

Курсовая работа по дисциплине “Технологии анализа данных и машинное обучение”

Цель работы

Продemonстрировать владение основными навыками работы с методами машинного обучения с учителем и без учителя, 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.

Задания для выполнения

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 кореллограмм 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 разделения (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. Например, при разработке темы “Описательный анализ данных ...” следует более подробно остановиться на пунктах 2,3,4,5, а пункты 7,8,9,10 могут отсутствовать или реализовываться для примера. При реализации тем “Машинное обучение в задачах ...” наоборот, пункты 2,3,4,5 должны реализоваться в необходимом объеме, а пункты 7,8,9,10 нужно раскрыть как можно более подробно. Пункты 1,3,6,11 являются обязательными для всех тем курсовых работ.

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 интерпретации.

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.

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. 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. 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. Улучшение модели. Студент должен продемонстрировать умение анализировать обученную модель и искать пути для ее совершенствования. Оценивается количество итераций совершенствования модели и их эффективность.
7. 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.

Сроки выполнения

Указаны в соответствующем [разделе](#) сайта Департамента.

Примерные темы курсовых работ

1. Deskриптивный анализ численного набора данных с использованием технологий визуализации.
2. Deskриптивный анализ категориального набора данных с использованием технологий визуализации.
3. Deskриптивный анализ смешанного набора данных с использованием технологий визуализации.

4. Сравнение методов регрессии на реальных наборах данных.
5. Сравнение методов классификации на реальных наборах данных.
6. Предварительный анализ данных и построение признаков в задачах обработки финансовой и экономической информации.
7. Предварительный анализ данных и построение признаков в задачах распознавания голоса.
8. Предварительный анализ данных и построение признаков в задачах распознавания текста.
9. Предварительный анализ данных и построение признаков в задачах распознавания объектов на фотографии.
10. Предварительный анализ данных и построение признаков в задачах сжатия информации.
11. Предварительный анализ данных и построение признаков в задачах повышения качества изображений.
12. Предварительный анализ данных и построение признаков в задачах идентификации личности по голосу.
13. Предварительный анализ данных и построение признаков в задачах идентификации личности по изображению.
14. Предварительный анализ данных и построение признаков в задачах обработки текстов на естественных языках.
15. Предварительный анализ данных и построение признаков в задачах машинного перевода.
16. Предварительный анализ данных и построение признаков в задачах распознавания темы текста.
17. Предварительный анализ данных и построение признаков в задачах нормализации слов текста.
18. Предварительный анализ данных и построение признаков в задачах извлечения знаний из аудиоданных.
19. Предварительный анализ данных и построение признаков в задачах медицинской диагностики.

- 20.Предварительный анализ данных и построение признаков в задачах извлечения знаний из видеоданных.
- 21.Предварительный анализ данных и построение признаков в задачах классификации текстов.
- 22.Предварительный анализ данных и построение признаков в задачах кредитного скоринга.
- 23.Предварительный анализ данных и построение признаков в задачах оценки активов.
- 24.Предварительный анализ данных и построение признаков в задачах предсказания оттока клиентов.
- 25.Предварительный анализ данных и построение признаков в задачах противодействия коррупции.
- 26.Предварительный анализ данных и построение признаков в задачах верификации финансовых транзакций.
- 27.Предварительный анализ данных и построение признаков в задачах анализа социальных графов.
- 28.Предварительный анализ данных и построение признаков в задачах визуализации информации.
- 29.Машинное обучение в задачах обработки финансовой и экономической информации.
- 30.Машинное обучение в задачах распознавания голоса.
- 31.Машинное обучение в задачах распознавания текста.
- 32.Машинное обучение в задачах распознавания объектов на фотографии.
- 33.Машинное обучение в задачах сжатия информации.
- 34.Машинное обучение в задачах повышения качества изображений.
- 35.Машинное обучение в задачах идентификации личности по голосу.
- 36.Машинное обучение в задачах идентификации личности по изображению.

- 37.Машинное обучение в задачах обработки текстов на естественных языках.
- 38.Машинное обучение в задачах машинного перевода.
- 39.Машинное обучение в задачах распознавания темы текста.
- 40.Машинное обучение в задачах нормализации слов текста.
- 41.Машинное обучение в задачах извлечения знаний из аудиоданных.
- 42.Машинное обучение в задачах медицинской диагностики.
- 43.Машинное обучение в задачах извлечения знаний из видеоданных.
- 44.Машинное обучение в задачах классификации текстов.
- 45.Машинное обучение в задачах кредитного скоринга.
- 46.Машинное обучение в задачах оценки активов.
- 47.Машинное обучение в задачах предсказания оттока клиентов.
- 48.Машинное обучение в задачах противодействия коррупции.
- 49.Машинное обучение в задачах верификации финансовых транзакций.
- 50.Машинное обучение в задачах анализа социальных графов.
- 51.Машинное обучение в задачах визуализации информации.
- 52.Исследование эффективности ансамблевых моделей на примере задачи классификации на реальных данных.
- 53.Исследование эффективности ансамблевых моделей на примере задачи регрессии на реальных данных.
- 54.Исследование эффективности различных методов шкалирования данных в задачах регрессии.
- 55.Исследование эффективности различных методов шкалирования данных в задачах классификации.
- 56.Исследование эффективности различных методов оптимизации гиперпараметров в задачах машинного обучения.

Пример титульного листа пояснительной записки

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»**

Департамент анализа данных и машинного обучения

Пояснительная записка к курсовой работе
по дисциплине “Технологии анализа данных и машинное обучение”
на тему:

«Машинное обучение в задачах анализа текстов»

Выполнил(а):

студент(ка) группы ПМ18-1
факультета информационных
технологий и анализа больших
данных

_____Иванова Е.Е.

Научный руководитель:

доцент, к.э.н. Макрушин С.В.

2022 г

Список литературы

Основная литература

- [М.В. Коротеев. Об основных задачах дескриптивного анализа данных.](#)
- [М.В. Коротеев. Учебное пособие по дисциплине “Анализ данных и машинное обучение” - 2018.](#)

Дополнительная литература

- A. Geron. Hand on Machine Learning with scikit-learn and Tensorflow - 2017 (564p)
- C. Albon. Machine learning with Python Handbook - 2018 (427p)
- L.P. Coelho, W. Richert. Building machine learning systems with Python - 2015 (326p)
- J. Grus. Data science from scratch - 2015 (330p)
- W. McKinney. Pandas: powerful Python data analysis toolkit - 2016 (1971p)

Видео-ресурсы

- [Классический курс по машинному обучению](#)
- [Конспекты](#)
- [Материалы на гитхабе](#)
- [Pandas tutorial](#)
- Канал [Python programmer](#) - много контента по программированию на питоне
- Серия [Data analysis](#) от Computerphile