УНИВЕРСИТЕТ ИТМО

Факультет программной инженерии и компьютерной техники

Направление подготовки 09.03.04 Программная инженерия

Дисциплина «Системы искусственного интеллекта»

**Лабораторная работа 5**

Студент

*Селянта Олег Дмитриевич*

*P3314*

Преподаватель

*Королёва Юлия Александровна*

Санкт-Петербург, 2024 г.

**Введение**

Данная лабораторная работа посвящена реализации алгоритма построения дерева решений для задачи бинарной классификации. В качестве данных используется датасет "Mushroom Classification" из репозитория UCI, содержащий информацию о характеристиках различных видов грибов. Целью работы является построение модели, способной предсказывать, является ли гриб ядовитым (класс 'p') или съедобным (класс 'e') на основе его признаков.

**1. Загрузка и подготовка данных**

* **Загрузка данных:** Датасет был загружен с помощью библиотеки ucimlrepo.
* **Анализ данных:**
  + Датасет содержит 22 признака (features), описывающих внешний вид гриба (форма шляпки, цвет, запах и др.), и целевую переменную (target variable) - "poisonous" (ядовитый/съедобный).
  + Признаки представлены в категориальном формате.
* **Предобработка данных:**
  + **Выбор признаков:** Для построения модели было случайным образом выбрано 5 признаков (округляем sqrt(22) до 5).
  + **Преобразование признаков:** Категориальные признаки были преобразованы в числовые с помощью one-hot encoding (pd.get\_dummies).
  + **Преобразование целевой переменной:** Целевая переменная 'poisonous' преобразована в числовой формат: 'p' -> 1 (ядовитый), 'e' -> 0 (съедобный).
  + **Разделение на обучающую и тестовую выборки:** Данные были разделены на обучающую (80%) и тестовую (20%) выборки.

**2. Реализация алгоритма дерева решений**

* **Функция расчета примеси Джини (gini\_impurity):** Реализована функция для расчета примеси Джини, которая используется для определения качества разделения данных в узлах дерева.
* **Функция поиска лучшего разделения (find\_best\_split):** Реализована функция для поиска наилучшего признака и порога для разделения данных в узле дерева. Критерием выбора является максимальный прирост информации.
* **Функция разделения данных (split\_data):** Реализована функция для разделения данных на две ветви в соответствии с выбранным признаком и порогом.
* **Функция расчета прироста информации (information\_gain):** Реализована функция для расчета прироста информации при разделении данных.
* **Класс узла дерева (TreeNode):** Создан класс для представления узла дерева, содержащий информацию о признаке и пороге разделения, левом и правом поддереве, значении класса (для листовых узлов) и количестве объектов каждого класса в узле.
* **Функция построения дерева (build\_tree):** Реализована рекурсивная функция для построения дерева решений. Функция на каждой итерации ищет лучшее разделение данных и строит узлы дерева до тех пор, пока не будет достигнут критерий остановки (максимальная глубина дерева или нулевой прирост информации).

**3. Обучение и оценка модели**

* **Обучение:** Дерево решений было обучено на обучающей выборке с ограничением максимальной глубины равной 3.
* **Предсказание:** Были получены предсказания для тестовой выборки с помощью обученной модели.
* **Оценка качества:** Для оценки качества модели были рассчитаны следующие метрики:
  + **Accuracy:** Доля правильных ответов модели. В данной работе accuracy составила 99.13%, что свидетельствует о высоком качестве модели.
  + **Precision:** Доля правильно предсказанных положительных ответов (ядовитых грибов) от общего количества предсказанных положительных ответов. Precision = 1.0, то есть все грибы, которые модель определила как ядовитые, действительно являются таковыми.
  + **Recall:** Доля правильно предсказанных положительных ответов от общего количества реальных положительных ответов. Recall = 0.982, то есть модель правильно классифицировала 98.2% ядовитых грибов.

**4. Визуализация и интерпретация результатов**

* **Вывод структуры дерева (print\_tree):** Структура обученного дерева решений была выведена в консоль для визуализации процесса принятия решений моделью.
* **Вывод правил классификации (print\_rules):** Правила классификации, извлеченные из дерева решений, были выведены в консоль в формате "ЕСЛИ ... ТО ...", что облегчает интерпретацию модели.

**5. Построение ROC- и PR-кривых**

* **Функция вычисления вероятностей (predict\_proba):** Реализована функция для предсказания вероятности принадлежности объекта к классу 1 (ядовитый гриб).
* **Функции построения точек для ROC- и PR-кривых (calculate\_roc\_points, calculate\_pr\_points):** Реализованы функции для расчета пар значений FPR-TPR (для ROC-кривой) и Precision-Recall (для PR-кривой) при различных порогах вероятности.
* **Расчет AUC:** Для обеих кривых была рассчитана площадь под кривой (AUC), являющаяся агрегированной мерой качества классификатора (пример реальный значений, которые выдала программа):
  + AUC ROC: 0.9164
  + AUC PR: 0.4672

Такие значения может могут быть связаны со спецификой данных и чувствительностью этой метрики. В целом, полученные результаты свидетельствуют о хорошей способности модели классифицировать грибы на ядовитые и съедобные

* **Построение графиков:** ROC- и PR-кривые были построены для визуальной оценки качества модели при различных порогах вероятности.

Пример графиков:

A graph with a line and a point

Description automatically generated with medium confidence

A graph with a line

Description automatically generated

**Заключение**

В ходе лабораторной работы был успешно реализован алгоритм построения дерева решений для классификации грибов на ядовитые и съедобные.

**Основные результаты:**

* **Высокая точность классификации:** Модель продемонстрировала высокое значение accuracy (99.13%), что говорит о ее способности эффективно разделять грибы на классы.
* **Интерпретируемость модели:** Структура дерева решений и извлеченные правила классификации позволяют легко интерпретировать логику работы модели.
* **Анализ ROC- и PR-кривых:** Построение и анализ ROC- и PR-кривых позволило оценить качество модели при различных порогах вероятности и получить более полное представление о ее производительности.

В целом, результаты работы подтверждают эффективность алгоритма дерева решений для решения задач классификации и демонстрируют важность предобработки данных, выбора признаков и оценки качества модели для достижения наилучших результатов.