

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

МЕТОДИЧЕСКИЕ УКАЗАНИЯ
к выполнению курсовой работы
по дисциплине «Групповой проект по интеллектуальному анализу данных»
для бакалавров направления 09.03.04 «Программная инженерия»

Ставрополь

2025

1. ВИД, СТРУКТУРА, СОДЕРЖАНИЕ И ОБЪЕМ КУРСОВОЙ РАБОТЫ

Выполнение курсовой работы по направлению программы бакалавриата «Программная инженерия» реализуется в восьмом семестре обучения как вид учебной деятельности по дисциплине «Групповой проект по интеллектуальному анализу данных», результатом которого является курсовая работа, предусмотренная учебным планом и выполняемый студентом самостоятельно под руководством преподавателя.

Курсовая работа студента - заключительный этап изучения дисциплины «Групповой проект по интеллектуальному анализу данных». Цель работы - систематизация и закрепление теоретических знаний, полученных за время обучения, а также приобретение и закрепление навыков самостоятельной работы. Курсовая работа, как правило, основывается на обобщении выполненных студентом лабораторных работ или представляет собой индивидуальное задание по изучаемой дисциплине и подготавливается к защите в завершающий период теоретического обучения.

2. ЦЕЛИ И ТЕМАТИКА КУРСОВЫХ РАБОТ

Основная цель выполнения курсовой работы: продемонстрировать владение основными навыками работы с методами интеллектуального анализа данных, владение основными инструментальными средствами библиотек языка программирования Python, методами и приемами подготовительного и описательного анализа данных, средствами визуализации данных, использования и усовершенствования обучаемых моделей, умение делать выводы из проведенного анализа.

Каждая курсовая работа строго индивидуальна и ориентирована на развитие у студента определенной части профессиональных навыков и умения творчески решать практические задачи. По результатам выполнения курсовой работы оформляется пояснительная записка.

Содержание пояснительной записки включает следующие разделы:

1. Титульный лист
2. Задание на курсовую работу
3. Содержание
4. Введение, в котором описывается актуальность разработки для конкретной прикладной области.
5. Анализ прикладной области, приводящий к конкретной постановке задачи с выделением совокупности понятий (интенционалы и экстенционалы понятий), подлежащих анализу методами интеллектуального ана-

лиза.

6. Постановка задачи, то есть:

А. Цель разработки в конкретной прикладной области для выбранных данных.

Б. Перечень и методы решаемых задач интеллектуального анализа для конкретных данных из выбранной прикладной области.

В. Общие требования к данным и результатам интеллектуального анализа на уровне вход/выход.

Г. Техническое задание с выбором конкретных методов решаемых задач интеллектуального анализа.

7. Проектирование:

А. Анализ аналогов решения обозначенных в предыдущем разделе задач.

Б. Выбор, системное и математическое описание конкретного метода (методов) интеллектуального анализа для решения поставленных задач. Описание мер точности и адекватности применяемых методов.

В. Описание контрольного примера для каждого метода интеллектуального анализа.

Г. Проектирование и описание системы хранения данных и систем знаний.

Д. Проектирование архитектуры программной системы.

8. Реализация:

А. Выбор инструментальных средств.

Б. Описание характеристик программной системы.

В. Разработка и описание пользовательского интерфейса.

Г. Описание используемых компонент системы.

9. Инструкция пользователю с копиями экранных форм.

10. Заключение, в котором отразить основные результаты, полученные в результате выполнения курсовой работы, а также чем отличается разработанный проект от аналогов, приведенных в пункте 5А.

11. Список используемых источников (литературы, электронных ресурсов, отчетов) не менее 15.

12. Приложения, включая листинг программного продукта

Основные этапы выполнения курсовой работы:

1. Выбрать набор данных для анализа в соответствии с выбранной темой курсовой работы. Описать этот набор и решаемую задачу.

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 точек данных для иллюстрации структуры данных.

3. При необходимости, преобразовать атрибуты исходного датасета

в числовые признаки. Этот этап сильно зависит от типа исследуемых данных и может включать в себя векторизацию текста, извлечение признаков их аудио и видео данных, преобразование изображений в плоский численный массив и другие преобразования.

4. Провести описательный анализ данных. Сделать выводы. Этот этап включает в себя определение шкалы измерения каждого признака, выявление аномальных значений, визуализацию распределения каждого признака, при необходимости - проверка на нормальность, построение корреллограмм и совместных распределений каждого признака с целевой переменной, выявление коррелированных признаков и признаков, не несущих информации для данной задачи.

5. Применить при необходимости к данным методы обучения без учителя: кластеризацию, понижение размерности и поиск аномалий. Сделать выводы.

6. Разделить набор данных на обучающую и тестовую выборки. Обосновать количественные характеристики и метод деления (временной, случайный, последовательный).

7. Обучить несколько моделей для решения выбранной задачи (для задач классификации - не менее 7 различных алгоритмов). Проанализировать результаты, сделать выводы.

8. Выбрать наиболее перспективную модель для решения поставленной задачи. Изменить гиперпараметры модели. Предпочтительно, провести Grid Search. Найти оптимальные гиперпараметры.

9. С учетом сделанных выводов провести усовершенствование моделей. Это можно осуществить с помощью введения регуляризации, изменение параметров модели (для параметрических моделей), введением суррогатных признаков, отбором признаков, нормализацией данных, ансамблированием моделей, изменением алгоритма предварительной обработки данных. Сравнить результаты.

10. Попробовать изменить порядок предобработки данных для повышения эффективности модели. Попробовать применить понижение размерности для создания суррогатных признаков. Сравнить результаты, сделать выводы.

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

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

Студенты могут предлагать свои инициативные темы. В таких случаях студенты должны представить необходимое обоснование целесообразности разработки выбранной темы.

Курсовая работа должна выполняться студентами в соответствии с графиком, разрабатываемым руководителем.

Примерный график выполнения курсовой работы по дисциплине «Ин-

теллектуальный анализ больших данных»:

1 неделя: Получение и анализ постановки задачи.

2 неделя: Согласование требований. Подготовка предварительной презентации с постановкой задачи пункты 1-4. Первая аттестация.

3 неделя: Разработка ТЗ.

4 неделя: Уточнение, модификация структуры приложения, согласование и описание.

5,6,7 недели: Проектирование приложения в программной среде.

7 неделя: Предварительный показ приложения, уточнение интерфейса. Вторая аттестация.

8 неделя: Отладка.

9-10 недели: Написание пояснительной записки.

11-12 недели: Сдача на предварительную проверку пояснительной записки и итоговой презентации.

13-15 недели: Доработка курсовой работы. Подготовка презентации по курсовой работе.

15-16 недели: Сдача курсовой работы на проверку в электронной форме в виде презентации.

17 неделя: Защита курсовой работы.

Курсовая работа выполняется по индивидуальным заданиям, которые обновляются ежегодно.

Исходные данные к курсовой работе по дисциплине «Групповой проект по интеллектуальному анализу данных» выбираются на одном из репозитариев, предназначенных для хранения больших массивов данных. После выбора исходного массива данных преподаватель определяет цель их обработки (кластеризация, классификация, прогнозирование и т.д.).

Работа выполняется в виде программного ноутбука Python Jupyter. Все пояснения, выводы и замечания, на которые необходимо обратить внимание должны присутствовать в работе в виде ячеек документации либо (менее предпочтительно) программных комментариев

Ответственность за принятые решения, качество исполнения презентации и пояснительной записки несет автор проекта. Руководитель курсовой работы несет ответственность за организацию и обеспеченность процесса проектирования, обеспечение контроля ритмичности работы, своевременности завершения, соответствия принимаемых инженерных решений уровню развития и современному состоянию отраслей.

ЗАЩИТА КУРСОВОЙ РАБОТЫ

Выполненная курсовая работа решением руководителя допускается к защите, о чем на титульном листе делается надпись «К защите».

Для допуска к защите пояснительная записка должна быть подписана студентом - автором проекта.

Для допуска к защите курсовой работы необходимо:

1. Представить презентацию по курсовой работе, содержащей основные этапы выполнения.

2. Продemonстрировать и объяснить работу разработанной программной системы.

3. Предоставить подписанную пояснительную записку с выполненным заданием в электронном и печатном вариантах.

4. Защита курсовой работы должна осуществляться в конце третьего семестра.

При защите курсовой работы студент в своем докладе должен раскрыть основные вопросы:

- 1) назначение, область применения объекта проектирования;
- 2) обоснование выбранных методов анализа данных и проектных решений;
- 3) полученные результаты и степень адекватности примененных методов интеллектуального анализа.

Время, отводимое студенту на доклад, должно быть ограничено (8-10 мин). Вопросы, задаваемые студенту на защите, не должны выходить за рамки тематики курсовой работы и тех конкретных задач, которые решались студентом в процессе выполнения курсовой работы.

Защита курсовых работ по комплексным программам должна в обязательном порядке осуществляться в один день при участии всех членов рабочей группы авторов работы. Защиту таких работ целесообразно организовывать в строгой последовательности авторов отдельных частей, логически вытекающих одна из другой. Порядок такой защиты должен быть оговорен заранее на стадии выдачи задания и доведен до каждого исполнителя.

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. (рекомендуется) Замеры времени. В целях анализа временной сложности алгоритмов. Все инструкции, запускающие цикл обучения модели должны содержать замер времени обучения. Замер можно производить с помощью магических инструкций Jupyter или (более предпочтительно) с использованием стандартной библиотеки Python. Сравнение моделей должно учитывать и время обучения.

4. Визуализация. Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.

5. (в зависимости от темы) Разнообразие моделей. Студент должен продемонстрировать умение работать с разнообразными моделями обучения, применимыми к одной задаче. Например, в задачах классификации существует как минимум десять наиболее применимых моделей. Оценивается число

алгоритмов, примененных студентом для одной и той же задачи.

6. (в зависимости от темы) Улучшение модели. Студент должен продемонстрировать умение анализировать обученную модель и искать пути для ее совершенствования. Оценивается количество итераций совершенствования модели и их эффективность.

7. (рекомендуется) Использование конвейеров. Студент должен продемонстрировать умение строить сложные последовательности операций при помощи программных конвейеров библиотеки `scikit learn`. Оценивается сложность и уместность использования контейнеров.

8. Предобработка данных. Работа должна содержать исчерпывающий алгоритм предварительной обработки данных. Он служит для того, чтобы исправить все несовершенства в данных и сделать набор данных как можно более пригодным для машинного обучения. Оценивается сложность и воспроизводимость процедуры предварительной обработки данных.

9. Использование метрик эффективности. Оценивается разнообразие и адекватность задаче примененных метрик эффективности (включая время обучения) а также полнота сравнения и правильность выводов из сравнения моделей по разным метрикам.

10. Валидность результатов. Студент должен продемонстрировать умение оценивать достоверность измерения метрик моделей и повышать ее с использованием перекрестной проверки (кросс-валидации). Использование `k-fold cross validation` является предпочтительным методом измерения эффективности модели. Если происходит выбор модели, то ее итоговая эффективность должна измеряться на чистом наборе данных.

Оценка за курсовую работу определяется суммой баллов, полученных на каждой аттестации выполненной работы. Всего три аттестации (см. график выполнения проекта).

Максимальное количество баллов за курсовую работу 15. Правила перехода к четырехбалльной системе оценивания представлены ниже:

«отлично» - сумма баллов 15-13,

«хорошо» - сумма баллов 10-12,

«удовлетворительно» - сумма баллов 9-10,

«неудовлетворительно» - сумма баллов менее 9.

При положительном результате защиты оценка записывается в зачетную книжку за подписью руководителя проекта и проставляется на титульном листе.

Студент, не представивший в установленный срок курсовую работу или не защитивший ее, считается имеющим академическую задолженность.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Ярушкина Н. Г., Афанасьева Т. В., Перфильева И. Г. Интеллектуальный анализ временных рядов : учебное пособие - М. : ИД «ФОРУМ»ИНФРА-М, 2012. - 160 с. - (Высшее образование).

2. Нечеткие гибридные системы. Теория и практика / под ред. Н. Г. Ярушкиной. - М. : ФИЗМАТЛИТ, 2007. - 208 с.
3. Ярушкина Н.Г., Афанасьева Т.В., Перфильева И.Г. Интеллектуальный анализ временных рядов : учебное пособие. - Ульяновск : УлГТУ, 2010. - 324 с.
4. Хайкин С. Нейронные сети: полный курс. - М. : Вильямс, 2006, 1104 с.
5. Ярушкина Н.Г. Основы теории нечетких и гибридных систем: учебное пособие. - М. : Финансы и статистика, 2004. - 320 с.
7. Андрейчиков А. В. Интеллектуальные информационные системы: учебник для вузов. - М. : Финансы и статистика, 2004. - 423 с.
8. Пупков К. А., Коньков В. Г. Интеллектуальные системы: (Исследование и создание) - М. : МГТУ им. Н. Э. Баумана, 2003. - 347 с.
9. Рассел С., Норвиг П. Искусственный интеллект: современный подход. - М. : Вильямс, 2005. - 1424 с.
10. Бокс Дж. Анализ временных рядов. Прогноз и управление. - М. : Мир, 1974.
11. Афанасьева Т.В., Ярушкина Н.Г. Нечеткое моделирование и анализ нечетких тенденций. - Ульяновск: УлГТУ, 2009.