

**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:	Дополнительные задачи анализа данных
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Системное программирование и прикладная математика Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: И.Г. Эрлих, канд. физ.-мат. наук, доцент

Программа обсуждена на заседании кафедры дискретной математики 05.06.2023

Аннотация

Курс знакомит слушателей с задачами анализа данных в дополнении к курсам прикладной статистики и машинного обучения. В рамках курса студенты научатся правильно обрабатывать сырые данные и визуализировать их. Также слушатели познакомятся с анализом временных рядов и способами выявления аномалий в данных. По курсу предполагаются теоретические задачи и работа с реальными данными с помощью языков Питон и R, в том числе соревнования на платформе Kaggle.

1. Цели и задачи

Цель дисциплины

Познакомить слушателей с задачами анализа данных в дополнении к курсам прикладной статистики и машинного обучения. Научить студентов правильно обрабатывать сырые данные и визуализировать их, в том числе с помощью современных методов понижения размерности пространства.

Задачи дисциплины

- Познакомить студентов с некоторыми применениями интеллектуального анализа данных и машинного обучения для решения бизнес-задач;
- научить студентов видеть проблемы, которые могут быть решены с помощью машинного обучения;
- научить студентов осуществлять постановку задач интеллектуального анализа данных.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения

научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- Постановки задач интеллектуального анализа данных;
- популярные алгоритмы интеллектуального анализа данных;
- современный технический уровень в развитии алгоритмов интеллектуального анализа данных.

уметь:

- Находить в описании задач из бизнеса задачи для интеллектуального анализа данных;
- осуществлять математическую постановку задач интеллектуального анализа данных.

владеть:

- Современными алгоритмами интеллектуального анализа данных;
- современным инструментарием для промышленного решения задач интеллектуального анализа данных.

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	Прогнозирование временных рядов	6	6		14

2	Экспоненциальное сглаживание, адаптивное Экспоненциальное сглаживание	6	6		14
3	Способы нелинейного прогнозирования временных рядов	6	6		15
4	Аномалии: выбросы и новизна	6	6		16
5	Постановка задачи последовательного анализа, сравнение с обычной процедурой проверки гипотез	6	6		16
Итого часов		30	30		75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 6 (Весенний)

1. Прогнозирование временных рядов

Автокорреляционная функция, кореллограмма, критерий Льюнга-Бокса. STL-декомпозиция временного ряда на тренд, сезонность и остатки. Стационарные временные ряды, критерий KPSS, преобразование Бокса-Кокса, дифференцирование ряда. Анализ остатков.

2. Экспоненциальное сглаживание, адаптивное Экспоненциальное сглаживание

Модель скользящего среднего MA и модель авторегрессии AR. Представление модели AR в виде модели MA(inf), стационарность в модели AR. Модели ARMA, ARIMA, оценка параметров модели. Подбор оптимальных гиперпараметров модели на основе автокорреляционной и частичной автокорреляционной функции. Учет сезонности и экзогенных факторов: модель SARIMAX.

3. Способы нелинейного прогнозирования временных рядов

Способы оценки качества, кросс-валидация для временных рядов.

4. Аномалии: выбросы и новизна

Детектирование аномалий: типы задач, подходы. Ящик с усами, критерий Граббса, эллиптическая оболочка (Elliptic Envelope), метод главных компонент, локальный уровень выброса (Local Outlier Factor), кластеризация с помощью DBSCAN, изолирующий лес (Isolation Forest), Robust Random Cut Forest, One Class SVM. Детектирование аномалий на основе нейронных сетей (автоэнкодеры), особенности построения для детекции аномалий. Аномалии во временных рядах, онлайн и оффлайн методы. Фильтрация, медианный фильтр. Метрические методы. Seasonal EDS и Seasonal Hybrid EDS. Адаптация Robust Random Cut Forest для работы в онлайн.

5. Постановка задачи последовательного анализа, сравнение с обычной процедурой проверки гипотез

Последовательный критерий отношения правдоподобия, примеры. Задача скорейшего обнаружения разладки, примеры применения. Статистики CUSUM, Ширяева-Робертса.

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

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6.Перечень рекомендуемой литературы

Основная литература

1. Экстремальные задачи теории графов и анализ данных [Текст] / А. М. Райгородский - М. Регулярная и хаотическая динамика, 2008
2. Геронтология in Silico: становление новой дисциплины. Математические модели, анализ данных и вычислительные эксперименты [Электронный ресурс] : сборник науч. тр. / под ред. Г. И. Марчука, В. Н. Анисимова, А. А. Романюхи, А. И. Яшина .— 3-е изд. (эл.) .— Электрон. текстовые дан. (1 файл pdf : 538 с.) .— М. : БИНОМ. Лаборатория знаний, 2015 .— Систем. требования : Adobe Reader XI ; экран 10" .— Электрон. версия печ. публикации .— Полный текст (Режим доступа : доступ из сети МФТИ).

Дополнительная литература

1. Наглядная математическая статистика [Текст] : учеб. пособие для вузов / М. Б. Лагутин .— 2-е изд., испр. — М. : Бином. Лаб. знаний, 2009 .— 472 с.

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

Не используются

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

На семинарских занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

Для контроля и коррекции знаний, обучающиеся могут использовать компьютерное тестирование.

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

Для успешного освоения данной дисциплины студенту необходимо:

- посещать семинары, при этом конспектирование материалов не является необходимым, поскольку основные материалы хранятся в кафедральной папке в облачном хранилище данных «Яндекс. Диск», к которому предоставлен доступ всем студентам кафедры;
- выполнять задания, задаваемые преподавателем на семинарах;
- выполнить итоговое письменное задание по дисциплине, которое вносит основной вклад в изучение дисциплины, а также в итоговую оценку по данному курсу.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Информатика и вычислительная техника
профиль подготовки:	Системное программирование и прикладная математика Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 6 (весенний) - Дифференцированный зачет

Разработчик: И.Г. Эрлих, канд. физ.-мат. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации

2. Показатели оценивания компетенций

В результате изучения дисциплины «Дополнительные задачи анализа данных» обучающийся должен:

знать:

- Постановки задач интеллектуального анализа данных;
- популярные алгоритмы интеллектуального анализа данных;
- современный технический уровень в развитии алгоритмов интеллектуального анализа данных.

уметь:

- Находить в описании задач из бизнеса задачи для интеллектуального анализа данных;
- осуществлять математическую постановку задач интеллектуального анализа данных.

владеть:

- Современными алгоритмами интеллектуального анализа данных;
- современным инструментарием для промышленного решения задач интеллектуального анализа данных.

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

Примеры задач:

1. Пусть временной ряд $\{y_t, t \in \mathbb{Z}\}$ с нулевым средним подчиняется модели авторегрессии $AR(2)$: $y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t$, где белый шум $\{\varepsilon_t\}$ не зависит от y_{t-1} , $i > 1$. Докажите, что условия $|\varphi_2| < 1$, $\varphi_1 + \varphi_2 < 1$, $\varphi_2 - \varphi_1 < 1$ являются необходимыми для того, чтобы ряд y_t являлся стационарным в широком смысле.
2. В файле содержится информация о максимальном спросе на электричество (Consumption) в штате Виктория (Австралия) за 30-минутные интервалы с 10 января 2000 в течении 115 дней, а также информация о температуре воздуха (Temperature) за эти же промежутки времени.
 - 1) Нарисуйте графики временных рядов температуры и потребления электричества. Верно ли, что спрос на электричество зависит от температуры воздуха? Для ответа на вопрос используйте коэффициенты корреляции, учитывая условия их применимости.
 - 2) Сколько типов сезонностей можно выделить в каждом из двух рядов (спрос на электричество и температура)?
 - 3) С помощью STL-декомпозиции в каждом ряде выделите тренд, все типы сезонности, остатки.
 - 4) С помощью критерия KPSS проверьте на стационарность исходные ряды и остатки, полученные после применения STL-декомпозиции.
 - 5) С помощью преобразований исходных рядов приведите их к стационарным.
 - 6) По графикам ACF и PACF подберите параметры модели $SARIMA(p, d, q) \times (P, D, Q)_s$.
 - 7) С помощью поиска по сетке вокруг выбранных параметров подберите оптимальные параметры по значению AIC.
 - 8) Учтите, что из сделанных ранее преобразований ряда нужно оставить лишь некоторые.
 - 9) Другие, например, одна из сезонностей будут учтены параметрами модели.
 - 10) Постройте прогнозы модели с оптимальными параметрами на неделю вперед.
 - 11) Посчитайте качество прогноза по сравнению с реальными данными на тестовом интервале, используя метрику MSE.
3. Вам выдаются 4 файла со статистикой посещаемости веб-страниц. Каждый файл соответствует одному месяцу с февраля по май. Данные сгруппированы по часам. Записывается только точка входа на сайт. Иначе говоря, сессия, в рамках которой пользователь посетил несколько страниц, учитывается только один раз. Задача --- найти аномалии в предоставленных временных рядах. Рассмотрите как отдельные ряды, так и их сумму.

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

1. Какие основные методы обработки данных?
2. В чем особенность несбалансированных классов?
3. Почему нельзя для прогнозирования временных рядов использовать простую линейную или нелинейную регрессию?
4. В чем заключается сезонное дифференцирование временного ряда?
5. Как автокорреляционная и частичная автокорреляционная функции помогают подобрать модель для прогнозирования временного ряда?
6. Какие существуют типы аномалий?
7. В чем принципиальное отличие аномалий?
8. В чем связь между случайным и изолирующими лесами?
9. В чем отличие между случайным и изолирующими лесами?
10. Является ли аномалией разладка во временном ряду?

Критерии оценивания

- оценка «отлично (10)» 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 (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» 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 (6)» выставляется студенту, если он 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)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

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

При проведении дифференцированного зачета, обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося на зачете не должен превышать двух астрономических часов. Во время проведения дифференцированного зачета, обучающиеся могут пользоваться программой дисциплины, а также справочной литературой и другими материалами.