**Федеральное государственное образовательное**
**бюджетное учреждение высшего образования**

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ**
**ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**
**(Финансовый университет)**

**Кафедра анализа данных и машинного обучения**

**Факультета информационных технологий и анализа больших данных**

**С.В. Макрушин, В.А. Малекова**

# Методические указания по выполнению курсовых работ (проектов) по дисциплине «Машинное обучение в семантическом и сетевом анализе»

**Москва 2024**

**Цель работы:**

Продемонстрировать владение моделями и алгоритмами машинного обучения в семантическом и сетевом анализе. В том числе их программная реализация (с использованием NetworkX, PyTorch и других библиотек) и применение к наборам данных, в том числе в практических приложениях, методами сбора и подготовки набора данных, анализа качества обучения модели глубокого обучения, умение делать выводы из проведенного анализа.

**Задачи для выполнения:**

1. Выбрать (собрать) набор данных для анализа в соответствии с выбранной темой курсовой работы. Описать этот набор и решаемую задачу. Определить метрики качества для модели, определить типичный уровень метрик для решения аналогичных задач, установить ориентиры для метрики.
2. Провести предварительный анализ и очистку данных. Этот этап включает в себя вывод информации о количественных характеристиках датасета, информацию об отсутствующих значениях, характеристиках и физическом смысле каждого атрибута данных, его значимости для предсказания целевой переменной, вывод нескольких элементов данных для иллюстрации структуры данных.
3. При необходимости выполнить преобразование данных. Этот этап сильно зависит от типа исследуемых данных и может включать в себя формирование графа, графа знаний, задание атрибутов, в том числе текстовых, создание векторного представления, извлечение признаков из исходных и другие преобразования.
4. Разделить набор данных на обучающую, тестовую и валидационную выборки. Обосновать количественные характеристики и метод разделения выборки.
5. Определить принципиальную архитектуру нейросетевой модели глубокого обучения, включая функцию ошибки и методы регуляризации, общую логику обучения модели. Предусмотреть возможные варианты модификации модели.
6. Реализовать модель с помощью фреймворка PyTorch или специализированных фреймоврков для работы с графовыми данными. Реализация должна включать функционал загрузки данных, вывода промежуточных и финальных метрик качества и диагностических параметров процесса обучения модели. Реализация должна предусматривать гибкость, позволяющую выполнять ограниченную модификацию модели для выбора наилучшей из альтернативных архитектур.
7. Настроить параметры входных данных и параметров для обучения. Провести обучение модели на подготовленных входных данных используя выбранные настройки и алгоритмы обучения.
8. Провести оценку качества модели с использованием предусмотренных метрик качества, проверить модель на переобучение.
9. Провести обучение и оценку качества для несколько альтернативных архитектур (вариаций архитектур) моделей (не менее 3) для решения выбранной задачи. Проанализировать результаты, сделать выводы, в том числе определить наилучшую архитектуру.
10. Выполнить тонкую настройку модели с помощью подбора значений гиперпараметров. Провести подбор не менее трех гиперпараметров, при этом как минимум для одного подбор значения параметра выполнить в автоматическом режиме (с помощью Grid Search или аналогичных методов).

11. Представить результаты обучения модели в наглядном виде (графики, линии обучения, таблицы сравнения моделей, таблицы классификации, и другие). Сделать выводы, сравнить с существующими аналогичными решениями, порассуждать о перспективах улучшения методов решения проблемы.

В зависимости от формулировки выбранной темы, объем и наличие пунктов их этого списка может варьироваться.

**Методические указания:**

1. Работа выполняется в виде программного ноутбука Python Jupyter. Пояснительная записка выполняется в виде текстового документа и должна включать в себя: титульный лист, текстовое описание проблемы, ссылку на публично доступный репозиторий с полным кодом выполнения работы, по необходимости пример кода для каждого этапа работы, текстовые выводы по каждому этапу и сформулированное заключение с результатами работы и их интерпретация. Подробнее типовая структура пояснительной записки приводится ниже.
2. Все пояснения, выводы и замечания, на которые необходимо обратить внимание должны присутствовать в работе в виде ячеек документации либо (менее предпочтительно) программных комментариев.
3. Работа должна выполняться студентом самостоятельно и индивидуально.
4. Оценка качества моделирования должна производиться с использованием определенных метрик. Их выбор должен быть описан и обоснован до начала моделирования. Плюсом работы является широкий набор метрик эффективности моделей.
5. Отчет работы производится в формате презентации. Слушатели (включая преподавателя) могут задавать вопросы представляющему свою работу студенту. Регламент презентации - 5 минут на выступление, 2 минуты на вопросы.

Типовая структура пояснительной записки. Традиционно курсовая работа состоит из следующих частей:

1. Содержание (оглавление)
2. Введение
3. Основная часть, разделенная на пронумерованные пункты (главы) и подпункты (параграфы)
4. Заключение
5. Список использованных источников
6. Содержание представляет собой план работы.

В содержании необходимо перечислить названия всех глав и параграфов работы. Главы и параграфы нумеруются арабскими цифрами (1, 1.1, …2, 2.1, 2.2, …). Введение и заключение не нумеруются. В конце названия глав и параграфов точка не ставится. Примерное содержание каждого пункта должно быть ясным из его названия.

Введение в курсовой работе должно кратко знакомить читателя с темой. Объем – от одной до трех страниц. Здесь описывается объект и предмет исследования, ставится цель и задачи исследования, определяется круг вопросов, на которые должен ответить курсовая работа. Введение рекомендуется начать с развернутого представления темы исследования.

Обоснованность выбора темы вытекает в актуальность работы, ее значение для практического опыта. Раскрытие проблемы, лежащей в основе курсового исследования, сделает актуальность работы более убедительной.

## 1. Введение

- Описание текущей проблемы
- Цель и задачи исследования

## 2. Обзор литературы

- Рассмотрение существующих методов и подходов
- Оценка предыдущих исследований по теме

## 3. Теоретическая часть

- Описание основных концепций и теорий
- Обоснование выбора методов и подходов

## 4. Методология

- Описание методологии исследования
- Обзор используемых технологий и инструментов

## 5. Практическая часть

- Реализация программного кода (если речь идет о программировании)
- Проведение вычислительных экспериментов или моделирования
- Анализ полученных результатов

## 6. Результаты и обсуждение

- Презентация полученных результатов
- Их интерпретация и обсуждение
- Выявление преимуществ и недостатков предложенного решения

## 7. Заключение

- Подведение общих итогов исследования
- Формулирование выводов по работе

## 8. Список использованных источников

- Перечень литературы, статей, руководств и других материалов

## 9. Приложения

- Дополнительные материалы, код, таблицы, графики и т.д.

Основная часть работы освещает вопросы работы, такие как:

- теоретическая часть работы: описание основных концепций и теорий и обоснование выбора методов и подходов;
- методология исследования и выбор технологического стека и архитектуры решения;
- практическая часть работы: описание программного кода решения; проведение вычислительных экспериментов или моделирования;
- анализ полученных результатов.

При делении пояснительной записки на пункты старайтесь делать их примерно одинаковыми по размеру.

При сборе материала для курсовой рекомендуется сразу оформлять библиографические данные источников. Рекомендуется использовать косвенное цитирование, так как пересказ различных информационных источников (а не дословные выдержки) экономит место, делает курсовую содержательно грамотнее. Прямое цитирование берите в кавычки и указывайте источник.

Результаты сравнительного анализа рекомендуется всегда представлять в табличном виде.

В заключении, как правило, не содержится новой информации. В нем повторяются выводы, вытекающие из содержания работы. Заключение в работе – это ответы на вопросы, которые поставлены во введении. Если были написаны краткие выводы по каждому пункту в основной части, их можно повторить.

Последний пункт – список использованных источников. Чтобы выполнить хорошую курсовую работу, необходимо использовать современные источники информации. В области стандартизации регулярно происходит обновление ранее существовавших стандартов, появление новых. Появляются также новые подходы и методологии по организации разработки м сопровождения ПО. Поэтому нужно внимательно проверять на сайтах ISO и ГОСТ Р, являются ли действующими найденные стандарты, подбирать различные издания, содержащие свежую информацию об интересующих вопросах. Рекомендуется использовать от пяти до десяти источников.

**Критерии оценки:**

1. Структурированность отчета. В работе должна прослеживаться четкая структура - подготовительный этап, анализ данных, построение простых моделей, сравнение и анализ моделей, выводы, построение моделей с учетом выводов, итоговый результат.
2. Наличие выводов. Работа должна содержать текстовые замечания, поясняющие каждый шаг работы студента: что делается, зачем и какую информацию это нам дает. Оценивается полнота и адекватность выводов.

3. Визуализация, анализ и представление результатов (наборов данных, тестирование реализованных алгоритмов, выполненное сравнение и т.д.). Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.
4. (в зависимости от темы) Разнообразие моделей. Студент должен продемонстрировать умение работать с разнообразными моделями, применимыми к одной задаче.
5. (в зависимости от темы) Оценка качества модели/алгоритма. Студент должен продемонстрировать умение реализовывать, улучшать и оценивать работоспособность созданной модели/алгоритма.
6. (в зависимости от темы) Объем и качество собранного набора данных.
7. Использование метрик эффективности и оценка валидности результатов. Оценивается разнообразие и адекватность задаче примененных метрик эффективности (включая время обучения); корректность проверки модели на переобучение; полнота сравнения и правильность выводов из сравнения моделей по разным метрикам.

## Сроки выполнения:

Указаны в соответствующем разделе сайта кафедры:
http://www.fa.ru/org/chair/findata/Pages/kurs.aspx

*Приложение 1*
*Примерные темы курсовых работ по дисциплине*
*«Машинное обучение в семантическом и сетевом анализе»*

**Создание графовых наборов данных:**

1. Создание согласованной базы знаний по узкой предметной области на основе крупномасштабной базы знаний или открытой базы данных.
2. Построение согласованной базы знаний по предметной области на основе нескольких источников данных.
3. Создание и анализ графового набора данных для заданной предметной области на основе крупномасштабной базы знаний или веб-скрейпинга.
4. Создание и анализ графового набора данных для заданной предметной области на основе открытой базы данных или веб-скрейпинга.
5. Создание и анализ графового набора данных о таксономии товаров, услуг или видов деятельности с помощью веб-скрэйпинга или анализа открытой базы данных.
6. Создание и анализ графового набора данных о художественных произведениях и метаинформации о них с помощью веб-скрэйпинга.
7. Эмпирическое сравнение различных техник сэмплирования для задач анализа крупного графа.
8. Создание и анализ графового набора данных о таксономии товаров заданной предметной области с помощью веб-скрэйпинга интернет-магазина.
9. Создание первичного источника информации для построения базы знаний на основе веб-скрейпинга и анализа и нормализации данных с помощью LLM.
10. Создание и использование набора данных об экономической активности компаний для задач машинного обучения

**Сообщества в сетях:**

11. Сравнение различных методов выделения сообществ на крупных графовых наборах данных.
12. Интерпретируемые методы выделения сообществ на атрибутированных графовых наборах данных.
13. Анализ качества и устойчивости выделения сообществ с помощью дискретного алгоритма распространения меток на искусственно сгенерированных данных
14. Реализация различных методов выделения сообществ, основанных на дискретном алгоритме распространения меток и их сравнительный анализ.
15. Реализация и сравнение различных спектральных методов выделения сообществ и их сравнительный анализ.
16. Реализация методов визуализации структуры сети с помощью спектральных методов анализа сети.
17. Оптимизация гиперпараметров методов выделения сообществ для алгоритмов, основанных на случайных блужданиях

18. Анализ динамики модулярности в многошаговых алгоритмах поиска сообществ в сетях.
19. Реализация алгоритма PageRank и анализ его применения для реальных сетей.

**Визуализация, интерпретация, кластеризация:**

20. Сравнительный анализ и развитие методов укладки графов.
21. Реализация интерактивных методов визуализации графов.
22. Сравнительный анализ методов визуализации крупных наборов эмбеддингов сложных объектов.
23. Сравнительный анализ методов построения интерпретируемых эмбеддингов сложных объектов.
24. Методы интерпретируемой кластеризация эмбеддингов сложных объект.

**Рекомендательные системы:**

25. Построение рекомендательной системы с помощью векторного представления графа.
26. Построение рекомендательной системы с помощью векторного представления графа знаний.
27. Построение рекомендательной системы с помощью графовых нейронных сетей использующих атрибуты товаров и пользователей.
28. Построение рекомендательной системы в области закупок с использованием графовых нейронных сетей
29. Построение рекомендательной системы в области закупок с использованием графовых нейронных сетей
30. Рекомендация исполнителей для выполнения контрактов на закупки с учетом географического расположения контрагентов
31. Рекомендация исполнителей для выполнения контрактов на закупки с учетом бюджета и времени выполнения контракта
32. Построение рекомендательной системы в области закупок с учетом характеристик контракта

**Эмбеддинги:**

33. Построение и сравнительный анализ эмбеддингов узлов графового набора данных с помощью современных методов.
34. Построение эмбеддингов узлов атрибутированного графового набора данных с помощью современных методов.
35. Построение эмбеддингов узлов графового набора данных из специализированной предметной области с помощью современных методов.
36. Построение эмбеддингов узлов графовой базы знаний с помощью современных методов построения эмбеддингов баз знаний.
37. Построение эмбеддингов узлов графового набора данных из специализированной предметной области с помощью графовых нейронных сетей.
38. Реализация и применение алгоритма для построения векторных представлений в графах знаний.

39. Сравнительный анализ качества решения задачи предсказания связи в графах знаний с помощью различных трансляционных алгоритмов.
40. Построение онтологических эмбеддингов графов знаний

**Классификация узлов:**

41. Сравнение решений задачи классификации узлов сети с использованием различных методов векторных представлений узлов.
42. Классификация узлов сети с помощью графовой нейронной сети, использующей атрибуты узлов.
43. Решение задачи предсказания частично заполненной числовой характеристики узлов сети с помощью графовой нейронной сети.
44. Предсказание центральности узлов по PageRank с помощью графовых нейронных сетей.

**Задачи NLP:**

45. Построение интеллектуального агента с помощью технологий LLM.
46. Автоматизация решения задач промпт-инжениринга для LLM.
47. Автоматизация оценки качества работы моделей на основе LLM с помощью применения оценок на основе технологий LLM.
48. Структурирование атрибутивной информации о большом наборе сложных объектов с помощью LLM.

*Приложение 2*
*Образец оформления титульного листа*

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**
**(ФИНАНСОВЫЙ УНИВЕРСИТЕТ)**

Кафедра анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных

***Дисциплина: «*Машинное обучение в семантическом и сетевом анализе»**
*Направление подготовки: «Прикладная математика и информатика»*
*Профиль: «Прикладное машинное обучение»*
*Факультет информационных технологий и анализа больших данных*
*Форма обучения очная*
*Учебный 2023/2024 год, 6 семестр*

**Курсовая работа на тему:**
«…….»

*Выполнил(а):*

студент(ка) группы ПМ21-1

Лашуков Т. Д.

*Научный руководитель:*

доцент к.э.н. Макрушин С. В.

**Москва 2024**

*Приложение 3*
*Образец оформления списка литературы*

**Основная литература:**


1. Кочкаров, А.А., Экспериментальная теория графов и алгоритмы анализа сетевых моделей : учебное пособие / А. А. Кочкаров, С. В. Макрушин, В. Е. Каменчук, Н. В. Блохин. — М.: КноРус, 2024. — 160 с.
2. Коротеев, М.В. Технологии анализа данных и машиннон обучение. Учебное пособие для самостоятельной работы. 1 семестр. / М.В. Коротеев – М.: Финансовый университет, департамент анализа данных, принятия решений и финансовых технологий, 2018. – 48 с.


**Дополнительная литература:**


1. Грас, Джоэл. Наука о данных с нуля [Текст] / Джоэл Грас; [пер. с англ. Андрея Логунова]. – Санкт – Петербург: БХВ-Петербург, 2017. – 336 с.

2. Geron, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems / A. Geron – O'Reilly Media, Inc., 2019. – 856 p.

3. Albon, C. Machine learning with Python Handbook / C. Albon. – O'Reilly Media, Inc, 2018. – 366 p.

4. Coelho, L.P. Building machine learning systems with Python / L.P. Coelho, W. Richert – Packt Publishing Ltd, 2015. – 326 p.

5. McKiney, W. Pandas: powerful Python data analysis toolkit / W. McKiney – O'Reilly Media, Inc., 2016. – 1971 p.