

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «МЭИ»



«УТВЕРЖДАЮ»

Проректор по научной работе

Драгунов В.К.

«24» декабря 2017 г.

Программа аспирантуры

Направление 09.06.01 – Информатика и вычислительная техника

Направленность (специальность) 05.13.15 – Вычислительные машины, комплексы и компьютерные сети

РАБОЧАЯ ПРОГРАММА

дисциплины по выбору

«Машинное обучение в инженерии знаний»

Индекс дисциплины по учебному плану: Б1.В.ДВ.4.1

Всего: 108 часов

Семестр 7, в том числе 6 часов – контактная работа,
84 часов – самостоятельная работа,
18 часов – контроль

Москва 2017

Программа составлена на основе федерального государственного образовательного стандарта высшего образования (уровень подготовки кадров высшей квалификации) по направлению подготовки 09.06.01 Информатика и вычислительная техника, утвержденного приказом Минобрнауки России от 30 июля 2014 г. № 875, и паспорта специальности, 05.13.15 Вычислительные машины, комплексы и компьютерные сети, указанной в номенклатуре специальностей научных работников, утвержденной приказом Минобрнауки России от 25 февраля 2009 г. № 59.

ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью изучения дисциплины является постижение актуальных проблем организации вычислений и подходов к их решению в направлении создания перспективных средств вычислительной техники.

Задачами дисциплины являются:

- ознакомление аспирантов с новыми направлениями развития вычислительных систем на основе решения во многом неочевидных проблем организации вычислений;
- демонстрация конкретных путей, методов и средств решения указанных проблем;
- развитие творческих способностей обучающихся в части построения формализованных моделей с их доведением до работающей программой реализации;
- направление обучающихся аспирантов на поиск новых архитектурных и структурных решений в построении преимущественно распределенных вычислительных систем;
- подтверждение работоспособности изученных средств посредством опытного макетирования в сетевой среде персональных компьютеров.

В процессе освоения дисциплины **формируются следующие компетенции:**

- 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);
- способность проектировать и осуществлять комплексные исследования, в том числе междисциплинарные, на основе целостного научного мировоззрения с использованием знаний в области истории и философии науки (УК-2);

- готовность участвовать в работе российских и международных исследовательских коллективов по решению научных и научно-образовательных задач (УК-3);
- 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 объективно оценивать результаты исследований и разработок, выполненных другими специалистами и в других научных учреждениях (ОПК-5);
- готовность к преподавательской деятельности по основным образовательным программам высшего образования (ОПК-8);
- знание современных теоретических и экспериментальных методов исследования и анализа современных вычислительных систем, комплексов и компьютерных сетей (ПК-2);
- знание традиционных и новых методов работы информационных систем (ПК-3).

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБРАЗОВАНИЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

В результате освоения дисциплины обучающийся должен демонстрировать следующие результаты образования:

знать:

- традиционные и новые методы работы информационных систем (ПК-3);
- методологию теоретических и экспериментальных исследований в области инженерии знаний (ОПК-1);

уметь:

- оценивать качество проектных решений аппаратных или программных средств вычислительной техники (ПК-2);
- критически анализировать и оценивать современные научные достижения и генерировать новые научные идеи (УК-1);
- проектировать и осуществлять комплексные научные исследования в области поискового проектирования вычислительных систем (УК-2);
- объективно оценивать результаты исследований и разработок, выполненных другими специалистами и в других научных учреждениях (ОПК-5);
- работать в научных и исследовательских коллективах (УК-3).

владеть:

- культурой научного исследования, в том числе с использованием современных информационно-коммуникационных технологий (ОПК-2);
- основными навыками технического обучения (ОПК-8).

КРАТКОЕ СОДЕРЖАНИЕ РАЗДЕЛОВ ДИСЦИПЛИНЫ

1. *Машинное обучение как направление развития инженерии знаний*

Машинное обучение (англ. *Machine Learning*) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Основные виды обучения. *Индуктивное обучение*, основанное на выявлении закономерностей в эмпирических данных. *Дедуктивное обучение* предусматривающее формализацию знаний экспертов с их переносом компьютер в виде базы знаний. Превалирующее влияние методов и средств индуктивного обучения в преодолении «информационного взрыва» в среде обитания современного сообщества. Типы решаемых задач: классификация, регрессия, прогнозирование, ранжирование. Основные понятия: модель алгоритмов, метод обучения, функция потерь, функционал качества, 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. Конкурсы по анализу данных [kaggle.com](https://www.kaggle.com). Полигон алгоритмов классификации. CRISP-DM — межотраслевой стандарт ведения проектов интеллектуального анализа данных.

2. *Типизация методов и средств машинного обучения*

Системообразующие признаки: индукция/ дедукция, с учителем/ без учителя, известное/заранее неизвестное число получаемых классов, искомые классы пересекающиеся/ непересекающиеся, задаваемая метрика: Евклида, Хемминга, Маханолобиса и др., по областям эффективного применения: задачи интерполяции/экстраполяции. Показатели эффективности обучения: на обучающей выборке/ на тестовой выборке. Эмпирический и обобщенный риски. Метрические методы классификации и регрессии. Гипотезы компактности и непрерывности. Обобщенный метрический классификатор. Метод ближайших соседей k NN и его обобщения. Подбор числа k по критерию скользящего контроля. с постоянной и переменной шириной окна.. Метод потенциальных функций и его связь с линейной моделью классификации. Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Ядерное сглаживание

3. *Методы машинного обучения «с учителем»*

Добывание знаний на основе машинного обучения. 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. Решающие списки и искомые деревья. 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.

4. Методы машинного обучения «без учителя»

Кластеризация. Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент. Кратчайший незамкнутый путь. Алгоритм ФОРЭЛ. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k -means). Обучение с подкреплением. Задача о многоруком бандите. Жадные и эpsilon-жадные стратегии. Метод UCB (upper confidence bound). Стратегия Softmax. Среда для экспериментов. Адаптивные стратегии на основе скользящих средних. Метод сравнения с подкреплением. Метод преследования. Общая постановка задачи обучения с подкреплением. Ценность состояния среды. Ценность действия. Метод временных разностей. Метод SARSA. Метод Q-обучения. Многошаговое TD-прогнозирование. Ранжирование. Постановка задачи обучения ранжированию. Примеры. Признаки в задаче ранжирования поисковой выдачи: текстовые, ссылочные, кликовые. TF-IDF. PageRank. Критерии качества ранжирования: Precision, MAP, AUC, DCG, NDCG, pFound. Ранговая классификация, OS-SVM. Парный подход: RankingSVM, RankNet, LambdaRank.

5. Машинное обучение в дальнейшем развитии новых информационных технологий

Теория обобщающей способности. Теория Вапника -Червоненкиса. Функционал равномерного отклонения частот ошибок. Функция роста, ёмкость семейства алгоритмов. Структурная минимизация риска. Эффект «бритвы Оккама». Радемахеровская сложность семейства алгоритмов. Комбинаторная теория переобучения. Функционал вероятности переобучения. Граф и оценки расслоения-связности. Методы отбора признаков. Сложность задачи отбора признаков. Полный перебор. Поиск в глубину, метод ветвей и границ. Усечённый поиск в ширину, многорядный

итерационный алгоритм. Случайный поиск и Случайный поиск с адаптацией. Проблема извлечения знаний из неструктурированных текстов. Технологии Text Mining. Автоматическая классификация текстовых документов и Web-страниц. Расширение возможностей на основе использования лингвистических построений естественного языка. Совершенствование поисковых механизмов Интернета. Эксперименты по обучению беспилотному вождению транспортных средств. Google-автомобиль и его технические характеристики.

ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ РЕЗУЛЬТАТОВ ОБРАЗОВАНИЯ ПО ДИСЦИПЛИНЕ

Промежуточная аттестация по итогам освоения дисциплины: 7-й семестр – дифференцированный зачет.

Вопросы для самоконтроля и проведения зачета

1. Что означает инженерия знаний? (ОПК-1)
2. Что представляет собою машинное обучение в области искусственного интеллекта?(ПК-3)
3. Что подразумевается под программным, лингвистическим и математическим обеспечением средств машинного обучения? (ПК-3)
4. Какова иерархическая структура уровней кластеризации изучаемых объектов? (ОПК-1)
5. Каковы структура и основные разновидности методов машинного обучения? (ОПК-8)
6. Перечислите основные этапы решения задач машинного обучения? (ОПК-8)
7. Какие компоненты включаются в математическое обеспечение? (ПК-2)
8. Какой математический аппарат используется в моделях разного иерархического уровня? (ПК-3)
9. Каковы требования к математическим моделям и численным методам? (ОПК-8)
10. Что включает в себя математическое обеспечение инструментальных средств машинного обучения? (ПК-3)
11. В чем заключается задача не параметрической оптимизации эмпирического риска? (ОПК-8)
12. Каким образом классифицируются задачи машинного обучения без учителя? (ПК-3)
13. Каковы особенности методов обучения с подкреплением? (ОПК-1)
14. Перечислите уже существующие аналитические платформы и редакторы задач машинного обучения? (ОПК-2)
15. Каковы возможности, достоинства и недостатки ПО PoluAnlist Deducter? (ПК-2)
16. Назовите основные характеристики доступных аппаратных и программных средств для реализации процессов извлечения знаний? (ПК-2)

17. Объясните суть и области эффективного применения метода k-ближайших соседей? (ПК-3)
18. В чем состоит отличие метода C- means от метода k-ближайших соседей? (ПК-3)
19. Назовите возможности и особенности применения методов обобщения данных в виде дерева? (ПК-3)
20. Приведите классификацию и охарактеризуйте средства реализации технологии прогнозирования эффективных решений? (ОПК-2)
21. Расскажите о назначении знаменитого метода Page Rank (разработчики С.Брин и Л.Пейдж), обеспечивающего эффективное ранжирование Web – страниц в Google? (ОПК-5)

Критерии оценки за освоение дисциплины определены в Инструктивном письме И-23 от 14 мая 2012 г.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Основная литература:

1. Применение решателя открытых задач: методические указания по курсу «Поисковое проектирование вычислительных систем» / Дзегеленок И.И., Кондратьев К.А., Аляева Ю.В.— М.: Издательский дом МЭИ, 2012—50 с. УДК 621.398 Д431

Дополнительная литература:

2. Интернетика. Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов . – М. : Эдиториал УРСС, 2009 . – 264 с. - ISBN 978-5-397-00497-8.
3. Дюк В.А., Самойленко А.П. Data Mining. Учебный курс.–СПб: Питер, 2001. – 368 с. - ISBN 5-318-00227-7.
4. Методы и модели анализа данных: OLAP и Data Mining : учебное пособие по специальности 071900 "Информационные системы и технологии" направления 654700 "Информационные системы" / А. А. Барсегян, и др. . – СПб. : БХВ-Петербург, 2004 . – 336 с. + CD-ROM . - ISBN 5-941575-22-X .

5. Пенроуз Роджер.–Новый ум короля: о компьютерах, мышлении и законах физики.– Едиториал УРСС–2011.–400 с. - ISBN 978-5-453-00013-5.