Домашнее задание Аудиторная работа
7	Оценка качества и современные направления	3	3	2	10	Домашнее задание Аудиторная работа Контрольная работа
	Зачет с оценкой			4		Коллоквиум
	Итого:	***30***	***30***	***10***	***120***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основы и PAC-обучение	Постановка задачи. Риск и эмпирический риск PAC-обучение. Конечные классы гипотез
2	Неравенства и меры сложности	Хёффдинга, Чернова, ограниченных разностей VC-размерность. Примеры вычисления Радемахеровская сложность. Границы обобщения.
3	SVM, ядра и ансамбли	Метод опорных векторов (SVM) Ядерный трюк. RKHS Алгоритм AdaBoost
4	Нейронные сети и регуляризация	VC-размерность сетей. NTK. Структурная минимизация риска. L1, L2
5	Теория маржи и онлайн-обучение	Границы, основанные на марже Экспертные советы. Перцептрон
6	Оценка качества и современные направления	Кросс-валидация, доверительные интервалы для метрик Обучение с подкреплением. Открытые проблемы
7	Оценка качества и современные направления	От теории обобщения к практике. Сравнительный анализ подходов и современные направления

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Введение в статистическое обучение с примерами на языке R : учебник / Д. Гарет, Д. Уиттон, Т. Хасти, Р. Тибширани ; пер. с англ. С. Э. Мастицкого. - 2-е изд. испр. - Москва : ДМК Пресс, 2017. - 456 с. - ISBN 978-5-97060-495-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/1873554

2. Мастицкий, С. Э. Статистический анализ и визуализация данных с помощью R : практическое руководство / С. Э. Мастицкий, В. К. Шитиков. - 2-е изд. - Москва : ДМК Пресс, 2023. - 497 с. - ISBN 978-5-89818-601-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/2108480

Дополнительная литература:

1. Тимофеева, А. Ю. Вероятностные основы методов и алгоритмов анализа данных : учебное пособие / А. Ю. Тимофеева, Е. А. Хайленко. - Новосибирск : Изд-во НГТУ, 2020. - 82 с. - ISBN 978-5-7782-4166-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/1866892

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модуле), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное

Электронный документ

Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Статистическая теория обучения» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекция, семинары, контрольные работы, домашние задания, коллоквиум, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар — это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Аудиторная работа – активная работа студента на семинаре или лекции, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в аудиторной работе студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения, аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Контрольная работа – письменная работа с набором задач, которые нужно решить за ограниченное время.

Цель контрольной работы - получить специальные знания по одной или нескольким темам дисциплины (модуля) и продемонстрировать навыки их практического применения.

Коллоквиум – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модуле) «Статистическая теория обучения»

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***зачета с оценкой***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	**Пятибалльная оценка**	**Оценка за зачет**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Зачтено	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный
9	Отлично	Зачтено	
8	Отлично	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Зачтено	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и
4	Удовлетворительно	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
3	Не сдан	Не зачтено	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «Статистическая теория обучения» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	25%	10	Набор задач по темам недели
Контрольные работы	35%	2	Письменная работа с набором задач, которые нужно решить за ограниченное время
Аудиторная работа	15%	15	Активная работа студента на семинаре или лекции
Зачет с оценкой	25%	1	Коллоквиум - устные ответы на вопросы, список которых известен студенту заранее

Формула расчёта итоговой оценки по дисциплине (модуле) «Статистическая теория обучения»: «0,25 × среднее за домашние задания + 0,35 × среднее за контрольные работы + 0,15 × среднее за аудиторную работу + 0,25 ×зачет с оценкой».

Текущий контроль успеваемости обучающихся по дисциплине (модуле)

Примерные домашние задания

1. Даны выборка и функция потерь $\ell(y,f(x))=(y-f(x))2$. Запишите выражение для эмпирического риска. В чём разница между эмпирическим и истинным риском?
2. Приведите пример задачи, в которой высокая точность на обучающей выборке не гарантирует хорошего качества на тесте. Объясните, почему это происходит.
3. Постройте пример, иллюстрирующий компромисс между смещением и дисперсией при выборе степени полиномиальной регрессии.
4. Даны выборка и функция потерь $\ell(y,f(x))=(y-f(x))2$. Запишите выражение для эмпирического риска. В чём разница между эмпирическим и истинным риском?
5. Приведите пример задачи, в которой высокая точность на обучающей выборке не гарантирует хорошего качества на тесте. Объясните, почему это происходит.
6. Постройте пример, иллюстрирующий компромисс между смещением и дисперсией при выборе степени полиномиальной регрессии.
7. Даны выборка и функция потерь $\ell(y,f(x))=(y-f(x))2$. Запишите выражение для эмпирического риска. В чём разница между эмпирическим и истинным риском?
8. Приведите пример задачи, в которой высокая точность на обучающей выборке не гарантирует хорошего качества на тесте. Объясните, почему это происходит.
9. Постройте пример, иллюстрирующий компромисс между смещением и дисперсией при выборе степени полиномиальной регрессии.
10. Вычислите выборочную рademахеровскую сложность для множества из двух функций: $h1(x)=1$, $h2(x)=-1$, при фиксированной выборке $x1,\ldots,xn$.

Примерные задания по контрольной работе

1. Дана модель бинарной классификации с функцией потерь 0-1. Запишите выражения для истинного и эмпирического рисков. В чём заключается проблема минимизации эмпирического риска?
2. Что такое PAC-обучаемость? Сформулируйте определение в терминах ε и δ.
3. Найдите VC-размерность класса всех прямоугольников на плоскости, ориентированных по осям координат. Приведите краткое обоснование.
4. С помощью неравенства Хёфдинга оцените минимальный объём выборки n, необходимый для того, чтобы с вероятностью не менее 0.95 эмпирический риск отличался от истинного не более чем на 0.05.
5. Объясните, как рademахеровская сложность используется для получения неравенства обобщающей способности. Запишите общую форму границы.
6. Какой вклад вносит регуляризация в компромисс между смещением и дисперсией? Приведите пример из линейной регрессии.
7. В чём заключается идея структурной минимизации риска (SRM)? Как она реализуется в методе опорных векторов?
8. Пусть H — конечный класс гипотез, $|H|=50$. Получите верхнюю оценку вероятности того, что для некоторой гипотезы $h\in H$ разница между истинным и эмпирическим рисками превысит ε.

Примерные задания для семинаров

1. Постройте пример, когда минимизация эмпирического риска приводит к переобучению.
2. Найдите VC-размерность класса всех кругов на плоскости.
3. Постройте множество из 3 точек, разбиваемое всеми способами с помощью пороговых функций на прямой.

Электронный документ

4. Выведите границу обобщающей способности для конечного класса гипотез с помощью неравенства объединения и Хёфдинга.
5. Обсудите, почему равномерная сходимость важна для теории обучения.
6. Сравните границы, основанные на VC-размерности и рademахеровской сложности.
7. Покажите, как L2-регуляризация влияет на решение в линейной регрессии.
8. Нарисуйте график зависимости обобщающей ошибки от сложности модели.
9. Обозначьте участки переобучения и недообучения.
10. Обсудите: как кросс-валидация помогает выбрать оптимальный уровень регуляризации?

Примерные вопросы для семинаров

Основы онлайн-обучения

1. Что такое модель mistake-bound в онлайн-обучении и как она отличается от традиционных моделей обучения?
2. Опишите алгоритм Halving и объясните, как он работает в задаче бинарной классификации.
3. Как алгоритм Weighted Majority использует веса экспертов для принятия решений?
4. Докажите верхнюю границу на число ошибок для алгоритма Halving в худшем случае.
5. Как доказывается граница ошибок для алгоритма Weighted Majority?
6. Что такое экспертные советы (Expert Advice) и как они моделируются в онлайн-обучении?
7. Опишите алгоритм Hedge и его параметры, влияющие на производительность.
8. Как алгоритм Hedge адаптируется к изменяющимся весам экспертов?
9. Что такое задача прогнозирования с экспертами и как она решается с помощью онлайн-алгоритмов?
10. Объясните сходимость перцептрона в онлайн-обучении и условия его гарантированной работы.
11. Как работает ядерный перцептрон и в чем суть ядерного трюка?
12. Приведите пример реализации перцептрона на Python с использованием Scikit-learn.
13. Как ядерный трюк применяется для нелинейных классификаторов в онлайн-обучении?
14. Сравните производительность алгоритмов Halving, Weighted Majority и Hedge на синтетических данных.
15. Какие практические ограничения существуют при реализации онлайн-алгоритмов в реальных приложениях?

Теория PAC-обучения

1. Что такое реализуемая модель в теории PAC-обучения и как она отличается от нереализуемой?
2. Определите VC-размерность и приведите примеры для интервалов на прямой.
3. Как вычислить VC-размерность для прямоугольников в двумерном пространстве?
4. Сформулируйте фундаментальную теорему PAC-обучения и объясните её компоненты.
5. Как оценить достаточный размер выборки для достижения заданной точности в PAC-обучении?
6. Что такое равномерные неравенства и как они связаны с теоремой Гливенко-Кантелли?
7. Приведите пример нереализуемой модели и объясните, почему она такова.
8. Как VC-размерность влияет на сложность гипотезного класса?
9. Докажите, что VC-размерность для линейных классификаторов в R^d равна d+1.
10. Что такое ε-неточная гипотеза в контексте PAC-обучения?

Электронный документ

11. Как применить теорему PAC для оценки обобщающей способности алгоритма?
12. Опишите процесс вычисления VC-размерности для полиномиальных классификаторов.
13. В чем разница между PAC-обучением и традиционной статистической оценкой?
14. Приведите практический пример оценки VC-размерности для линейных моделей с помощью кода.
15. Как теория PAC-обучения применяется в задачах машинного обучения с большими данными?

Радемахеровская сложность

1. Что такое радемахеровская сложность и как она определяется для класса функций?
2. Объясните процесс симметризации в контексте радемахеровской сложности.
3. Как радемахеровская сложность связана с обобщающей способностью модели?
4. Сформулируйте неравенство ограниченных разностей и его применение.
5. Как оценить радемахеровскую сложность для линейных функций?
6. Что такое техники сжатия данных и как они используются в анализе сложности?
7. Приведите пример алгоритма сжатия для линейных моделей.
8. Как радемахеровская сложность применяется к оценке сложности нейросетей?
9. Объясните свойства радемахеровской сложности (конвексность, субаддитивность).
10. Как вычислить верхнюю границу радемахеровской сложности для класса выпуклых функций?
11. В чем разница между радемахеровской и VC-размерностью?
12. Приведите пример, где радемахеровская сложность дает более точную оценку, чем VC-размерность.
13. Как техники сжатия помогают в практической оценке обобщающей способности?
14. Опишите применение радемахеровской сложности к глубоким нейросетям.
15. Какие ограничения существуют при использовании радемахеровской сложности для сложных моделей?

Маржинальные методы

1. Что такое теория маржи в контексте SVM и как она влияет на обобщающую способность?
2. Как границы обобщения зависят от ширины маржи в SVM?
3. Сравните SVM и логистическую регрессию по их подходу к марже.
4. Как вычислить маржу для заданного классификатора?
5. Что такое мягкая маржа в SVM и когда она применяется?
6. Объясните роль ядер в маржинальных методах для SVM.
7. Как границы обобщения для маржинальных методов доказываются?
8. Приведите пример, где SVM с широкой маржой превосходит логистическую регрессию.
9. Что такое потеря hinge в SVM и как она связана с маржой?
10. Как оптимизировать маржу в задаче многоклассовой классификации?
11. Опишите практическую реализацию сравнения SVM и логистической регрессии на датасете.
12. В чем преимущества маржинальных методов для высокоразмерных данных?
13. Как маржа влияет на устойчивость модели к шуму в данных?
14. Приведите формулу для границы обобщения через маржу для линейных SVM.
15. Какие модификации маржинальных методов используются в современных алгоритмах?

Ядра и RKHS

1. Что такое теорема представления в контексте ядер и RKHS?
2. Объясните ядерный трюк и его роль в нелинейном обучении.
3. Какие свойства должны удовлетворять функции ядер?
4. Опишите RBF-ядро и его параметры.

Электронный документ

5. Как работает полиномиальное ядро и когда оно применяется?
6. Что такое RKHS (воспроизводящее ядро Гильбертово пространство) и как оно строится?
7. Приведите пример применения ядерного трюка к линейной регрессии.
8. Как вычислить грам-матрицу для заданного ядра?
9. В чем разница между положительно определенными и условно положительно определенными ядрами?
10. Объясните, как ядра позволяют решать задачи в высокомерных пространствах без явного отображения.
11. Как оптимизировать параметры RBF-ядра для лучшей производительности?
12. Приведите примеры популярных ядер помимо RBF и полиномиального.
13. Как RKHS используется в задачах регрессии с ядрами?
14. Что такое ядерные методы в задачах кластеризации?
15. Какие ограничения существуют при использовании ядер в больших наборах данных?

Бустинг

1. Опишите алгоритм AdaBoost и его основные шаги.
2. Как AdaBoost строит ансамбль слабых классификаторов?
3. Докажите границы ошибок для AdaBoost в терминах экспоненциальной потери.
4. Что такое слабый ученик в контексте AdaBoost?
5. Как AdaBoost адаптирует веса примеров на каждой итерации?
6. Приведите пример реализации AdaBoost на деревьях решений в Python.
7. В чем разница между AdaBoost и другими методами ансамблирования?
8. Как оценить производительность AdaBoost на несбалансированных данных?
9. Объясните роль экспоненциальной потери в AdaBoost.
10. Что такое переобучение в бустинге и как его избежать?
11. Как AdaBoost применяется к задачам регрессии?
12. Приведите практический пример сравнения AdaBoost с другими алгоритмами на датасете.
13. Как границы ошибок AdaBoost связаны с VC-размерностью?
14. Что такое градиентный бустинг и как он отличается от AdaBoost?
15. Какие модификации AdaBoost используются в современных библиотеках машинного обучения?

Нейросети и NTK

1. Что такое невыпуклость в оптимизации нейросетей и как она влияет на обучение?
2. Объясните концепцию lazy-режима в нейросетях.
3. Что такое ядерная теория нейросетей (NTK) и как она моделирует обучение?
4. Как NTK аппроксимирует поведение нейросети в бесконечной ширине?
5. В чем разница между традиционным обучением нейросетей и NTK-подходом?
6. Как невыпуклость приводит к проблемам в градиентном спуске?
7. Опишите, как NTK используется для анализа сходимости нейросетей.
8. Приведите пример, где lazy-режим упрощает анализ нейросети.
9. Как NTK связан с гауссовскими процессами?
10. Что такое бесконечная ширина в NTK и её implications?
11. Как невыпуклость влияет на ландшафт потерь в глубоких сетях?
12. Объясните, как NTK помогает в теоретическом анализе глубокого обучения.
13. Приведите практический пример применения NTK к простой нейросети.
14. В чем преимущества NTK для понимания обобщающей способности?
15. Какие ограничения существуют в применении NTK к реальным архитектурам?

Современные аспекты

1. Что такое теория маржи для SVM в современных контекстах?
2. Как границы обобщения через маржу применяются к глубоким моделям?

3. Сравните SVM и логистическую регрессию по их границам обобщения.
4. Как маржа влияет на устойчивость к adversarial примерам?
5. Опишите практическую реализацию сравнения SVM и логистической регрессии на реальном датасете.
6. Что такое маржинальные методы в контексте больших данных?
7. Как границы обобщения для маржи доказываются в современных теориях?
8. Приведите пример, где маржинальные методы превосходят другие подходы.
9. Что такое мягкая маржа и её роль в регуляризации?
10. Как оптимизировать маржу для многоклассовых задач?
11. В чем преимущества маржинальных методов для интерпретируемости?
12. Как маржа связана с понятием robustness в машинном обучении?
13. Приведите формулу для современной границы обобщения через маржу.
14. Как маржинальные методы интегрируются с нейросетями?
15. Какие вызовы возникают при применении маржинальных методов к высокоразмерным данным?

Заключение и подготовка к итоговой аттестации

1. Что такое Dropout и как он работает как метод регуляризации?
2. Как Dropout влияет на обобщающую способность нейросетей?
3. Объясните стабильность градиентного спуска и факторы, её влияющие.
4. Как регуляризация помогает предотвратить переобучение?
5. Приведите пример применения Dropout в практике глубокого обучения.
6. Что такое вариации Dropout (например, DropConnect) и их преимущества?
7. Как оценить стабильность градиентного спуска на практике?
8. В чем разница между L1 и L2 регуляризацией в контексте итоговой аттестации?
9. Как Dropout комбинируется с другими методами регуляризации?
10. Объясните, как нестабильность градиентного спуска приводит к проблемам в обучении.
11. Приведите пример, где регуляризация улучшает производительность модели.
12. Как подготовиться к вопросам по Dropout на экзамене?
13. Что такое адаптивные методы оптимизации и их связь со стабильностью?
14. Как оценить влияние регуляризации на границы обобщения?
15. Какие ключевые темы из курса следует повторить для успешной аттестации?

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Назовите модель, где алгоритмы обучаются последовательно, минимизируя число ошибок на каждом шаге.	mistake-bound	УК-1
2	Какой алгоритм онлайн-обучения использует взвешивание экспертов для прогнозирования?	Weighted Majority	УК-1
3	Укажите алгоритм, который экспоненциально уменьшает веса экспертов при ошибках.	Hedge	УК-1
4	Назовите алгоритм, гарантирующий сходимость для линейно разделимых данных.	Перцептрон	УК-1
5	В PAC-обучении укажите меру сложности гипотезного класса, определяющую его мощность.	VC-размерность	УК-2
6	Назовите теорему, связывающую VC-размерность с обобщающей способностью.	Фундаментальная теорема PAC-обучения	УК-2
7	Укажите неравенство, устанавливающее равномерную сходимость эмпирических мер.	Гливенко-Кантелли	УК-2
8	Назовите метод симметризации, используемый в	Симметризация	УК-2

	радемажеровской сложности.		
9	В радемажеровской сложности укажите неравенство для оценки разностей функций.	Неравенство ограниченных разностей	ОПК-1
10	Назовите метод, позволяющий оценивать обобщение через ширину разделяющей полосы.	Теория маржи для SVM	ОПК-1
11	Укажите теорему, которая позволяет представлять функции через ядра.	Теорема представления	ОПК-1
12	Назовите популярное ядро, основанное на евклидовом расстоянии.	RBF	ОПК-1
13	Укажите алгоритм бустинга, итеративно корректирующий веса примеров.	AdaBoost	ПК-1
14	В нейросетях назовите режим, где обучение похоже на ядерные методы.	lazy-режим	ПК-1
15	Назовите теорию, описывающую динамику обучения широких нейросетей.	NTK	ПК-1
16	Укажите технику регуляризации, случайно отключающую нейроны во время обучения.	Dropout	ПК-1
17	Назовите свойство, описывающее устойчивость градиентного спуска к шумам.	Стабильность градиентного спуска	ПК-2
18	В маржинальных методах укажите модель, максимизирующую маржу между классами.	SVM	ПК-2
19	Укажите трюк, позволяющий вычислять скалярные произведения в высокомерных пространствах.	Ядерный трюк	ПК-2
20	Назовите алгоритм, применяемый для сжатия данных в линейных моделях.	Алгоритмы сжатия для линейных моделей	ПК-2