**Лабораторные работы (32 часа) и**

**Курсовой проект по дисциплине «Методы машинного обучения»**

1. Установка пакетов: базового Python 4 Shell или более нового, простейшего Python-редактора IDLE, PyСharm Community, Anaconda3 или более нового пакета. Составление простой программы (но не менее 20 строк, например, вывод графиков) и демонстрация ее работы (можно скриншоты + код) на каждом из установленных пакетов. Установка основных пакетов из стека SciPy, Установка Spider и Jupyter Notebook. Демонстрация работы пакетов на тестовых примерах. Поиск и выбор датасета предметной области, например, из репозитория Kaggle. Входной датасет должен включать не менее 20000 записей и не менее 10 признаков (информационных полей). Выбрать возможные информационные поля для целевых векторов задачи классификации. Вывести на экран основные информационные характеристики входного датасета.
2. Исследовать данные входного датасета. Произвести анализ каждого из входных признаков. Исследовать:
   * кодирование данных приведением их к числовым значениям;
   * взаимную корреляцию входных данных;
   * создать тепловую карту признаков.
   * корреляцию каждого признака со значениями целевого вектора (векторов);
3. (4-х часовая работа, обязательна только для получения повышенного рейтинга) Исследовать данные входного датасета. Произвести анализ каждого из входных признаков. Исследовать:
   * построить графики частоты встречаемости значений каждого из признаков для положительного значения целевого вектора (1) и для отрицательного (0), разместить их на одном графике;
   * произвести дискретизацию признаков с изменяемым коэффициентом дискретизации (все значения признака на графической оси его значений разделить на «высокое значение», «среднее значение», «низкое значение»).
4. Определить целевые векторы для решения на выбранном датасете задачи регрессионного анализа. Произвести предварительную подготовку данных для обучения, включающую:
   * + удаление информационных выбросов;
     + исключение противоречий;
     + исключение некорректных значений;
     + обработку пропусков.
5. Определить целевые векторы для решения на выбранном датасете задачи регрессионного анализа. Произвести предварительную подготовку данных для обучения, включающую:
   * + поиск и замену неявных дубликатов;
     + нормализацию данных датасета.
6. Программно реализовать разбиение графиков каждого признака датасета на интервалы с различной частотой встречаемости их в датасете для положительных и отрицательных значений целевого вектора. Дискретность разбиения должна зависеть от поведения признака и иметь возможность настраиваться с помощью переменного коэффициента (коэффициентов) подробности (4-х часовая работа).
7. Оптимизировать пространство признаков, например, методом главных компонент, PCA.
8. Разделить данные на три подвыборки: обучающую, тестовую и валидационную (отложенную). Провести обучение ML-моделей следующими методами:
   * + k ближайших соседей;
     + случайного леса.

Для каждой из моделей опробовать различные параметры их настройки и провести сравнение методов, вычислив характеристики их точности с обязательным использованием графиков ROC AUC.

1. Сбалансировать датафрейм методом imblearn.
2. Разделить данные на три подвыборки: обучающую, тестовую и валидационную (отложенную). Провести обучение ML-моделей следующими методами:
   * + деревьев решений;
     + опорных векторов.

Для каждой из моделей опробовать различные параметры их настройки и провести сравнение методов, вычислив характеристики их точности с обязательным использованием графиков ROC AUC.

1. Разделить данные на три подвыборки: обучающую, тестовую и валидационную (отложенную). Провести обучение ML-моделей следующими методами:
   * + градиентного спуска;
     + логистической регрессии;
     + алгоритмами бустинга.

Для каждой из моделей опробовать различные параметры их настройки и провести сравнение методов, вычислив характеристики их точности с обязательным использованием графиков ROC AUC.

1. Реализовать какой-либо из методов регрессионного анализа.
2. Реализовать какой-либо из методов кластеризации с заданным и неизвестным числом классов.
3. Для задач классификации реализовать модуль автотюнинга гиперпараметров.
4. Реализовать простой ансамбль моделей с метамоделью логистической регрессии. Реализовать методы: стекинг, беггинг, бустинг.
5. Произвести калибровку какой-либо выбранной модели.
6. Реализовать простейшую нейронную сеть для выбранной предметной области. Описать все этапы выполненных работ со сравнительными характеристиками, сделать самостоятельные выводы и оформить пояснительную записку к курсовой работе.
7. Оформить пояснительную записку к курсовой работе.

Задание составил профессор каф. ВПМ И.Ю.Каширин