АННОТАЦИЯ
рабочей программы дисциплины

Методы машинного обучения

1. Общая трудоёмкость

Трудоёмкость дисциплины составляет 5 зачётных единиц (180 часов), из них 18 часов лекционных занятий, 36 часов практических занятий, 126 часов самостоятельной работы.

2. Место дисциплины в структуре образовательной программы

Дисциплина относится к модулю обязательных профессиональных дисциплин обязательной части образовательной программы.

Данная дисциплина опирается на базовые знания, умения и навыки, формируемые при получении предшествующего уровня образования.

Знания, умения и навыки, формируемые данной дисциплиной, потребуются при освоении следующих элементов образовательной программы:

- Математические методы анализа больших данных.
- Технологии анализа больших данных.
- Экспертные системы и базы знаний.
- Программирование на языке Python.
- Математические методы и модели поддержки принятия решений.
- Информационный поиск и обработка естественного языка.
- Нейронные сети и глубокое обучение.

3. Цель изучения дисциплины

Формирование у обучающихся способности совершенствовать, разрабатывать и внедрять новые методы, модели, алгоритмы машинного обучения.

4. Содержание дисциплины

Модуль 1. Математическое и алгоритмическое обеспечение аналитики больших данных

Введение в аналитику больших данных (Big Data). Обзор задач интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей. Понятие Data Mining. OLAP и Data Mining. Данные, информация и знания. Единая методология обнаружения знаний. Задача регрессии. Обучение с учителем. Задача классификации. Обучение без учителя. Задача кластеризации. Задача анализа ассоциаций и последовательностей. Программное обеспечение для интеллектуального анализа данных: Weka, R, SAS Enterprise Miner, SPSS Modeler. Большие данные. Масштабируемые алгоритмы. Hadoop, MapReduce. Microsoft Azure Machine Learning, RapidMiner.

Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные вектора, байесовские классификаторы. Индукция деревьев решений. Информационный выигрыш. Индекс Gini. «Обрезка» деревьев: предредукция и постредукция. Решающие правила. Алгоритмы ID3, CART, C4.5. Алгоритм «случайный лес». Алгоритмы ограниченного перебора. Метод опорных векторов, линейная и нелинейная разделимость. Байесовская и наивная байесовская классификация.

Оценка эффективности и сравнительный анализ моделей обучения. Подготовка данных. Выбор значимых признаков. Наборы данных. Типы данных. Шкалы измерений. Форматы хранения данных. Качество данных. Очистка данных. Работа с дубликатами и пропущенными значениями. Снижение размерности данных. Интеграция данных. BI и визуализация данных. Методы отбора значимых признаков. Фильтры. Оболочки. Встроенные методы. Матрица несоответствий. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, специфичность. Обучающее множество. Независимое тестовое множество.

Подтверждающее множество. Проблема переобучения. Метод удержания. Метод перекрестной проверки. Метод «без одного». Стратификация данных. Метод самонастройки (бутстреп). Матрица стоимостей ошибок. Диаграмма выигрыша. Диаграмма роста. Кривая ошибок (ROC-кривая). AUC. Изолинии точности.

Ансамблирование классификаторов. Ансамбли (комитеты) моделей. Бэггинг. Бэггинг с рандомизацией. Последовательно обучающиеся классификаторы. Бустинг (усиление) ансамбля классификаторов. Алгоритм AdaBoost. Стэкинг.

Методы кластерного анализа. Алгоритмы машинного обучения: метод k-средних, EM, Cobweb. Типологический и таксономический анализ. Статистические методы кластеризации. EM-алгоритм. Метод k-средних. Меры расстояний. Иерархические методы кластеризации. Визуализация кластеров. Дендрограммы. Диаграммы рассеивания. Самоорганизующиеся карты Кохонена. Концептуальная кластеризация. Алгоритм Cobweb. Графовые методы кластеризации. Выделение связных компонент. Нечеткая кластеризация. FCM-алгоритм.

Методы анализа ассоциаций и последовательностей. Поиск часто встречающихся наборов элементов. Меры интересности: поддержка, достоверность, лифт. уверенность. Алгоритм Apriori. Задача анализа рыночных корзин. Количественные и нечеткие ассоциативные правила. Поиск последовательных шаблонов.

Введение в методы сетевого анализа (Social Network Analysis). Анализ связей. Понятия социальной сети и социального графа. Теория «малого мира». Диаметр, радиус, коэффициент кластеризации. Центральность: по степени, по близости, на основе собственных векторов. Позиционный и ролевой анализ в социальной сети.

Модуль 2. Прикладной анализ данных

Машинное обучение и большие данные для решения прикладных задач. Анализ рыночных корзин, «умные» рекламные объявления, совместная фильтрация, анализ тональности, чат-боты и др.

Персонализация методами машинного обучения. Массовая кастомизация, индивидуальный маркетинг. Методы и техники персонализации в интернете. Полуавтоматическая и автоматическая персонализация. Понятие адаптивного веб-ресурса. Цели, методы и приемы адаптации. Архитектура адаптивного веб-ресурса. Оверлейная и стереотипная модели пользователя. Формальная постановка задачи персонализации. Оценка эффективности рекомендаций. Рекомендации на основе композиционного правила нечеткого вывода. Методы совместной фильтрации. Теоретико-графовый метод хортинга.

5. Дополнительная полезная информация

Дисциплина предназначена для формирования элементов следующих компетенций образовательной программы:

ПК-1. Способен адаптировать и применять методы и алгоритмы машинного обучения для решения прикладных задач в различных предметных областях

ПК-2. Способен руководить проектами по созданию систем искусственного интеллекта с применением новых методов и алгоритмов машинного обучения со стороны заказчика

По дисциплине предусмотрены следующие методы обучения и интерактивные формы проведения занятий: лекции-визуализации с использованием презентационного материала; практические занятия, которые способствуют разнообразному (индивидуальному, групповому, коллективному) изучению (усвоению) учебных вопросов (проблем), активному взаимодействию обучающихся и преподавателя, живому обмену мнениями между ними, нацеленному на выработку правильного понимания содержания изучаемой темы и способов ее практического использования.

Наряду с традиционными образовательными технологиями, для реализации дисциплины могут использоваться технологии электронного обучения и дистанционные образовательные технологий в электронной информационно-образовательной среде университета. Лекционные и практические занятия могут проводиться с использованием платформ Microsoft Teams, Cisco, Moodle (BigBlueButton) и др.

Формой промежуточной аттестации является дифференцированный зачет (1 семестр).