Курсовой проект по дисциплине «Машинное обучение»

Цель работы

Продемонстрировать 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 их 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 интерпретацие.
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.

Примерные темы курсовых проектов

1. Дескриптивный анализ данных о пассажирах авиакомпаний с использованием технологий визуализации
2. Аугментация табличных данных и её влияние на качество ML-моделей
3. Моделирование динамики распространения информации в социальных сетях
4. Сегментация медицинских изображений УЗИ с использованием машинного обучения
5. Bootstrap доверительные интервалы BTC halving регрессий
6. Calibrated Logistic Regression для rug pull классификации. Imbalanced Precision-Recall
7. CatBoost+LightGBM stacking ансамбли для анализа волатильности на финансовый рынках
8. Ridge+Lasso при мультиколлинеарности on-chain (NVT/MVRV/SOPR)
9. SHAP/LIME интерпретируемость BlackBox моделей для ЦБ compliance. XAI для крипто-аудитов
10. Wavelet denoising + PCA шумоподавление orderbook данных. Signal-to-noise ratio

11. Байесовская оптимизация SMAC3 для LightGBM крипто-класификаторов
12. Гиперпараметры ансамблей: Optuna vs. Hyperopt для XGBoost
13. Оптимизация борьбы с дисбалансом классов: SMOTE+ADASYN
14. Байесовы нейронные сети (Bayesian MLP) в задачах машинного обучения
15. Нейронные ансамбли неразличимых решающих деревьев (NODE) в задачах машинного обучения
16. Диагностика заболеваний сердца на основе клинических данных: сравнение моделей и влияние балансировки классов
17. Прогнозирование оттока клиентов телеком-оператора с использованием ансамблевых методов и интерпретации моделей
18. Байесовская оптимизация для настройки моделей машинного обучения
19. Использование машинного обучения для анализа политических сетей
20. Машинное обучение для анализа и моделирования экономических сетей
21. Многокритериальная оптимизация с помощью машинного обучения
22. Разработка рекомендаций по диверсификации портфеля криптовалют с учетом риска и доходности
23. Эволюционные алгоритмы в оптимизации моделей машинного обучения
24. Методы машинного обучения в задачах управления энергетическими системами
25. Применение методов машинного обучения в задачах управления предприятием
26. Разработка модели мониторинга состояния банка по интегральным показателям
27. Разработка модели МО анализа рыночной корзины для выявления устойчивых наборов товаров, приобретаемых клиентами в супермаркете
28. Автоматическое построение признаков для прогнозирования оттока корпоративных клиентов в сфере программного обеспечения
29. Применение машинного обучения для оценки экологических, социальных и управленческих рисков компаний
30. Прогнозирование динамики децентрализованных финансовых протоколов с использованием ансамблевых моделей на основе временных рядов
31. Разработка алгоритмов адаптивного подбора учебных материалов для построения индивидуальной образовательной траектории студента
32. Сравнительный анализ методов борьбы с дисбалансом классов в задачах медицинской диагностики по изображениям

33. Калибровка и оценка вероятностных предсказаний в задачах бинарной классификации с использованием методов машинного обучения
34. Прогнозирование наличия сердечно-сосудистых заболеваний по медицинским данным с использованием методов машинного обучения
35. Исследование эффективности классификаторов на основе линейного дискриминантного анализа в сравнении с современными методами
36. Кластеризация многомерных данных с использованием технологии визуализации на базе кривых Эндрюса
37. Метод визуальной оценки качества кластеризации на основе кривых Эндрюса
38. Применение информационных критериев Акаике и байесовского для выбора наилучшей модели.
39. Применение классификаторов на базе линейного дискриминантного анализа в задачах машинного обучения
40. Регрессионный анализ сложных данных: нелинейные и сегментные модели
41. Сравнение алгоритмов пошаговой регрессии в задачах машинного обучения.
42. Сравнение методов нелинейной и сегментной регрессий.
43. Сравнение методов построения доверительных интервалов для прогнозов регрессионных моделей (на примере бутстрапа и аналитических методов)
44. Сравнительный анализ алгоритмов машинного обучения для классификации заболеваний на медицинских данных
45. Анализ сообщений в социальных сетях методами машинного обучения
46. Автоматическая классификация текстов по темам с использованием методов машинного обучения
47. Анализ метода главных компонент и его применение для снижения размерности данных
48. Анализ социальных сетей с помощью методов машинного обучения.
49. Ансамбли на основе стекинга в задачах классификации
50. Влияние методов предобработки данных и генерации признаков на качество моделей машинного обучения
51. Выявление сообществ в сетях с использованием алгоритмов кластеризации.
52. Использование машинного обучения для предсказания волатильности криптовалют
53. Исследование эффективности ансамблевых моделей в задаче предсказания оттока клиентов

54. Исследование эффективности ансамблевых моделей на примере задачи классификации на реальных данных.
55. Исследование эффективности ансамблевых моделей на примере задачи регрессии на реальных данных.
56. Исследование эффективности различных методов оптимизации гиперпараметров в задачах машинного обучения.
57. Исследование эффективности различных методов шкалирования данных в задачах классификации.
58. Исследование эффективности различных методов шкалирования данных в задачах регрессии.
59. Классификация изображений одежды с применением методов глубокого обучения
60. Классификация изображений рукописных цифр и символов с использованием классических ML-методов (без глубокого обучения)
61. Классификация клиентов по уровню кредитного риска с использованием методов машинного обучения
62. Классификация тональности текстовых отзывов о фильмах с использованием методов обработки естественного языка
63. Машинное обучение в задачах анализа социальных графов.
64. Машинное обучение в задачах верификации финансовых транзакций.
65. Машинное обучение в задачах визуализации информации
66. Машинное обучение в задачах идентификации личности по голосу.
67. Машинное обучение в задачах идентификации личности по изображению.
68. Машинное обучение в задачах интерпретируемой визуализации моделей
69. Машинное обучение в задачах классификации текстов
70. Машинное обучение в задачах медицинской диагностики.
71. Машинное обучение в задачах обработки финансовой и экономической информации.
72. Машинное обучение в задачах предсказания оттока клиентов.
73. Машинное обучение в задачах распознавания объектов на фотографии
74. Машинное обучение в задачах распознавания темы текста.
75. Моделирование временных рядов с помощью машинного обучения
76. Обнаружение аномальных наблюдений в многомерных данных с применением алгоритмов машинного обучения
77. Обнаружение мошеннических транзакций в банковских данных с использованием алгоритмов машинного обучения
78. Обнаружение мошеннических транзакций с помощью методов обучения без учителя и аномалий
79. Обоснование применения метрик качества решений в системах искусственного интеллекта

80. Определение эмоций по изображениям лиц
81. Оптимизация гиперпараметров в ансамблях моделей
82. Оптимизация методов борьбы с дисбалансом классов в машинном обучении
83. Применение методов искусственного интеллекта при формировании портфеля акций на фондовом рынке
84. Прогнозирование оттока клиентов методами машинного обучения: сравнительный анализ алгоритмов
85. Прогнозирование цен криптовалют с использованием моделей машинного обучения
86. Распознавание дорожных знаков методами машинного обучения
87. Реализация методов активной выборки данных
88. Реализация методов снижения шума в данных
89. Сравнение методов регрессий при наличии мультиколлинеарных признаков.
90. Сравнение моделей при обучении на малом объёме реальных и синтетических данных
91. Сравнительный анализ методов бустинга в задачах регрессии
92. Сравнительный анализ методов регуляризации для улучшения обобщающей способности линейных моделей
93. Сравнительный анализ регуляризованных методов регрессии в условиях мультиколлинеарности признаков
94. Техника синтетического оверсемплинга (SMOTE) в задачах классификации при сильном дисбалансе классов

Пример титульного листа пояснительной записки

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

Кафедра искусственного интеллекта
Факультета информационных технологий и анализа больших данных

Пояснительная записка к курсовому проекту

по дисциплине «Машинное обучение»

на тему:

«Машинное обучение в задачах анализа текстов»

Выполнил(а):

студент(ка) группы ПМ20-1 факультета информационных технологий и анализа больших данных

_____Иванова Е.Е.

Научный руководитель:

доцент, к.э.н. Макрушин С.В.

_____

Москва 2025