

**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)
по дисциплине:	Введение в прикладной анализ данных
по направлению:	Инфокоммуникационные технологии и системы связи
профиль подготовки:	Телекоммуникационные сети и системы
	Физтех-школа Радиотехники и Компьютерных Технологий
	кафедра проблем передачи информации и анализа данных
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 0 час.

семинары: 0 час.

лабораторные занятия: 30 час.

Самостоятельная работа: 60 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: М.Г. Беляев, ассистент

Программа обсуждена на заседании кафедры проблем передачи информации и анализа данных 01.02.2024

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.

1. Цели и задачи

Цель дисциплины

Дать студентам обзор современных задач анализа данных и обучить методам и навыкам решения таких задач.

Задачи дисциплины

- изучение постановок стандартных задач анализа данных;
- знакомство с библиотеками анализа данных для языка python;
- изучение методов решения задач анализа данных;
- самостоятельное решение прикладных задач методами анализа данных.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
ОПК-1 Способен представлять современную научную картину мира, выявлять естественнонаучную сущность проблем своей профессиональной деятельности, определять пути их решения и оценивать эффективность сделанного выбора	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные и прикладные научные знания в области естественных наук
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
ОПК-2 Способен реализовывать новые принципы и методы исследования современных инфокоммуникационных систем и сетей различных типов передачи, распределения, обработки и хранения информации	ОПК-2.1 Знает принципы и методы исследования современных инфокоммуникационных систем и умеет оценивать их достоинства и недостатки
	ОПК-2.2 Владеет навыками реализации новых принципов и методов исследования в современных инфокоммуникационных системах и сетях
ОПК-3 Способен приобретать, обрабатывать и использовать новую информацию в своей предметной области, предлагать новые идеи и подходы к решению задач своей профессиональной деятельности	ОПК-3.1 Умеет использовать современные информационные и компьютерные технологии, средства коммуникаций при поиске научно-технической информации в своей профессиональной деятельности
	ОПК-3.2 Способен системно анализировать полученную информацию, использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки и техники
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
	ПК-2.2 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и (или) участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны знать:

- основные постановки задач анализа данных;
- основные методы решения задач анализа данных.

уметь:

- пользоваться стандартными библиотеками анализа данных;
- решать прикладные задачи анализа данных из различных областей.

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Обзор основных прикладные задач анализа данных			3	6
2	Прикладные пакеты для решения задач анализа данных			8	16
3	Задача классификации			8	16
4	Задачи обучения без учителя			5	10
5	Задача регрессии			3	6
6	Подготовка к решению прикладных задач			3	6
Итого часов				30	60
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 2 (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.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная мультимедийным оборудованием (проектор или плазменная панель), доска, компьютерный класс.

6.Перечень рекомендуемой литературы

Основная литература

Фонд литературы базовой кафедры (организации):

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. Айвазян С.А., Бухштабер В.М., Енюков С.А., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности // М.: Финансы и статистика, 1989.

Дополнительная литература

Фонд литературы базовой кафедры (организации):

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. Bishop, Christopher M. Pattern recognition and machine learning // New York: Springer, 2006.
4. Steele J., Piinsky 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.

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

1. <http://lib.mipt.ru> – электронная библиотека Физтеха.
2. <http://www.edu.ru> – федеральный портал «Российское образование».
3. <http://benran.ru> – библиотека по естественным наукам Российской академии наук.
4. <http://www.statsoft.ru/home/textbook/default.htm>.

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лабораторных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен, с одной стороны, овладеть общими понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике. В результате изучения дисциплины студент должен знать основные определения, понятия, методы решения задач.

Успешное освоение курса требует напряженной самостоятельной работы студента. В программе курса отведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы;
- проработку учебного материала (по конспектам занятий, учебной и научной литературе), подготовку ответов на вопросы, предназначенные для самостоятельного изучения, решение задач;
- подготовка к дифференцированному зачёту.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов следует обращаться за консультациями к преподавателю.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Инфокоммуникационные технологии и системы связи
профиль подготовки: Телекоммуникационные сети и системы
Физтех-школа Радиотехники и Компьютерных Технологий
кафедра проблем передачи информации и анализа данных
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Разработчик: М.Г. Беляев, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий	УК-1.2 Осуществляет поиск вариантов решения поставленной проблемной ситуации на основе доступных источников информации
ОПК-1 Способен представлять современную научную картину мира, выявлять естественнонаучную сущность проблем своей профессиональной деятельности, определять пути их решения и оценивать эффективность сделанного выбора	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные и прикладные научные знания в области естественных наук
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
ОПК-2 Способен реализовывать новые принципы и методы исследования современных инфокоммуникационных систем и сетей различных типов передачи, распределения, обработки и хранения информации	ОПК-2.1 Знает принципы и методы исследования современных инфокоммуникационных систем и умеет оценивать их достоинства и недостатки
	ОПК-2.2 Владеет навыками реализации новых принципов и методов исследования в современных инфокоммуникационных системах и сетях
ОПК-3 Способен приобретать, обрабатывать и использовать новую информацию в своей предметной области, предлагать новые идеи и подходы к решению задач своей профессиональной деятельности	ОПК-3.1 Умеет использовать современные информационные и компьютерные технологии, средства коммуникаций при поиске научно-технической информации в своей профессиональной деятельности
	ОПК-3.2 Способен системно анализировать полученную информацию, использовать исследовательские методы при решении новых задач, применяя знания в различных областях науки и техники
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Способен планировать и проводить научные исследования самостоятельно или в составе научного коллектива
	ПК-2.2 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и (или) участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «Введение в прикладной анализ данных» обучающийся должен:

знать:

- основные постановки задач анализа данных;
- основные методы решения задач анализа данных.

уметь:

- пользоваться стандартными библиотеками анализа данных;
- решать прикладные задачи анализа данных из различных областей.

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлой лабораторной работы.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Задача классификации. Метрики качества классификации.
2. Решающие деревья для регрессии и классификации. Алгоритмы построения и выбора глубины дерева.
3. Градиентный бустинг на деревьях.
4. Алгоритмы агрегации моделей.
5. Метод главных компонент.
6. Нелинейные методы снижения размерности.
7. Методы кластеризации данных.
8. Основные методы решения задачи регрессии.
9. Методология решения практических задач анализа данных.
10. Методы генерации признаков в различных задачах анализа данных (текста, аудио).

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Дифференцированный зачёт проводится в устной форме.

При проведении устного дифференцированного зачёта обучающемуся предоставляется 30 минут на подготовку.

Во время проведения дифференцированного зачёта обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой и проч.