

**Федеральное государственное учреждение
«Федеральный исследовательский центр «Информатика и управление» Российской
академии наук»
(ФИЦ ИУ РАН)**

Утверждена

Ученым советом ФИЦ ИУ РАН,
протокол № 1 от «27» ноября 2015 г.
Председатель Ученого совета,
директор ФИЦ ИУ РАН
_____ И.А. Соколов
«30» ноября 2015 г.

РАБОЧАЯ ПРОГРАММА

**УЧЕБНОЙ ДИСЦИПЛИНЫ
«Теория обучения машин»**

Направление подготовки
09.06.01 Информатика и вычислительная техника

Профиль (направленность программы)
15.13.17 Теоретические основы информатики

Квалификация выпускника
Исследователь. Преподаватель-исследователь

Форма обучения
очная

Москва, 2015

Направление подготовки: 09.06.01 Информатика и вычислительная техника

Профиль (направленность программы): 15.13.17 Теоретические основы информатики

Дисциплина: «Теория обучения машин»

Форма обучения: очная

Рабочая программа составлена с учетом ФГОС ВО по направлению подготовки 15.13.17 Теоретические основы информатики, утвержденного приказом Министерства образования и науки Российской Федерации от 30 июля 2014 года № 875, зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33685.

РАБОЧАЯ ПРОГРАММА РЕКОМЕНДОВАНА

Отделом интеллектуальных систем ФИЦ ИУ РАН

Руководитель отдела _____ /Воронцов К.В./

« ____ » _____ 201_ г.

ИСПОЛНИТЕЛИ (разработчики программы):

В.В.Воронцов, ФИЦ ИУ РАН, в.н.с., д.ф.-м.н., профессор

Рабочая программа зарегистрирована в аспирантуре под учетным номером _____ на правах учебно-методического издания.

Начальник отдела докторантуры и аспирантуры _____ / Клименко С..И. /

Оглавление

АННОТАЦИЯ	4
1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ	4
2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ	4
3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ	6
3.1. Структура дисциплины	6
3.2. Содержание разделов дисциплины	7
3.3. Семинарские занятия	19
3.4. Практические занятия	19
3.5. Самостоятельная работа.....	19
4. ТЕКУЩАЯ И ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ.	
ФОНД ОЦЕНОЧНЫХ СРЕДСТВ	21
5. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ	22
6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ	24

АННОТАЦИЯ

Дисциплина «Теория обучения машин» реализуется в рамках Блока 1 Основной профессиональной образовательной программы высшего образования - программ подготовки научно-педагогических кадров в аспирантуре Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по направлению подготовки 09.06.01 Информатика и вычислительная техника, профиль (направленность программы) 15.13.17 Теоретические основы информатики аспирантам очной формы обучения.

Рабочая программа разработана с учетом требований ФГОС ВО по направлению подготовки 09.06.01 Информатика и вычислительная техника, утвержденного приказом Министерства образования и науки Российской Федерации от 30 июля 2014 года № 875, зарегистрировано в Минюсте Российской Федерации 20 августа 2014 года № 33685.

Основным источником материалов для формирования содержания программы являются: материалы конференций, симпозиумов, семинаров, Интернет-ресурсы, научные издания и монографические исследования и публикации.

Общая трудоемкость дисциплины по учебному плану составляет - 2 зач.ед. (72 часов), из них лекций - 36 час., семинарских занятий – 0 час., практических занятий – 0 час., часов самостоятельной работы – 36 час. Дисциплина реализуется на 1 курсе, 2 семестре, продолжительность обучения – 1 семестр.

Текущая аттестация проводится не менее 2 раз в соответствии с заданиями и формами контроля, предусмотренными настоящей программой.

Промежуточная оценка знания осуществляется в период зачетно-экзаменационной сессии в форме:зачет.

1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели и задачи дисциплины «Теория обучения машин»

Цель:

Целью дисциплины является изучение аспирантами статистических и алгоритмических основ анализа сигналов и многомерных массивов данных, а также знакомство с практическими приложениями статистических методов анализа сигналов и многомерных массивов данных. Целью дисциплины является изучение основ теории обучения машин, современных методов восстановления зависимостей по эмпирическим данным, включая дискриминантный, кластерный и регрессионный анализ, овладение навыками практического решения задач интеллектуального анализа данных.

Задачи:

- освоение аспирантами базовых знаний (понятий, концепций, методов и моделей),
- приобретение теоретических знаний и практических умений и навыков, проведение собственных теоретических исследований,
- консультирование студентов в области машинного обучения и интеллектуального анализа данных.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс изучения дисциплины «Теория обучения машин» направлен на формирование компетенций или отдельных их элементов в соответствии с ФГОС ВО по направлению подготовки 09.06.01 Информатика и вычислительная техника, профиль (направленность программы) 15.13.17 Теоретические основы информатики аспирантам очной формы обучения:

а) универсальных (УК)

- способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях (УК-1);
- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
- способность следовать этическим нормам в профессиональной деятельности (УК-5);
- способность планировать и решать задачи собственного профессионального и личностного развития (УК-6).

б) общепрофессиональных (ОПК):

- владение методологией теоретических и экспериментальных исследований в области профессиональной деятельности (ОПК-1);
- владение культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий (ОПК-2);
- способность к разработке новых методов исследования и их применению в самостоятельной научно-исследовательской деятельности в области профессиональной деятельности (ОПК-3);
- готовность организовать работу исследовательского коллектива в области профессиональной деятельности (ОПК-4);
- способность представлять полученные результаты научно-исследовательской деятельности на высоком уровне и с учетом соблюдения авторских прав (ОПК-6);
- владение методами проведения патентных исследований, лицензирования и защиты авторских прав при создании инновационных продуктов в области профессиональной деятельности (ОПК-7);
- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-8).

в) профессиональных (ПК):

- готовность использовать знание основных методов искусственного интеллекта в последующей профессиональной деятельности в качестве научных сотрудников, преподавателей вузов, инженеров, технологов (ПК-1);
- готовность выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности в области моделирования и анализа сложных естественных и искусственных систем (ПК-3);
- способность к созданию математических и информационных моделей исследуемых процессов, явлений и объектов, относящихся к профессиональной сфере (ПК-4);
- способность применять на практике умения и навыки в организации исследовательских работ и проводить научные исследования, готовность к участию в инновационной деятельности (ПК-5).

В результате освоения дисциплины «Теория обучения машин» обучающийся должен:

Знать:

- фундаментальные понятия, современные подходы, методы и проблемы машинного обучения и интеллектуального анализа данных.

Уметь:

- понять и формализовать поставленную задачу анализа данных;
- использовать современные методы машинного обучения для практического решения задач анализа данных;
- при необходимости, продиктованной особенностями поставленной задачи, создавать новые методы машинного обучения;

- проводить численные эксперименты на модельных и реальных данных и интерпретировать их результаты;
- представлять результаты исследований в устной и письменной форме.

Владеть:

- навыками освоения большого объема информации и решения сложных теоретических и практических задач анализа данных;
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов и методов;
- предметным языком машинного обучения и интеллектуального анализа данных, навыками описания решения задач и представления полученных результатов.

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

3.1. Структура дисциплины

В курсе рассматриваются основные классы задач и методов машинного обучения и интеллектуального анализа данных: классификация, кластеризация, регрессия, понижение размерности. Изучаются методы их решения, как классические, так и новые, созданные за последние 10–15 лет. Упор делается на глубокое понимание математических основ, взаимосвязей, достоинств и ограничений рассматриваемых методов. Отдельные теоремы приводятся с доказательствами. Все методы излагаются по единой схеме: исходные идеи и эвристики, их формализация и математическая теория, описание алгоритма в виде слабо формализованного псевдокода, анализ достоинств, недостатков и границ применимости, пути устранения недостатков, сравнение с другими методами.

Распределение трудоемкости дисциплины по видам учебных работ

Вид учебной работы	Трудоемкость					
	общая		Из них			
	Зач. Ед.	Час.	Лекц.	Прак.	Сем.	Сам.р.
ОБЩАЯ ТРУДОЕМКОСТЬ по Учебному плану	2	108	36			36
<i>Аудиторные занятия</i>						
Лекции (Л)	1,0	36	36			
Практические занятия (ПЗ)						
Семинары (С)						
<i>Самостоятельная работа (СР) без учёта промежуточного контроля:</i>						
Самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к семинарским и практическим занятиям) и самостоятельное изучение тем дисциплины	1,0	36				36
<i>Вид контроля:</i> зачет (является составной частью кандидатского экзамена)						

3.2. Содержание разделов дисциплины

Общее содержание дисциплины

№ раздела	Наименование темы (раздела)	Содержание темы (раздела)	Количество часов
1	Основные понятия и примеры прикладных задач	Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных. Полигон алгоритмов классификации. Вероятностная постановка задачи обучения по прецедентам, принцип максимума правдоподобия и его связь с принципом минимизации эмпирического риска. Разновидности функций потерь и их вероятностная интерпретация.	4
2	Байесовский классификатор.	Непараметрическое оценивание плотности Принцип максимума апостериорной вероятности. Функционал среднего риска. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода. Наивный байесовский классификатор. Ядерная оценка плотности Парзена-Розенблатта. Одномерный и многомерный случаи. Метод парзеновского окна. Выбор функции ядра. Выбор ширины окна, переменная ширина окна. Робастное оценивание плотности. Непараметрический наивный байесовский классификатор.	4
3	Параметрическое оценивание плотности	Нормальный дискриминантный анализ. Многомерное нормальное распределение, геометрическая интерпретация. Выборочные оценки параметров	4

		многомерного нормального распределения. Матричное дифференцирование. Вывод оценок параметров многомерного нормального распределения. Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Линейный дискриминант Фишера. Проблемы мультиколлинеарности и переобучения. Регуляризация ковариационной матрицы. Робастное оценивание. Цензурирование выборки (отсев объектов-выбросов). Параметрический наивный байесовский классификатор. Метод редукции размерности Шурыгина.	
4	Разделение смеси распределений	Смесь распределений. ЕМ-алгоритм: основная идея, понятие скрытых переменных. «Вывод» алгоритма без обоснования сходимости. Псевдокод ЕМ-алгоритма. Критерий останова. Выбор начального приближения. Выбор числа компонентов смеси. Стохастический ЕМ-алгоритм. Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение ЕМ-алгоритма для её настройки.	3
5	Метрические методы классификации	Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Обобщённый метрический классификатор, понятие отступа. Метод потенциальных функций, градиентный алгоритм. Отбор эталонных объектов. Псевдокод: алгоритм СТОЛП. Функция конкурентного сходства, алгоритм FRiS-СТОЛП. Функционал полного скользящего контроля, формула быстрого вычисления для метода 1NN. Профиль компактности.	3
6	Линейные методы классификации	Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия. Квадратичная функция потерь, метод наименьших квадратов, связь с линейным дискриминантом Фишера. Метод стохастического градиента и частные случаи: адаптивный линейный	4

		элемент ADALINE, персептрон Розенблатта, правило Хэбба. Теорема Новикова о сходимости. Доказательство теоремы Новикова Эвристики: инициализация весов, порядок предъявления объектов, выбор величины градиентного шага, «выбивание» из локальных минимумов. Проблема переобучения, редукция весов (weight decay). Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели. Квадратичный (гауссовский) и лапласовский регуляризаторы.	
7	Логистическая регрессия	Гипотеза экспоненциальности функций правдоподобия классов. Теорема о линейности байесовского оптимального классификатора. Оценивание апостериорных вероятностей классов с помощью сигмоидной функции активации. Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Снова метод стохастического градиента, сглаженное правило Хэбба. Пример прикладной задачи: кредитный скоринг. Бинаризация признаков. Скоринговые карты и оценивание вероятности дефолта. Риск кредитного портфеля банка. Настройка порога решающего правила по критерию числа ошибок I и II рода. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации площади под ROC-кривой.	4
8	Многослойные нейронные сети	Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключаящего или. Полнота двухслойных сетей в пространстве булевых функций. Теоремы Колмогорова, Стоуна, Горбаня (без доказательства). Алгоритм обратного распространения ошибок. Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети. Метод послойной настройки сети. Подбор структуры сети: методы	4

		постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage).	
9	Сети Кохонена	Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена. Сети встречного распространения, их применение для кусочно-постоянной и гладкой аппроксимации функций.	2
10	Методы кластеризации	Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связанных компонент. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means). Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи. Алгоритм построения дендрограммы. Определение числа кластеров. Свойства сжатия/растяжения, монотонности и редуцируемости. Псевдокод редуцирующей версии алгоритма.	4
11	Критерии выбора моделей	Внутренние и внешние критерии. Эмпирические и аналитические оценки функционала полного скользящего контроля. Скользящий контроль, разновидности эмпирических оценок скользящего контроля. Критерий непротиворечивости. Регуляризация. Критерий Акаике (AIC). Байесовский информационный критерий (BIC). Агрегированные и многоступенчатые критерии.	2
12	Теория обобщающей способности	Теория Вапника-Червоненкиса. Функционал равномерного отклонения частот ошибок. Функция роста, ёмкость семейства алгоритмов. Структурная минимизация риска. Оценка «бритвы Оккама». Радемахеровская сложность семейства алгоритмов. Комбинаторная теория переобучения. Функционал вероятности переобучения. Граф расслоения-связности. Оценки	4

		расслоения-связности.	
13	Методы отбора признаков	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).	2
14	Композиции классификаторов, бустинг	Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование. Алгоритм AdaBoost. Экспоненциальная аппроксимация пороговой функции потерь. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга. Обобщение бустинга как процесса градиентного спуска. Теорема сходимости. Алгоритм AnyBoost. Простое голосование (комитет большинства). Эвристический алгоритм ComBoost. Идентификация нетипичных объектов (выбросов). Обобщение на большое число классов. Решающий список (комитет старшинства). Эвристический алгоритм. Стратегия выбора классов для базовых алгоритмов. Стохастические методы: бэггинг и метод случайных подпространств. Нелинейные алгоритмические композиции. Смесь экспертов, область компетентности алгоритма. Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический. Построение смесей экспертов с помощью ЕМ-алгоритма.	6
15	Логические методы классификации	Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. Асимптотическая эквивалентность статистического и энтропийного определения. Сравнение областей эвристических и статистических закономерностей. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости. «Градиентный» алгоритм синтеза	6

		конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Бинаризация признаков. Алгоритм разбиения области значений признака на информативные зоны. Решающий список. Жадный алгоритм синтеза списка. Решающее дерево. Псевдокод: жадный алгоритм ID3. Недостатки алгоритма и способы их устранения. Проблема переобучения. Редукция решающих деревьев: предредукция и постредукция. Преобразование решающего дерева в решающий список. Чередующиеся решающие деревья (alternating decision tree). Невнимательные решающие деревья (oblivious decision tree). Решающий лес и бустинг над решающими деревьями. Взвешенное голосование закономерностей. Методы синтеза конъюнктивных закономерностей. Алгоритмы КОРА и ТЭМП. Эвристики, обеспечивающие различность и полезность закономерностей. Построение Парето-оптимального фронта закономерностей. Выравнивание распределения отступов. Применение алгоритма бустинга AdaBoost к закономерностям. Критерий информативности в бустинге. Примеры прикладных задач: кредитный скоринг, прогнозирование ухода клиентов.	
16	Алгоритмы вычисления оценок.	Принцип частичной прецедентности. Структура Алгоритмов вычисления оценок. Тупиковые тесты. Тупиковые представительные наборы. Проблема оптимизации АВО. АВО как композиция метрических закономерностей. Применение бустинга, ТЭМП и СПА для оптимизации АВО.	2
17	Поиск ассоциативных правил	Понятие ассоциативного правила и его связь с понятием логической закономерности. Примеры прикладных задач: анализ рыночных корзин, выделение терминов и тематики текстов. Алгоритм APriori. Два этапа: поиск частых наборов и рекурсивное порождение ассоциативных правил. Недостатки и пути	4

		усовершенствования алгоритма APriori. Алгоритм FP-growth. Понятия FP-дерева и условного FP-дерева. Два этапа поиска частых наборов в FP-growth: построение FP-дерева и рекурсивное порождение частых наборов. Общее представление о динамических и иерархических методах поиска ассоциативных правил.	
18	Задачи с частичным обучением	Постановка задачи Semisupervised Learning, примеры приложений. Простые эвристические методы: self-training, co-training, co-learning. Адаптация алгоритмов кластеризации для решения задач с частичным обучением. Кратчайший незамкнутый путь. Алгоритм Ланса-Уильямса. Алгоритм k-средних. Трансдуктивный метод опорных векторов TSVM. Алгоритм Expectation-Regularization на основе многоклассовой регуляризированной логистической регрессии.	2
19	Коллаборативная фильтрация	Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты—объекты. Корреляционные методы user-based, item-based. Латентные методы на основе би-кластеризации. Алгоритм Брегмана. Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных. Метод стохастического градиента. Неотрицательные матричные разложения. Вероятностный латентный семантический анализ PLSA. EM-алгоритм для PLSA. Эксперименты на данных конкурса «Интернет-математика» 2005.	3
20	Тематическое моделирование	Задачи тематического моделирования, коллекции текстовых документов и матрица документы—слова. Перплексия как мера качества тематической модели. Задача тематического поиска. Униграммная модель документа. Метод максимума правдоподобия и метод максимума апостериорной вероятности. Применение метода множителей Лагранжа. Вероятностный латентный семантический анализ PLSA. EM-алгоритм. Инкрементное добавление новых документов (folding-in). Задача с частичным обучением. Латентное размещение Дирихле.	3

		Сглаженная частотная оценка вероятности. Сэмплирование Гиббса. Оптимизация гиперпараметров. Робастная тематическая модель с фоновой и шумовой компонентой. Эксперименты по сравнению робастных и регуляризованных моделей.	
21	Обучение с подкреплением	Задача о многоруком бандите. Жадные и эpsilon-жадные стратегии. Среда для экспериментов. Метод сравнения с подкреплением. Метод преследования. Адаптивные стратегии на основе скользящих средних. Уравнения Беллмана. Оптимальные стратегии. Динамическое программирование. Метод итераций по ценностям и по стратегиям. Методы временных разностей: TD, SARSA, Q-метод. Многошаговое TD-прогнозирование. Адаптивный полужадный метод VDBE.	2

Лекционный курс

Порядковый номер лекции	Раздел, тема учебного курса, содержание лекции	Трудоемкость	
		час.	зач. ед.*
1	1. Основные понятия и примеры прикладных задач Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал. типы задач: классификация, регрессия, прогнозирование, кластеризация. Примеры прикладных задач. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль. Методика экспериментального исследования и сравнения алгоритмов на модельных и реальных данных.	1	
2	Вероятностная постановка задачи обучения по прецедентам, принцип максимума правдоподобия и его связь с принципом минимизации эмпирического риска. Разновидности функций потерь и их вероятностная интерпретация.	1	
3	2. Байесовский классификатор. Непараметрическое оценивание плотности. Принцип максимума апостериорной вероятности. Функционал среднего риска. Ошибки I и II рода. Теорема об оптимальности байесовского классификатора. Оценивание плотности распределения: три основных подхода.	1	
4	Наивный байесовский классификатор. Ядерная оценка плотности Парзена-Розенблатта. Одномерный и многомерный случаи. Метод парзеновского окна. Выбор функции ядра. Выбор ширины	1	

5	окна, переменная ширина окна. Робастное оценивание плотности. Непараметрический наивный байесовский классификатор.	1	
6	3. Параметрическое оценивание плотности Нормальный дискриминантный анализ. Многомерное нормальное распределение, геометрическая интерпретация. Выборочные оценки параметров многомерного нормального распределения. Матричное дифференцирование. Вывод оценок параметров многомерного нормального распределения. Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Линейный дискриминант Фишера.	1	
7	Проблемы мультиколлинеарности и переобучения. Регуляризация ковариационной матрицы. Робастное оценивание. Цензурирование выборки (отсев объектов-выбросов). Параметрический наивный байесовский классификатор. Метод редукции размерности Шурыгина.	1	
8	4. Разделение смеси распределений Смесь распределений. ЕМ-алгоритм: основная идея, понятие скрытых переменных. «Вывод» алгоритма без обоснования сходимости. Псевдокод ЕМ-алгоритма. Критерий останова. Выбор начального приближения. Выбор числа компонентов смеси.	1	
9	Стохастический ЕМ-алгоритм. Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF) и применение ЕМ-алгоритма для её настройки.	1	
10	5. Метрические методы классификации Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Обобщённый метрический классификатор, понятие отступа.	1	
11	Метод потенциальных функций, градиентный алгоритм. Отбор эталонных объектов. Псевдокод: алгоритм СТОЛП. Функция конкурентного сходства, алгоритм FRiS-СТОЛП. Функционал полного скользящего контроля, формула быстрого вычисления для метода 1NN. Профиль компактности.	1	
12	6. Линейные методы классификации Линейный классификатор, непрерывные аппроксимации пороговой функции потерь. Связь с методом максимума правдоподобия. Квадратичная функция потерь, метод наименьших квадратов, связь с линейным дискриминантом Фишера. Метод стохастического градиента и частные случаи: адаптивный линейный элемент ADALINE, персептрон Розенблатта, правило Хэбба. Теорема Новикова о сходимости. Проблема переобучения, редукция весов (weight decay). Байесовская регуляризация. Принцип максимума совместного правдоподобия данных и модели.	1	

13	Квадратичный (гауссовский) и лапласовский регуляризаторы. 7. Логистическая регрессия Гипотеза экспоненциальности функций правдоподобия классов. Теорема о линейности байесовского оптимального классификатора. Оценивание апостериорных вероятностей классов с помощью сигмоидной функции активации.	1	
14	Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Настройка порога решающего правила по критерию числа ошибок I и II рода. Кривая ошибок (ROC curve). Алгоритм эффективного построения ROC-кривой. Градиентный метод максимизации площади под ROC-кривой.	1	
15	8. Многослойные нейронные сети Модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключяющего или. Полнота двухслойных сетей в пространстве булевых функций. Теоремы Колмогорова, Стоуна, Горбаня.	1	
16	Алгоритм обратного распространения ошибок. Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети. Метод послойной настройки сети. Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage).	1	
17	9. Сети Кохонена Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена.	1	
18	Сети встречного распространения, их применение для кусочно-постоянной и гладкой аппроксимации функций. 10. Методы кластеризации Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур.	1	
19	Графовые алгоритмы кластеризации. Выделение связанных компонент. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means). Агломеративная кластеризация, Алгоритм Ланса-Вильямса и его частные случаи. Алгоритм построения дендрограммы. Определение числа кластеров.	1	
20	Свойства сжатия/растяжения, монотонности и редуцируемости. 11. Критерии выбора моделей Внутренние и внешние критерии. Эмпирические и аналитические оценки функционала полного скользящего контроля. Скользящий контроль, разновидности эмпирических оценок скользящего контроля.	1	

21	Критерий непротиворечивости. Регуляризация. Критерий Акаике (AIC). Байесовский информационный критерий (BIC). Агрегированные и многоступенчатые критерии.	1	
22	12. Теория обобщающей способности Теория Вапника-Червоненкиса. Функционал равномерного отклонения частот ошибок. Функция роста, ёмкость семейства алгоритмов. Структурная минимизация риска. Оценка «бритвы Оккама».	1	
23	Радемахеровская сложность семейства алгоритмов. Комбинаторная теория переобучения. Функционал вероятности переобучения. Граф расслоения-связности. Оценки расслоения-связности.	1	
24	13. Методы отбора признаков 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).	1	
25	14. Композиции классификаторов, бустинг Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция. Взвешенное голосование. Алгоритм AdaBoost. Экспоненциальная аппроксимация пороговой функции потерь. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга.	1	
26	Обобщение бустинга как процесса градиентного спуска. Теорема сходимости. Алгоритм AnyBoost. Простое голосование (комитет большинства). Эвристический алгоритм ComBoost. Идентификация нетипичных объектов (выбросов). Обобщение на большое число классов. Решающий список (комитет старшинства). Эвристический алгоритм. Стратегия выбора классов для базовых алгоритмов.	1	
27	Стохастические методы: бэггинг и метод случайных подпространств. Нелинейные алгоритмические композиции. Смесь экспертов, область компетентности алгоритма. Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический. Построение смесей экспертов с помощью ЕМ-алгоритма.	1	
	15. Логические методы классификации Понятие логической закономерности. Эвристическое, статистическое, энтропийное определение информативности. Асимптотическая эквивалентность статистического и энтропийного определения. Разновидности закономерностей: конъюнкции пороговых предикатов (гиперпараллелепипеды), синдромные правила, шары, гиперплоскости.	1	

28	«Градиентный» алгоритм синтеза конъюнкций, частные случаи. Бинаризация признаков. Алгоритм разбиения области значений признака на информативные зоны. Решающий список. Жадный алгоритм синтеза списка. Решающее дерево. Редукция решающих деревьев: предредукция и постредукция. Преобразование решающего дерева в решающий список. Чередующиеся решающие деревья.	1	
29	Невнимательные решающие деревья. Решающий лес и бустинг над решающими деревьями.	1	
30	Взвешенное голосование закономерностей. Методы синтеза конъюнктивных закономерностей. Алгоритмы КОРА и ТЭМП. Эвристики, обеспечивающие различность и полезность закономерностей. Построение Парето-оптимального фронта закономерностей. Выравнивание распределения отступов.	1	
31	Применение алгоритма бустинга AdaBoost к закономерностям. Критерий информативности в бустинге. 16. Алгоритмы вычисления оценок. Принцип частичной прецедентности. Структура Алгоритмов вычисления оценок. Тупиковые тесты. Тупиковые представительные наборы. Проблема оптимизации АВО. АВО как композиция метрических закономерностей.	1	
32	Применение бустинга, ТЭМП и СПА для оптимизации АВО. 17. Поиск ассоциативных правил. Понятие ассоциативного правила и его связь с понятием логической закономерности. Примеры прикладных задач: анализ рыночных корзин, выделение терминов и тематики текстов. Алгоритм APriori. Два этапа: поиск частых наборов и рекурсивное порождение ассоциативных правил. Недостатки и пути усовершенствования алгоритма APriori. Алгоритм FP-growth. Понятия FP-дерева и условного FP-дерева. Два этапа поиска частых наборов в FP-growth: построение FP-дерева и рекурсивное порождение частых наборов.	1	
33	Общее представление о динамических и иерархических методах поиска ассоциативных правил. 18. Задачи с частичным обучением. Постановка задачи Semisupervised Learning, примеры приложений. Простые эвристические методы: self-training, co-training, co-learning. Адаптация алгоритмов кластеризации для решения задач с частичным обучением. Кратчайший незамкнутый путь. Алгоритм Ланса-Уильямса. Алгоритм k-средних. Трансдуктивный метод опорных векторов TSVM. Алгоритм Expectation-Regularization на основе многоклассовой регуляризированной логистической регрессии.	1	

34	19. Коллаборативная фильтрация Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты—объекты. Корреляционные методы user-based, item-based. Латентные методы на основе би-кластеризации. Алгоритм Брегмана. Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных. Метод стохастического градиента. Неотрицательные матричные разложения. Вероятностный латентный семантический анализ PLSA. EM-алгоритм для PLSA. Эксперименты на данных конкурса «Интернет-математика» 2005.	1	
35	20. Тематическое моделирование Задачи тематического моделирования, коллекции текстовых документов и матрица документы—слова. Перплексия как мера качества тематической модели. Задача тематического поиска. Униграммная модель документа. Метод максимума правдоподобия и метод максимума апостериорной вероятности. Применение метода множителей Лагранжа. Вероятностный латентный семантический анализ PLSA. EM-алгоритм. Инкрементное добавление новых документов (folding-in). Задача с частичным обучением. Латентное размещение Дирихле. Сглаженная частотная оценка вероятности. Сэмплирование Гиббса. Оптимизация гиперпараметров. Робастная тематическая модель с фоновой и шумовой компонентой. Эксперименты по сравнению робастных и регуляризованных моделей.	1	
36	21. Обучение с подкреплением Задача о многоруком бандите. Жадные и эпсилон-жадные стратегии. Среда для экспериментов. Метод сравнения с подкреплением. Метод преследования. Адаптивные стратегии на основе скользящих средних. Уравнения Беллмана. Оптимальные стратегии. Динамическое программирование. Метод итераций по ценностям и по стратегиям. Методы временных разностей: TD, SARSA, Q-метод. Многошаговое TD-прогнозирование. Адаптивный полужадный метод VDBE.	1	
ИТОГО		36	1,0

3.3. Семинарские занятия

Не предусмотрены

3.4. Практические занятия

Не предусмотрены

3.5. Самостоятельная работа аспирантов

Внеаудиторная самостоятельная работа аспирантов включает следующие виды деятельности:

- конспектирование и реферирование первоисточников и другой научной и учебной литературы;
- проработку учебного материала (по конспектам, учебной и научной литературе);
- изучение учебного материала, перенесенного с аудиторных занятий на самостоятельную проработку.

Содержание и объем самостоятельной работы аспирантов

Разделы и темы рабочей программы самостоятельного изучения	Перечень домашних заданий и других вопросов для самостоятельного изучения	Сроки выполнения (№ недели)	Трудоемкость	
			час.	зач. ед.
Байесовский классификатор.	Проработка учебного материала (по конспектам, учебной и научной литературе)	2н	2	0,06
Параметрическое оценивание плотности	Проработка учебного материала (по конспектам, учебной и научной литературе)	3н	2	0,06
Разделение смеси распределений	Проработка учебного материала (по конспектам, учебной и научной литературе)	4н	2	0,06
Метрические методы классификации	Проработка учебного материала (по конспектам, учебной и научной литературе)	4-5н	2	0,06
Линейные методы классификации	Проработка учебного материала (по конспектам, учебной и научной литературе)	6н	2	0,06
Логистическая регрессия	Проработка учебного материала (по конспектам, учебной и научной литературе)	7н	2	0,06
Многослойные нейронные сети Сети Кохонена	Проработка учебного материала (по конспектам, учебной и научной литературе)	8н	2	0,06
Методы кластеризации	Проработка учебного материала (по конспектам, учебной и научной литературе)	9н	2	0,06
Критерии выбора моделей	Проработка учебного материала (по конспектам,	10н	2	0,06

	учебной и научной литературе)			
Теория обобщающей способности Методы отбора признаков	Проработка учебного материала (по конспектам, учебной и научной литературе)	11н	2	0,06
Композиции классификаторов, бустинг	Проработка учебного материала (по конспектам, учебной и научной литературе)	12н	3	0,08
Логические методы классификации	Проработка учебного материала (по конспектам, учебной и научной литературе)	13н	3	0,08
Алгоритмы вычисления оценок.	Проработка учебного материала (по конспектам, учебной и научной литературе)	14н	2	0,06
Поиск ассоциативных правил	Проработка учебного материала (по конспектам, учебной и научной литературе)	15н	2	0,06
Задачи с частичным обучением	Проработка учебного материала (по конспектам, учебной и научной литературе)	16н	2	0,06
Коллаборативная фильтрация	Проработка учебного материала (по конспектам, учебной и научной литературе)	17н	1	0,03
Тематическое моделирование Обучение с подкреплением	Проработка учебного материала (по конспектам, учебной и научной литературе)	18н	1	0,03

4. ТЕКУЩАЯ И ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Текущая аттестация аспирантов. Текущая аттестация аспирантов проводится в соответствии с локальным актом ФИЦ ИУ РАН - Положением о текущей, промежуточной и итоговой аттестации аспирантов ФИЦ ИУ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной.

Текущая аттестация по дисциплине проводится в форме опроса, а также оценки вопроса-ответа в рамках участия обучающихся в дискуссиях и различных контрольных

мероприятиях по оцениванию фактических результатов обучения, осуществляемых преподавателем, ведущим дисциплину.

Объектами оценивания выступают:

- учебная дисциплина – активность на занятиях, своевременность выполнения различных видов заданий, посещаемость занятий;
- степень усвоения теоретических знаний и уровень овладения практическими умениями и навыками по всем видам учебной работы, проводимых в рамках семинаров, практических занятий и самостоятельной работы.

Оценивание обучающегося на занятиях осуществляется с использованием нормативных оценок по 4-х бальной системе (5-отлично, 4-хорошо, 3-удовлетворительно, 2-неудовлетворительно).

Промежуточная аттестация аспирантов. Промежуточная аттестация аспирантов по дисциплине проводится в соответствии с локальным актом ФИЦ ИУ РАН - Положением о текущей, промежуточной и итоговой аттестации аспирантов ФИЦ ИУ РАН по программам высшего образования – программам подготовки научно-педагогических кадров в аспирантуре и является обязательной.

Промежуточная аттестация по дисциплине осуществляется в форме экзамена в период зачетно-экзаменационной сессии в соответствии с Графиком учебного процесса по приказу (распоряжению заместителя директора по научной работе). Аспирант допускается к экзамену в случае выполнения аспирантом всех учебных заданий и мероприятий, предусмотренных настоящей программой. В случае наличия учебной задолженности (пропущенных занятий и (или) невыполненных заданий) аспирант отрабатывает пропущенные занятия и выполняет задания.

Оценивание аспиранта на промежуточной аттестации в форме зачета.

Оценка зачета (нормативная)	Требования к знаниям и критерии выставления оценок
Зачет	Аспирант при ответе демонстрирует содержание тем учебной дисциплины, владеет основными понятиями, имеет представление об особенностях теории машинного обучения, обладает навыком формализации задач анализа данных, способен проводить численные эксперименты на модельных и реальных данных и интерпретировать их результаты, изучил основные методы интеллектуального анализа данных.
Не зачет	Аспирант при ответе демонстрирует плохое знание значительной части основного материала. Не информирован или слабо разбирается в проблемах, и или не в состоянии наметить пути их решения.

Вопросы для контроля знаний:

1. В чем разница задач классификации, регрессии, прогнозирования, кластеризации? Приведите примеры.
2. Какова взаимосвязь принципа максимума правдоподобия и принципа минимизации эмпирического риска?
3. В чем разница между ошибками I и II рода?
4. Какие три основных подхода к оцениванию плотности распределения вам знакомы?
5. В чем состоит метод редукции размерности Шурыгина?
6. Приведите основные шаги/блоки ЕМ-алгоритма. Каков его критерий останова?

7. Опишите основную идею метода ближайших соседей.
8. Как производится отбор эталонных объектов с помощью алгоритмов СТОЛП/FriS-СТОЛП?
9. Какова природа связи линейного классификатора метода максимума правдоподобия?
10. Сформулируйте теорему Новикова о сходимости.
11. В чем состоит проблем переобучения?
12. Как решается задача настройки порога решающего правила по критерию числа ошибок I и II рода?
13. Опишите алгоритм обратного распространения ошибок для многослойных нейронных сетей.
14. Какие стратегии обучения для сетей Кохонена вам известны? Опишите их принципы.
15. Приведите несколько примеров задач кластеризации. Какие алгоритмы кластеризации вам известны?
16. Сформулируйте основные критерии выбора моделей.
17. Приведите основные положения комбинаторной теории обучения.
18. В чем основная сложность задачи отбора признаков? Какие алгоритмы могут применяться для решения этой задачи?
19. Опишите основные принципы работы алгоритма AdaBoost.
20. Какие различные определения информативности вам известны?
21. Опишите основные алгоритмы логической классификации (достоинства, недостатки).
22. В чем состоит принцип частичной прецедентности?
23. Охарактеризуйте связь ассоциативного правила с понятием логической закономерности.
24. Приведите пример и формальную постановку задачи Semisupervised Learning?
25. Как адаптировать алгоритмы кластеризации для решения задач с частичным обучением?
26. В чем состоит задача коллаборативной фильтрации? Какие основные методы решения этой задачи?
27. Опишите метод максимума правдоподобия и метод максимума апостериорной вероятности для задачи тематического моделирования.
28. Сформулируйте задачу обучения с подкреплением. Какие основные алгоритмы решения этой задачи вам известны?

5. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Основная литература

- [1] Воронцов К. В. Математические методы обучения по прецедентам. 2012.
[http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_\(курс_лекций,_К.В.Воронцов\)](http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_(курс_лекций,_К.В.Воронцов)).
- [2] Мерков А. Б. Распознавание образов. Введение в методы статистического обучения. Едиториал УРСС. 2011. 256 стр.

- [3] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer: Data Mining, Inference, and Prediction. — 2nd ed. — Springer-Verlag. 2009. — 746 p.
- [4] C. M. Bishop. Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics. 2006. — 738 p.

Дополнительная литература

- [5] Айвазян С. А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Классификация и снижение размерности. — М. Финансы и статистика. 1989.
- [6] Айвазян С. А., Енюков И.С., Мешалкин Л.Д. Исследование зависимостей. — М. Финансы и статистика. 1985.
- [7] Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. — М.: Юнити, 1998.
- [8] Вагин В. Н., Головина Е. Ю., Загорянская А. А, Фомина М. В. Достоверный и правдоподобный вывод в интеллектуальных системах. — М.: Физматлит. 2004.
- [9] Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука. 1979.
- [10] Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. <http://www.ccas.ru/voron>.
- [11] Головкин В. А. Нейронные сети: обучение, организация и применение. — М.: ИПРЖР. 2001.
- [12] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.
- [13] Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.
- [14] Ивахненко А. Г., Юрачковский Ю. П. Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1987.
- [15] Журавлёв Ю.И. Рязанов В.В. Сенько О.В. РАСПОЗНАВАНИЕ. Математические методы. Программная система. Применения. — Москва: Фазис, 2006.
- [16] Журавлев Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. — 1978. — Т. 33. — С. 5–68.
- [17] Казанцев В. С. Задачи классификации и их программное обеспечение. — М. Наука. 1990.
- [18] Лоусон Ч, Хенсон Р. Численное решение задач метода наименьших квадратов. — М. Наука. 1986.
- [19] Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика: начальный курс. М.: Дело. 2004.
- [20] Саттон Р.С., Барто Э.Г. Обучение с подкреплением. — БИНОМ, 2011.
- [21] Хардле В. Прикладная непараметрическая регрессия. — М.: Мир. 1993.
- [22] Шурыгин А. М. Прикладная стохастика: робастность, оценивание, прогноз. — М. Финансы и статистика. 2000.
- [23] Burges C. J. C. A tutorial on support vector machines for pattern recognition // Data Mining and Knowledge Discovery. — 1998. — Vol. 2, no. 2. — Pp. 121–167. <http://citeseer.ist.psu.edu/burges98tutorial.html>.
- [24] Martin J. K. An exact probability metric for decision tree splitting and stopping // Machine Learning. — 1997. — Vol. 28, no. 2-3. — Pp. 257–291. <http://citeseer.ist.psu.edu/martin97exact.html>.
- [25] Marchand M., Shawe-Taylor J. Learning with the set covering machine // Proc. 18th International Conf. on Machine Learning. — Morgan Kaufmann, San Francisco, CA, 2001. — Pp. 345–352. <http://citeseer.ist.psu.edu/452556.html>.
- [26] Schapire R. The boosting approach to machine learning: An overview // MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA. — 2001. <http://citeseer.ist.psu.edu/schapire02boosting.html>.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

В соответствии с федеральными государственными требованиями к структуре основной профессиональной образовательной программы послевузовского профессионального образования по специальности 15.13.17 программа специальности «Теоретические основы информатики» предусматривает широкое использование в учебном процессе активных и интерактивных форм проведения занятий (компьютерное моделирование и визуализация, разбор конкретных ситуаций) в сочетании с внеаудиторной работой с целью формирования и развития профессиональных навыков обучающихся.

ИСПОЛНИТЕЛИ (разработчики программы):

д.ф.-м.н., проф. К.В.Воронцов

« ____ » _____ 2015 г.