

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

**УТВЕРЖДЕНО**

**Директор физтех-школы  
электроники, фотоники и  
молекулярной физики**

**В.В. Иванов**

**Рабочая программа дисциплины (модуля)**

<b>по дисциплине:</b>	Физико-химические методы исследования объектов как источников больших баз данных
<b>по направлению:</b>	Прикладные математика и физика
<b>профиль подготовки:</b>	Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы Физтех-школа Электроники, Фотоники и Молекулярной Физики кафедра химической физики
<b>курс:</b>	1
<b>квалификация:</b>	магистр

Семестры, формы промежуточной аттестации:

1 (осенний) - Зачет

2 (весенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 60 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: С.О. Травин, канд. хим. наук

Программа обсуждена на заседании кафедры химической физики 27.05.2021

## Аннотация

Курс "Физико-химические методы исследования объектов как источников больших баз данных" предусматривает ознакомление обучающихся с базовыми проблемами работы с большими данными и методами их решения; применение алгоритмов обработки больших данных к физико-химическим исследованиям, использование справочных баз данных; возможности применения современных физико-химических и математических методов в культурологии.

### 1. Цели и задачи

#### Цель дисциплины

ознакомление обучающихся с базовыми проблемами работы с большими данными и методами их решения; применение алгоритмов обработки больших данных к физико-химическим исследованиям, использование справочных баз данных; возможности применения современных физико-химических и математических методов в культурологии.

#### Задачи дисциплины

Формирование представлений о работе с реляционными базами данных, извлечении очистке и трансформации данных, полученных из различных источников, превращение информации в знание и создание моделей исследуемых процессов. Знакомство с современными алгоритмами машинной обработки 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-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
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

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны знать:

- теоретические основы алгоритмов работы с большими данными, сильные и слабые стороны статистических методов;
- основные приемы извлечения, трансформации, очистки и хранения информации в базах данных;
- язык запросов SQL и инструментарий научно-ориентированных библиотек языка Python;
- типовые методы кластеризации, классификации и обнаружения скрытых тенденций, использование этих методов для прогнозирования;
- специфику различных физико-химических методов и баз данных, применяемых при их использовании;
- применения машинного обучения и нейросетей к построению моделей изучаемых явлений и процессов.

уметь:

- планировать стратегию исследования состава вещества и идентификации его компонент;
- обрабатывать экспериментальные данные, полученные с помощью физико-химических методов исследования вещества с использованием основных методологических принципов;
- использовать современные методики сбора, очистки и обработки данных;
- готовить наглядные презентации полученных результатов.

владеть:

- практическими навыками использования языка запросов SQL и написания простейших кодов на Python;
- навыками поиска в химических базах данных;
- типовыми приемами обработки и анализа результатов физико-химического и вычислительного эксперимента;
- методологией сопоставления и критической интерпретации массива данных, полученных всей совокупностью использованных физико-химических и математических методов исследования строения и состава вещества.

#### 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	Большие данные и революция в сознании	2			1
2	Большие данные и культурное наследие	2			1
3	Физико-химические методы анализа как источник больших данных	2			1
4	Об основах Data Science – добыча и очистка данных	2			1
5	Понятие о базах данных	2			1
6	Знакомство с хемиоинформатикой	2			1
7	Молекулярные дескрипторы - продолжение	2			1
8	Вероятность против детерминизма	2			1
9	Введение в язык SQL	4			2
10	Знакомство с библиотеками языка Python	4			2
11	Знакомство с регрессией	2			1
12	Нелинейная регрессия	2			1
13	Метод главных компонент	2			1
14	Системы поддержки принятия решений, коллаборативная фильтрация	4			4
15	Кластеризация методом k-средних	4			4
16	Метод k-ближайших соседей и обнаружение аномалий	4			4
17	Ассоциативные правила	4			4
18	Анализ социальных сетей	2			2
19	Дерево решений	2			2
20	Случайные леса	2			2

21	А/В-тестирование и многорукие бандиты	2			2
22	Нейронные сети	6			6
Итого часов		60			45
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

##### Семестр: 1 (Осенний)

##### 1. Большие данные и революция в сознании

Наука о данных - вводная информация. Что такое большие данные. Поиск скрытых закономерностей – превращение информации в знание. Визуализация и презентация данных, наглядное представление массивов различной информации. Типы визуализации, преобразованы в форму, усиливающую восприятие и анализ информации. Работа с неструктурированной информацией. Организация и хранение больших данных.

##### 2. Большие данные и культурное наследие

Большие данные, применительно к объектам культурного наследия могут быть использованы, главным образом, в следующих трех направлениях:

- 1) Цифровизация памятников культуры (картин, фресок, скульптур, архитектурных сооружений). Эти данные впоследствии окажутся исключительно полезны при реставрации объектов, подборе материалов, их идентификации, установлении совместимости.
- 2) Предотвращение преступлений в сфере культурного наследия. Выявление подделок, учет авторских и исторически значимых копий, препятствие вводу в оборот через аукционы и музеи фальшивых произведений, пресечение незаконного оборота похищенных произведений искусства.
- 3) Каталогизация и систематизация объектов культурного наследия. Классификация объектов, популяризация знаний, развитие туризма, повышение посещаемости музеев.

##### 3. Физико-химические методы анализа как источник больших данных

Обзор вопросов сбора и использования больших данных в современной аналитической химии. Такие данные характеризуются значительными объемами, потоками и разнообразием. Их генерация и манипуляции с ними сопровождают анализ биопроб, образцов другого происхождения, в первую очередь, методом хроматографии и масс-спектрометрии. Большие данные, полученные с помощью этих методов, обеспечивают многоаналитический анализ образцов, хотя характеристики обнаружения, идентификации и количественной оценки удовлетворительны не для всех анализируемых веществ. Применение простых аналитических систем также может сопровождаться накоплением больших объемов данных. Огромный объем информации содержится в больших химических базах данных, использование которых необходимо

при нецелевом анализе. При отборе кандидатов для идентификации учитывается распространенность (частота цитирования) химических веществ; идентификация включает использование эталонных масс-спектральных библиотек. Методы обработки, анализа и представления данных (статистика, хемометрия) эволюционируют с ростом объема информации. Технические характеристики компьютеров и их сетей улучшаются опережающими темпами, создавая потенциал для развития методов анализа и открытия новых возможностей.

##### 4. Об основах Data Science – добыча и очистка данных

Подготовка данных Знакомство с Data Mining. Очистка данных. Дублирование и пропуски. Процедуры ETL. Проблемы с выбором системы физических единиц. Подводные камни, о которых знают все, но попадают многие. Формат данных, типы переменных, выбор переменных, конструирование признаков, неполные данные, выбор алгоритма, обучение без учителя, обучение с учителем, обучение с подкреплением, другие факторы, настройка параметров, оценка результатов, метрики классификации, метрика регрессии, валидация, краткие итоги

## 5. Понятие о базах данных

Терминология, используемая в теории БД на стадии проектирования и практической работы. Сведения о БД как важнейшем компоненте информационных систем. Реляционные базы данных, таблицы, атрибуты и отношения. Ключевые элементы таблицы, первичные и вторичные ключи. Связи между таблицами. Типы связи "один ко многим" (1:N) и "один к одному" (1:1). Хранимые процедуры и триггеры. Основные платформы, поддерживающие крупные хранилища данных (SAP, Oracle, Microsoft, Galaktika, 1С и т.д.) Классификация баз данных: централизованные и распределенные, с локальным доступом и базы данных с сетевым доступом. Понятие транзакции. Для всех современных баз данных можно организовать сетевой доступ с многопользовательским режимом работы

## 6. Знакомство с хемиоинформатикой

Химическое пространство описывает все возможные молекулы, а также многомерные концептуальные пространства, представляющие структурное разнообразие этих молекул. Часть этого химического пространства доступна в общедоступных базах данных от тысяч до миллиардов соединений. Использование этих баз данных для поиска лекарств представляет собой типичную проблему больших данных, ограниченную вычислительной мощностью, возможностями хранения данных и доступа к ним. Здесь мы рассмотрим последние разработки нашей лаборатории, включая прогресс в универсальных базах химических данных (GDB) и фрагмент подмножество БПД-17, инструментов для лиганда на основе виртуального скрининга на ближайшей поиски соседа, такие, как наш мульти-отпечатков пальцев браузер на базе ZINC для выбора соединений приобретаемых скрининга, и их применения, чтобы обнаружить мощные и селективные ингибиторы кальциевого канала TRPV6 и Auroga в киназы, в polypharmacology браузер (ППБ) для прогнозирования мимо эффектов, и, наконец, интерактивной 3D-визуализации химического пространства, используя наши онлайн инструменты WebDrugCS и WebMolCS.

## 7. Молекулярные дескрипторы - продолжение

Внутреннее и внешнее представление химической информации.

Концепция молекулярных дескрипторов. Свободно доступные компьютерные программы расчета дескрипторов. Создание и работа с химическими базами данных. Свободно доступные базы данных: PubChem, PDB, ZINC, NCI, DrugBank, BindingDB, ChemSpider, Kegg и др.

Навигация в химическом пространстве данных: базовые понятия, область применения. Методы навигации в химическом пространстве данных. Дизайн "сфокусированных библиотек". Виртуальный скрининг. Классификационные методы машинного обучения, используемые в химической информатике. Оценка качества классификационных моделей. Расчет молекулярных дескрипторов и разработка классификационных моделей. Регрессионные методы машинного обучения, используемые в химической информатике

## 8. Вероятность против детерминизма

Вероятность и достоверность. Парадоксы теории вероятности. Черные лебеди и не нормальные (негауссовы) распределения с тяжелыми хвостами. Примеры распределений «ранг - размер» типа Ципфа Мандельброта. История анализа текстов, применение частотного анализа. Примеры из экономики, географии, военных действий. Оценки, проводимые с помощью ранговых распределений. Моделирование – главный путь превращения информации в знания. Иерархия моделей. В каких случаях уместен статистический анализ. Отличие науки от бизнеса. Парадокс Бертрона. Парадокс Симпсона. Отличие «ошибки» эксперимента от «неверной интерпретации» эксперимента

## 9. Введение в язык SQL

Язык SQL появился в 1974 г. при выполнении исследовательского проекта System R компании IBM. Тогда и долгое время спустя считалось, что язык SQL является практической реализацией идей реляционной модели данных, а многие пользователи языка и сегодня придерживаются того же мнения.

Поддерживаемые в SQL типы данных и преобразование типов. Создание пользовательских типов данных. Понятие выражения и оператора в SQL. Понятие домена - набор допустимых значений для одного или нескольких атрибутов. Если в таблице базы данных или в нескольких таблицах присутствуют столбцы, обладающие одними и теми же характеристиками, можно описать тип такого столбца и его поведение через домен, а затем поставить в соответствие каждому из одинаковых столбцов имя домена. Домен определяет все потенциальные значения, которые могут быть присвоены атрибуту.

Создание баз данных. Запросы к базе данных. выборка данных, СУБД, операции, результат выполнения запроса, выражение, параметр, фильтрация строк, группировка строк, список, критерий отбора, предикат, значение, дублирующие записи, операции реляционной алгебры, запись, WHERE, пользователь, FROM, условия поиска, сравнение, диапазон, принадлежность множеству, соответствие шаблону, значение NUL.

Приоритет операций. объединение отношений, пересечение, разность, декартово произведение, выборка, проекция, соединение, теоретико-множественные операции, операция проекции, операции реляционной алгебры, соединение по эквивалентности, внешнее соединение, полусоединение, кортеж отношения, деление

## 10. Знакомство с библиотеками языка Python

Библиотека Pandas

<https://habr.com/ru/company/ruvds/blog/494720/>

Знакомство с библиотекой языка Python. Scikit

Знакомство с библиотекой языка Python. Numpy

## 11. Знакомство с регрессией

Метод наименьших квадратов. Простая линейная регрессия. Вычисление коэффициентов регрессии и ошибок их определения. Линейная регрессия со scikit-learn. Пример решения задачи множественной регрессии с помощью Python. Полное руководство по линейной регрессии в Scikit-Learn.

## 12. Нелинейная регрессия

Сводимость нелинейной регрессии к линейной. Метод Ньютона –Рафсона, итерационное приближение. Метод Левенберга-Марквардта. Оценка ошибок определения численных параметров нелинейной регрессии.

## 13. Метод главных компонент

Подготовка данных. Сглаживание спектров ЭПР. Ковариационная матрица. Нормальная форма Фробениуса. Алгоритм Данилевского. Примеры распознавания спектров. Пример изучения пищевой ценности, главные компоненты, пример: анализ пищевых групп, ограничения, обзор темы.

#### 14. Системы поддержки принятия решений, коллаборативная фильтрация

Проблема рекомендации товаров или услуг. Прогнозирование неизвестных предпочтений. Алгоритмы «от клиента» User Based, UB CF и «от продукта» Item Based, IB CF. Меры схожести (косинусная мера, коэффициент корреляции пирсона, евклидово расстояние, коэффициент танимото, манхэттенское расстояние). SVD-разложение (Singular Value Decomposition), оно же сингулярное разложение. Проблема холодного старта. Белые вороны.

#### 15. Кластеризация методом к-средних

Поиск кластеров разбиение множества элементов векторного пространства на известное число кластеров. Пример: профили кинозрителей, определение кластеров, сколько кластеров существует? Что включают кластеры? Ограничения. Пример кластеризации изображений движущихся объектов.

#### 16. Метод k-ближайших соседей и обнаружение аномалий

Теория метода, интуитивная основа KNN – самый простой из всех контролируемых алгоритмов машинного обучения. Вычисление расстояний от каждой новой точки данных до всех других обучающих точек данных. Пищевая экспертиза, яблоко от яблони недалеко падает, пример поиска истинных различий. Обнаружение аномалий. Ограничения. Реализация алгоритма KNN с помощью Scikit-Learn. Алгоритм KNN плохо работает с категориальными объектами

#### 17. Ассоциативные правила

Составление схемы отношений, пример: геополитика в торговле оружием, Лувенский метод, алгоритм PageRank, ограничения, обзор темы

#### 18. Анализ социальных сетей

Составление схемы отношений, пример: геополитика в торговле оружием, Лувенский метод, алгоритм PageRank, ограничения, обзор темы

#### 19. Дерево решений

Дерево решений – наглядная пошаговая инструкция, что делать в какой ситуации. эффективное разделение выборки путем пошагового уменьшения энтропии. Создание дерева решений. Прогноз выживания в катастрофе. Пример: спасение с тонущего «Титаника». Ограничения.

#### 20. Случайные леса

Случайный лес — является одним из немногих универсальных алгоритмов, дошёл до нашего времени в «первозданном виде» и никакие эвристики не смогли его существенно улучшить. Идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим Масштабирование данных и разделение на обучающую и тестовую выборки. Пример: предсказание криминальной активности. Ансамбли. Бэггинг

#### 21. A/B-тестирование и многорукие бандиты

Основы A/B-тестирования, предназначение метода. Методика проведения тестов. Ограничения A/B-тестирования. Стратегия снижения эпсилона. Анализ результатов, оценка значимости. Пример: многорукие бандиты. Забавный факт: ставка на победителя

## 22. Нейронные сети

Обзор темы

Создание мозга

Пример: распознавание рукописных цифр

Компоненты нейронной сети

Правила активации

Ограничения

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

Учебная аудитория, снабженная доской, экраном, проектором.

## 6. Перечень рекомендуемой литературы

### Основная литература

1. Джон Форман Много цифр Анализ больших данных при помощи Excel, Пер. с англ. А. Соколовой. - М.: Альпина Паблишер, 2016.-461 с.
2. Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман, Анализ больших наборов данных / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 498 с.

### Дополнительная литература

1. Силен Дэви, Мейсман Арно, Али Мохамед, / Основы Data Science и Big Data. Python и наука о данных — СПб.: Питер, 2017. 336 с.: ил. — (Серия «Библиотека программиста»).
2. Дейтел Пол, Дейтел Харви //Python: Искусственный интеллект, большие данные и облачные вычисления. — СПб.: Питер, 2020. — 864 с.: ил. — (Серия «Для профессионалов»).
3. Жан-Батист Мишель, Эрец Эйден, / Незведанная территория. Как «большие данные» помогают раскрывать тайны прошлого и предсказывать будущее нашей культуры, / Серия «Наука XXI век» [http://www.litres.ru/pages/biblio\\_book/?art=14564549](http://www.litres.ru/pages/biblio_book/?art=14564549), /АСТ; Москва; 2016,
4. Виктор Майер-Шенбергер, Кеннет Кукьер Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / пер. с англ. Инны Гайдюк. — М. : Манн, Иванов и Фербер, 2014. — 240 с.
5. Талеб, Нассим Николас (2015), Черный лебедь. Под знаком непредсказуемости, КоЛибри,

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

<https://mipt.ru/science/labs/radiophotonics/obuchenie/lazery.php>

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

Не предусмотрено.

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

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике. В результате изучения дисциплины студент должен знать основные определения дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

- посещения всех занятий, предусмотренных учебным планом по дисциплине;
- ведения конспекта занятий;



– напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

– чтение рекомендованной литературы;

– проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;

– решение задач, предлагаемых студентам на занятиях;

– подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

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

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Прикладные математика и физика
<b>профиль подготовки:</b>	Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы Физтех-школа Электроники, Фотоники и Молекулярной Физики кафедра химической физики
<b>курс:</b>	<u>1</u>
<b>квалификация:</b>	магистр
Семестры, формы промежуточной аттестации:	
1 (осенний) - Зачет	
2 (весенний) - Экзамен	
<b>Разработчик:</b>	С.О. Травин, канд. хим. наук

## 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-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними
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

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Физико-химические методы исследования объектов как источников больших баз данных» обучающийся должен:

### знать:

- теоретические основы алгоритмов работы с большими данными, сильные и слабые стороны статистических методов;
- основные приемы извлечения, трансформации, очистки и хранения информации в базах данных;
- язык запросов SQL и инструментарий научно-ориентированных библиотек языка Python;
- типовые методы кластеризации, классификации и обнаружения скрытых тенденций, использование этих методов для прогнозирования;
- специфику различных физико-химических методов и баз данных, применяемых при их использовании;
- применения машинного обучения и нейросетей к построению моделей изучаемых явлений и процессов.

### уметь:

- планировать стратегию исследования состава вещества и идентификации его компонент;
- обрабатывать экспериментальные данные, полученные с помощью физико-химических методов исследования вещества с использованием основных методологических принципов;
- использовать современные методики сбора, очистки и обработки данных;
- готовить наглядные презентации полученных результатов.

### владеть:

- практическими навыками использования языка запросов SQL и написания простейших кодов на Python;
- навыками поиска в химических базах данных;
- типовыми приемами обработки и анализа результатов физико-химического и вычислительного эксперимента;
- методологией сопоставления и критической интерпретации массива данных, полученных всей совокупностью использованных физико-химических и математических методов исследования строения и состава вещества.

## 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 лекции или в конце занятия по пройденной теме.

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

Вопросы к зачету:

1. Какие четыре ключевых шага предполагает исследование в рамках науки о данных?
2. В чем заключается этап подготовки данных?
3. Как осуществляется выбор алгоритмов для моделирования данных?
4. Как проходит настройка алгоритмов для оптимизации моделей?
5. Как оценивается точность моделей?
6. Что такое кластеризация методом k-средних?
7. Как определяется число кластеров k?
8. Какие два шага используются для группировки элементов данных?
9. Когда происходит остановка этих двух шагов алгоритма?
10. Для каких типов кластеров лучше всего работает кластеризация методом k-средних и почему?
11. Что такое кластеризация методом k-средних?
12. Как определяется число кластеров k?

Вопросы к экзамену:

1. Какие четыре ключевых шага предполагает исследование в рамках науки о данных?
2. В чем заключается этап подготовки данных?
3. Как осуществляется выбор алгоритмов для моделирования данных?
4. Как проходит настройка алгоритмов для оптимизации моделей?
5. Как оценивается точность моделей?
6. Что такое кластеризация методом k-средних?
7. Как определяется число кластеров k?
8. Какие два шага используются для группировки элементов данных?
9. Когда происходит остановка этих двух шагов алгоритма?
10. Для каких типов кластеров лучше всего работает кластеризация методом k-средних и почему?
11. Что такое кластеризация методом k-средних?
12. Как определяется число кластеров k?
13. Какие два шага используются для группировки элементов данных?
14. Когда происходит остановка этих двух шагов алгоритма?
15. Для каких типов кластеров лучше всего работает кластеризация методом k-средних и почему?
16. Что выявляют ассоциативные правила?
17. Каковы три основных способа оценки ассоциации?
18. Что такое "поддержка"?
19. Что такое "достоверность"?
20. Что такое "лифт"?
21. В чем заключается принцип a priori
22. Как определяется метод для анализа социальных сетей?
23. Что такое Лувенский метод?
24. Когда лучше всего работает Лувенский метод?
25. Как работает алгоритм PageRank?
26. В чем преимущества и недостатки PageRank
27. Для чего используется регрессионный анализ?
28. Как определяются предиктор и вес предиктора?
29. Что показывает вес предиктора?

30. Как выводится линия тренда?
31. В каких случаях регрессионный анализ работает лучше всего?
32. Как определяется метод для анализа социальных сетей?
33. Что такое Лувенский метод?
34. Когда лучше всего работает Лувенский метод?
35. Как работает алгоритм PageRank?
36. В чем преимущества и недостатки PageRank
37. В чем методика экспертизы по соседям
38. Что представляет собой метод k-ближайших соседей?
39. Что такое кросс-валидация?
40. Как определяется число k в методе k-ближайших соседей?
41. Когда лучше всего работает метод k-ближайших соседей?
42. Что является верным признаком возможных аномалий?
43. Что такое опорный вектор?
44. Как работает метод опорных векторов (МОВ)?
45. Что такое функция ядра?
46. По отношению к каким значениям устойчив МОВ?
47. Когда лучше всего работает МОВ?
48. Какой прогноз создает дерево решений?
49. В чем заключается процесс рекурсивного деления?
50. Каков критерий остановки рекурсивного деления?
51. Каковы достоинства и недостатки деревьев решений?
52. Какая альтернатива используется для преодоления недостатков дерева решений?
53. Что такое бэггинг?
54. Что такое ансамблирование?
55. Как случайные леса задействуют бэггинг и ансамблирование?
56. Сравните прогнозы деревьев решений и случайных лесов.
57. Как оцениваются предикторы случайных лесов?
58. Опишите общую архитектуру нейронных сетей. Как определяются нейроны, слои нейронов и их количество?
59. Как происходит активация нейронов первого слоя и что формируется в последнем слое нейронной сети?
60. Что такое правило активации?
61. Какой процесс называется методом обратного распространения ошибки?
62. Когда нейронные сети работают лучше всего?

Примеры экзаменационных билетов.

Пример 1.

1. Какие два шага используются для группировки элементов данных?
2. Как оценивается точность моделей?

Пример 2.

1. Когда лучше всего работает Лувенский метод?
2. Как работает алгоритм PageRank?

#### 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**

Зачет в осеннем семестре выставляется при посещении не менее 80% лекций, а также в случае успешного ответа на заключительном коллоквиуме.

При проведении устного экзамена обучающемуся предоставляется 1 час на подготовку. Опрос обучающегося на экзамене не должен превышать одного астрономического часа.