Практическая работа №10: «Деревья принятия решений».

Оглавление

Цель работы 1
Задачи работы 1
Перечень обеспечивающих средств 1
Общие теоретические сведения 1
Описание метода 1
Алгоритм 2
Задание 4
Требования к отчету 4
Литература 4

Цель работы

Получить практические навыки решения задач регрессии и классификации с помощью деревьев принятия решений.

Задачи работы

1. Сравнить несколько моделей для решения задачи регрессии с помощью деревьев принятия решений.
2. Сравнить несколько моделей для решения задачи классификации с помощью деревьев принятия решений.

Перечень обеспечивающих средств

1. ПК.
2. Учебно-методическая литература.
3. Задания для самостоятельного выполнения.

Общие теоретические сведения

Описание метода

Дерево решений – визуальная модель алгоритма принятия решения на основании набора правил. Правила формулируются с помощью отдельных параметров элементов из набора данных.

Внутренние узлы – задают правила выбора ветви.

Листья – содержат отдельные элементы набора данных, например, относящиеся к одному и тому же классу.

Набор данных в задаче обучения с учителем можно представить в виде дерева принятия решений, а затем использовать полученное дерево для решения задач классификации или регрессии для новых данных.

Количество деревьев, которые можно построить из имеющегося набора данных, велико, поэтому необходим некий алгоритм, позволяющий строить такие деревья эффективно.

Алгоритм

Общий алгоритм построения дерева:

1. Помещаем весь набор данных в первый узел.

2. Для каждого узла вычисляем значение некоторого параметра и, если оно не равно нулю, то ищем такое разбиение данных в узле, которое максимизирует среднее уменьшение этого значения.

3. Найденное разбиение сохраняем как правило в узле, а разбитые данные помещаем в два новых дочерних узла.

В качестве параметра разбиения может использоваться энтропия или коэффициент Джини.

Энтропия Шенонна: $S_{sh} = -\sum_{i=1}^{K} p_i \log_2 p_i$.

Коэффициент Джини (Gini impurity): $I_G = 1 - \sum_{i=1}^{K} p_i^2$.

Критерии остановки алгоритма:

- Ранняя остановка – остановка при достижении некоторой доли правильной классификации или другого критерия.

- Ограничение глубины – остановка при достижении заранее заданной максимальной длины ветвей.
- Минимальное количество элементов в листе – остановка при достижении заранее заданного числа элементов в каждом листе.

Задание

Пояснение

Для сохранения результатов данной работы вам понадобится файл ipynb. Если требуется, для удобства можно создать также второй файл формата doc/docx. Названия файла или файлов должны иметь вид «*Фамилия* – задание 10».

Часть 1

- Обновите свой репозиторий, созданный в практической работе №1, из оригинального репозитория: https://github.com/mosalov/Notebook_For_AI_Main.

Часть 2

- Откройте свой репозиторий в Binder (https://mybinder.org/).
- Откройте файл «2021 Весенний семестр\task3.ipynb».
- Изучите, при необходимости – выполните повторно, приведённый в файле код.
- Выполните два задания, приведённых в ячейках в конце ноутбука.
- Сохраните код в ipynb-файле. При необходимости пояснения опишите в doc/docx-файле.

Требования к отчету

Готовые файлы загрузите в свой репозиторий, созданный в практическом задании №1 по пути: «Notebook_For_AI_Main/2021 Весенний семестр/Практическое задание 3/», и сделайте пул-реквест.

Литература

1. https://habr.com/ru/company/ods/blog/322534/
2. https://habr.com/ru/post/171759/
3. https://habr.com/ru/post/116385/