Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»

УТВЕРЖДЕНО
Директор физтех-школы
прикладной математики
и информатики

А.М. Райгородский

Рабочая программа дисциплины (модуля)

Дисциплина:	Введение в прикладной анализ данных
Направление:	Прикладные математика и физика
Профиль подготовки:	Математическая физика, компьютерные технологии и математическое моделирование в экономике Физтех-школа прикладной математики и информатики Кафедра проблем передачи информации и анализа данных
Курс:	3
Квалификация:	Бакалавр

Семестр, формы промежуточной аттестации: 6 (Весенний) – Дифференцированный зачёт

Аудиторных часов: 30 всего, в том числе:
лекции: 0 час.
практические (семинарские) занятия: 0 час.
лабораторные занятия: 30 час.

Самостоятельная работа: 15 час. всего, в том числе:
задания, курсовые работы: 0 час.

Подготовка к экзамену: 0 час.

Всего часов: 45, всего зач.ед.: 1

Программу составили: **М.Г. Беляев, кандидат физико-математических наук**
М.Е. Панов, кандидат физико-математических наук
А.И. Курмуков, кандидат компьютерных наук

Аннотация

В рамках этого вводного курса проводится знакомство с основными концепциями машинного обучения, включая методы разведочного анализа данных, работы с признаками и классические алгоритмы обучения с учителем и для табличных данных. Курс носит прикладной характер, все темы иллюстрируются с помощью интерактивных python примеров; домашние задания и финальных проект выполняются в аналогичном формате. Основная цель курса -

дать систематический обзор основных методов машинного обучения для табличных данных и навыки использования основных программных библиотек, реализующих эти методы.

Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 6 (Весенний)

1. Обзор основных прикладные задач анализа данных. Примеры задач из повседневной жизни.

2. Прикладные пакеты для решения задач анализа данных.

Основные понятия языка Python, структуры данных, конструкции языка. Библиотека матричных вычислений numpy. Работа в интерактивной среде ipython-notebook.
Предварительный визуальный анализ параметров задачи, эвристическая проверка значимости параметров. Библиотека визуализации seaborn.
Исследование задачи предсказания выживаемости пассажиров Титаника по формальным характеристикам (пол, класс каюты, ...).
Решение задач анализа данных с помощью языка Python. Библиотеки scikit-learn, pandas, scipy, statmodels.
Задача разбиения текстов новостей на группы.

3. Задача классификации.

Постановка задачи классификации, обзор основных методов ее решения. Метрики качества классификации (точность/специфичность, ROC-кривая, площадь под кривой).
Логические алгоритмы. Решающие деревья, решающие списки. Понятие информативности, методы поиска информативных закономерностей.
Агрегирование моделей. Ансамбли решающих деревьев. Градиентный бустинг.
Задача классификации тау-тау распада бозона Хиггса.

4. Задачи обучения без учителя.

Снижение размерности. Метод главных компонент. Обзор основных идей нелинейных методов снижения размерности.
Задача генерация профилей крыла самолета по заданной выборке данных, ее решение методами снижения размерности.
Кластеризация данных. Основные подходы и методы кластеризации, кластеризация на основе зависимостей.
Использование методов кластеризации в задаче распознавания цифр.

5. Задача регрессии.

Постановка задачи регрессии, основные линейные и нелинейные методы ее решения.
Задача моделирования распределения давления по профилю крыла самолета.

6. Подготовка к решению прикладных задач.

Методы генерации признаков в различных задачах анализа данных (текста, аудио).
Методология решения прикладных задач и написания отчетов.

Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

Основная литература
1. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning (second edition) // New York: Springer, 2009.
2. Leskovec J., Rajaraman A., Ullman J.D. Mining of massive datasets // Cambridge University Press, 2014.
3. Bishop, Christopher M. Pattern recognition and machine learning // New York: Springer, 2006.
4. Steele J., Iliinsky N. Beautiful Visualization: Looking at Data through the Eyes of Experts // "O'Reilly Media, Inc.", 2010.
5. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython // O'Reilly Media, 2012.
6. Айвазян С.А., Бухштабер В.М., Енюков С.А., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности // М.: Финансы и статистика, 1989.
7. Ту Дж., Гонсалес Р. Принципы распознавания образов // М.: Мир, 1978.

Дополнительная литература
1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных // М.: Финансы и статистика, 1983.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей // М.: Финансы и статистика, 1985.
3. Дьяконов А.Г. Чему не учат в анализе данных и машинном обучении // Учебное пособие. http://alexanderdyakonov.narod.ru/lpot4emu.pdf
4. Дронов С.В. Многомерный статистический анализ // Учебное пособие. Барнаул: Изд-во Алт. гос. ун-та, 2003.