Приложение 10 ОП ВО

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ
Интеллектуальный анализ данных
наименование дисциплин (модуля)/практики

Автор: Артамонова И. Б.

Код и наименование направления подготовки, профиля:

38.04.05 Бизнес-информатика, профиль Бизнес-аналитика

Квалификация (степень) выпускника: Магистр

Форма обучения: Очная

Цель освоения дисциплины:
Сформировать компетенции:
ПК-4 -способностью разрабатывать стратегию развития архитектуры предприятия
ПК-9 - способностью разрабатывать и внедрять компоненты архитектуры предприятия

План курса:

Раздел 1 Введение, основные понятия анализа данных
Введение в машинное обучение и анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Постановки задач машинного обучения. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.
Раздел 2 Математические объекты и методы в анализе данных
Линейная алгебра и анализ данных. Линейные пространства, их примеры из машинного обучения (признаки в кредитом скоринге, векторные представления текстов). Коллинеарность и линейная независимость. Скалярное произведение, косинус угла, примеры их применения. Векторы и матрицы, операции над ними. Матричное умножение. Системы линейных уравнений. Обратная матрица.
Математический анализ и анализ данных (на примере парной линейной регрессии и МНК). Производная и градиент, их свойства и интерпретации. Типы функций: непрерывные, разрывные, гладкие. Градиентный спуск. Выпуклые функции и их особое место в оптимизации.
Теория вероятностей и анализ данных. Случайные величины. Дискретные и непрерывные распределения, их свойства. Примеры распределений и их важность в анализе данных: биномиальное, пуассоновское, нормальное, экспоненциальное. Характеристики распределений: среднее, медиана, дисперсия, квантили. Пример их использования при генерации признаков. Центральная предельная теорема.
Математическая статистика и анализ данных. Оценивание параметров распределений. Метод максимального правдоподобия. Пример использования: анализ текстов и наивный байесовский классификатор. Доверительные интервалы и бутстрэппинг.
Раздел 3 Линейная регрессия и классификация
Линейная регрессия. Квадратичная функция потерь и предположение о нормальном распределении шума. Метод наименьших квадратов: аналитическое решение и оптимизационный подход. Стохастический градиентный спуск. Тонкости градиентного спуска: размер шага, начальное приближение, нормировка признаков. Проблема

переобучения. Регуляризация.
Линейная классификация. Аппроксимация дискретной функции потерь. Отступ. Примеры аппроксимаций, их особенности. Градиентный спуск, регуляризация. Классификация и оценки принадлежности классам. Кредитный скоринг. Логистическая регрессия: откуда берется такая функция потерь и почему она позволяет предсказывать вероятности. Максимизация зазора как пример регуляризации и устранения неоднозначности решения.

Раздел 4 Оценивание качества алгоритмов

Регрессия: квадратичные и абсолютные потери, абсолютные логарифмические отклонения. Примеры использования.
Классификация: доля верных ответов, ее недостатки. Точность и полнота, их объединение: арифметическое среднее, минимум, гармоническое среднее (F-мера).
Оценки принадлежности классам: площади под кривыми. AUC-ROC, AUC-PRC, их свойства.
Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.
Практические особенности кросс-валидации. Стратификация. Потенциальные проблемы с разбиением зависимой или динамической выборки.

Раздел 5 Логические методы

Логические методы и их интерпретируемость. Простейший пример: список решений. Пример решающего списка для задачи фильтрации нежелательных сообщений. Деревья решений. Проблема построения оптимального дерева решений. Жадный алгоритм, основные его параметры.
Построение деревьев решений. Критерий ветвления. Выбор оптимального разбиения в задачах регрессии. Сложности выбора разбиения в задаче классификации. Примеры критериев: энтропийный (прирост информации), Джини и их модификации. Критерии завершения построения. Регуляризация и стрижка деревьев.

Раздел 6 Композиции алгоритмов

Простейший пример: уменьшение дисперсии при усреднении алгоритмов методом бутстреп. Блендинг алгоритмов.Понятие смещения и разброса (иллюстрация на примере линейных методов и решающих деревьев). Уменьшение разброса с помощью усреднения. Случайный лес. Оценка out-of-bag.

Раздел 7 Особенности реальных данных

Неполнота и противоречивость. Шумы и выбросы в данных. Методы поиска выбросов. Пропуски в данных, методы их восстановления. Несбалансированные выборки: проблемы и методы борьбы. Задача отбора признаков, примеры подходов.

Раздел 8 Анализ частых множеств признаков и ассоциативных правил

Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритм Априори. Меры “интересности правил”.

Раздел 9. Кластеризация данных

Простые эвристические подходы. Алгоритм K-Means. Проблема устойчивости результатов и важность грамотной инициализации, алгоритм K-Means++. Выбор числа кластеров. Оценка качества кластеризации.

Формы текущего контроля и промежуточной аттестации:

Форма промежуточной аттестации – зачет с оценкой.
В результате освоения дисциплины обучающийся знает, умеет, владеет:

Код этапа освоения компетенции	**Результаты обучения**
ПК-4.2	На уровне знаний: знать: - содержание документационного обеспечения управления (делопроизводства);

ПК-9.1	- содержание и порядок оформления основных организационно-распорядительных документов; - основы организации документооборота на предприятии, в организации, учреждении; - методы и средства автоматизация делопроизводства; классификацию и виды систем управления электронным документооборотом; - состояние и перспективы развития систем управления электронным документооборотом.
	На уровне умений: уметь - отрабатывать основные организационно-распорядительные документы с использованием средств автоматизации; - использовать инструментальные средства компьютерных технологий для эффективной организации и ведения делопроизводства; - разрабатывать отдельные прототипы средств автоматизации делопроизводства
	На уровне навыков: - навыками разработки основных организационно-распорядительных документов с использованием средств автоматизации; - навыками разработки отдельных прототипов средств автоматизации делопроизводства. - владеть навыками разработки основных организационно-распорядительных документов с использованием средств автоматизации; - навыками разработки отдельных прототипов средств автоматизации делопроизводства.
	На уровне знаний знать: - содержание документационного обеспечения управления (делопроизводства); - содержание и порядок оформления основных организационно-распорядительных документов; - основы организации документооборота на предприятии, в организации, учреждении; - методы и средства автоматизация делопроизводства; - классификацию и виды систем управления электронным документооборотом; - состояние и перспективы развития систем управления электронным документооборотом. На уровне умений: уметь - отрабатывать основные организационно-распорядительные документы с использованием средств автоматизации; - использовать инструментальные средства компьютерных технологий для эффективной организации и ведения делопроизводства; разрабатывать отдельные прототипы средств автоматизации делопроизводства
	На уровне навыков: - владеть навыками разработки основных организационно-распорядительных документов с использованием средств автоматизации; - навыками разработки отдельных прототипов средств автоматизации делопроизводства.

Информационные технологии, программное обеспечение, материально-техническая база, оценочные средства, необходимые для освоения дисциплины, адаптированы для обучающихся инвалидов и обучающихся с ограниченными возможностями здоровья.

Основная литература:

1. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014(http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)
2. Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010 (http://www.hse.ru/data/2010/10/14/1223126254/Mirkin_All.pdf)
3. Белов В.С. Информационно-аналитические системы. Основы проектирования и применения. Учебное пособие. - М: МЭСИ, 2004.

4. А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. Учебное пособие. - СПб.: Изд-во : БХВ-Петербург, 2007.