Аннотация

Основной задачей курса является знакомство слушателей с предметной областью. Необходимо дать основные определения и познакомить с терминологией, обсудить прикладные задачи. Рассматриваются основные задачи Data Mining, а также популярные алгоритмы на основе методов машинного обучения для их решения. Для формирования комплексной картины курс нацелен на начальный уровень слушателей, старается использовать больше алгоритмов и дает механизм понимания как их настраивать и использовать. Наконец, важной задачей курса является наработка практического опыты работы с промышленными системами Data mining. В частности, рассматривается язык программирования Python, а также популярные пакеты расширений языка, используемых в машинном обучении. В курсе рассматриваются современные алгоритмы и методы интеллектуального анализа данных для решения задач поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования. В первой части курса, посвященной изучению методов обучения без учителя, рассматриваются: задача поиска ассоциативных правил и основные применяемые для этого алгоритмы: apriori и fp-tree; задача выявления скрытых структур в данных на основе тематического моделирования, в частности метод главных компонент, кластеризация переменных, самоорганизующиеся отображения, неотрицательная матричная факторизация; задача кластеризации данных на основе иерархических, метрических и вероятностных методов. Также обсуждаются методы предобработки данных для эффективного решения данных задач. Вторая часть курса посвящена изучению методов прогнозирования, используемых в системах интеллектуального анализа данных, связанные с этим проблемы, алгоритмы и терминология. Рассматриваются следующие вопросы: понятие проклятия размерности и проблема переобучения; вопросы и критерии для оценки и выбора моделей с использованием валидации и кросс-валидации; алгоритмы и методы необходимой предобработки данных для решения задачи прогнозирования. Далее рассматриваются наиболее популярные и современные алгоритмы и модели машинного обучения и прикладной статистики для решения задач прогнозирования в системах интеллектуального анализа данных, в частности: линейные регрессионные модели; пошаговые методы отбора переменных, регуляризация, преобразование пространства признаков для решения задач прогнозирования; нелинейные регрессионные модели, сплайны, локальная взвешенная регрессия; нейронные сети, их типовые архитектуры RBF и MLP, алгоритмы ранней остановки обучения, методы оптимизации для обучения нейронных сетей; метод опорных векторов для бинарной классификации, виды ядерных функций, алгоритмы оптимизации для обучения модели на основе опорных векторов; деревья решений, алгоритмы и критерии поиска разбиения при их построении, вопросы управления процессом роста и обрубания ветвей деревьев для борьбы с переобучением; ансамбли моделей на основе бустинга и бэгинга, случайный лес и градиентный бустинг.