МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ М. В. ЛОМОНОСОВА
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ
ИНТЕЛЛЕКТУАЛЬНЫЕ ТЕХНОЛОГИИ В МЕДИЦИНЕ

Основные задачи и методы анализа данных

СЕМЕСТРОВОЕ ЗАДАНИЕ

Выполнила:

Симакова Инна, 533 группа

Москва, 2023

Оглавление

1. Постановка задачи

Задача: классификация - задача разделения объектов на классы. Данная задача является примером задачи обучения с учителем (когда есть "ответы" для некоторых объектов). Будет рассмотрена задача классификации на несколько непересающихся классов.

Постановка задачи классфикации:

X - множество объектов, Y - множество ответов, $y : X -> Y$ - неизвестная зависимость.

Дано:

$x_1, ..., x_l \subset X$ - обучающая выборка,

$y_i = y(x_i), i = 1, ..., l$ - множество ответов.

Найти: $a : X -> Y$ - решающую функцию, приближающую y на всем множестве X.

Каждый объект для такой задачи описывается набором признаков (и по сути является вектором), будут рассматриваться номинальные признаки.

2. Данные

Датасет: Wine_dataset - винный набор данных.

Эти данные являются результатами химического анализа вин, выращенных в одном и том же регионе Италии тремя разными культиваторами. Было проведено тринадцать различных измерений, проведенных для различных компонентов, содержащихся в трех типах вина.

Объекты - вина (всего 178), которые разделены на 3 класса, количество признаков каждого объекта - 13.

Классы: класс_0 (59), класс_1 (71), класс_2 (48))

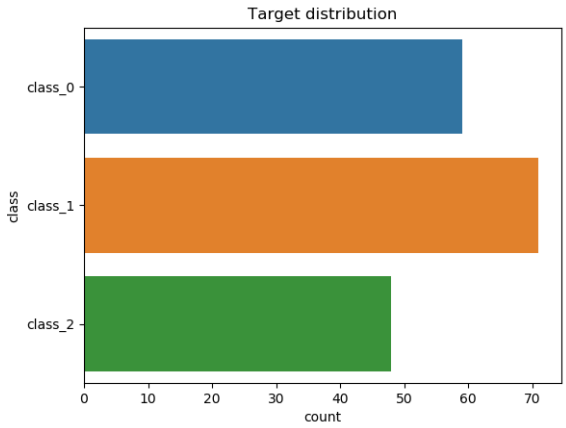

Рис. 1. Распределение классов.

Признаки:

Алкоголь, яблочная кислота, пепел, щелочность золы, магний, общие фенолы, флаваноиды, нефлаваноидные фенолы, проантоцианы, интенсивность цвета, оттенок, OD280 / OD315 разбавленных вин, пролин.

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04
...	...	...	...	...	...	...	...	...	...	...	...
173	13.71	5.65	2.45	20.5	95.0	1.68	0.61	0.52	1.06	7.70	0.64
174	13.40	3.91	2.48	23.0	102.0	1.80	0.75	0.43	1.41	7.30	0.70
175	13.27	4.28	2.26	20.0	120.0	1.59	0.69	0.43	1.35	10.20	0.59
176	13.17	2.59	2.37	20.0	120.0	1.65	0.68	0.53	1.46	9.30	0.60
177	14.13	4.10	2.74	24.5	96.0	2.05	0.76	0.56	1.35	9.20	0.61

Рис. 2. Примеры признаков и их значений.

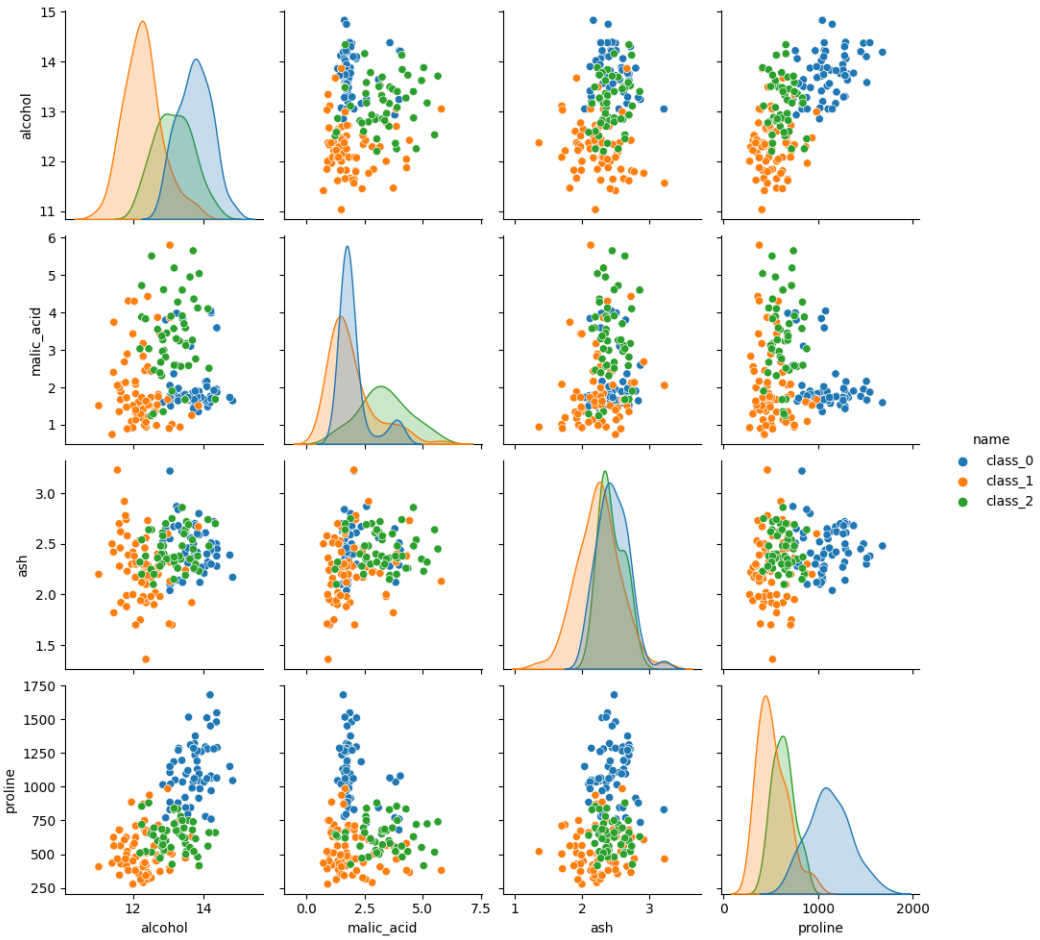

Рис. 3. Примеры распределения признаков для объектов разных классов.

Предобработка данных:

1) Была выполнена нормализация данных (см. рис.4).

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids
0	1.518613	-0.562250	0.232053	-1.169593	1.913905	0.808997	1.034819
1	0.246290	-0.499413	-0.827996	-2.490847	0.018145	0.568648	0.733629
2	0.196879	0.021231	1.109334	-0.268738	0.088358	0.808997	1.215533
3	1.691550	-0.346811	0.487926	-0.809251	0.930918	2.491446	1.466525
4	0.295700	0.227694	1.840403	0.451946	1.281985	0.808997	0.663351

Рис. 4. Примеры значений признаков после нормализации.

2) С помощью использования метода опорных векторов было произведено понижение размерности признакового пространства. Вместо 13 признаков в задаче рассматриваются 7 (количество важных признаков было получено с помощью GridSearchCV).

3. Модели

1 модель: Метод парзеновского окна (лекция 6)

Метод парзеновского окна — метрический метод классификации, в котором веса признаков задаются с помощью ядра K: $w(i,x) = K(\frac{\rho(x,x^{(i)}}{h})$.

В основе подхода лежит идея о том, что плотность выше в тех точках, рядом с которыми находится большое количество объектов выборки. Поэтому у близких точек веса выше.

Метрика ("расстояние между объектами") для данного метода:

$$a(x; X^l, h, K) = \arg\max_{y \in Y} \sum_{i=1}^{l} [y = y_i] K(\frac{\rho(x, x_i)}{h}),$$

где $K(r)$ - ядро, h - ширина окна.

2 модель: Решающее дерево (Decision Tree Classifier, лекция 8)

Решающее дерево предсказывает значение целевой переменной с помощью применения последовательности простых решающих правил (которые называются предикатами).

Решение о том, к какому классу будет отнесён текущий объект выборки, будет приниматься с помощью прохода от корня дерева к некоторому листу.

В каждом узле этого дерева находится предикат. Если предикат верен для текущего примера из выборки, мы переходим в правого потомка, если нет — в левого.

Пример дерева, полученного при реализации метода представлен в разделе результатов.

4. Результаты

Был произведен подбор гиперпараметров для моделей: число ближайших соседей для метода парзеновского окна равно 20, максимальная глубина решающего дерева - 2, критерий - энтропия. Значение ширины окна было выбрано эмпирически: 0.5.

Метод парзеновского окна нестабильно работал для 7 признаков и было принято решение сократить количество признаков до 2. Это не повлияло на сравнение методов, так как глубина решающего дерева равна 2 (что эквивалентно анализу 2 признаков).

Значения метрик для метода парзеновского окна:

	precision	recall	f1-score	support
0	0.86	1.00	0.93	19
1	1.00	0.91	0.95	22
2	1.00	0.92	0.96	13
accuracy			0.94	54
macro avg	0.95	0.94	0.95	54
weighted avg	0.95	0.94	0.95	54

Значения метрик для решающего дерева:

	precision	recall	f1-score	support
0	0.90	1.00	0.95	19
1	1.00	0.91	0.95	22
2	1.00	1.00	1.00	13
accuracy			0.96	54
macro avg	0.97	0.97	0.97	54
weighted avg	0.97	0.96	0.96	54

Визуализация результатов представлена ниже.

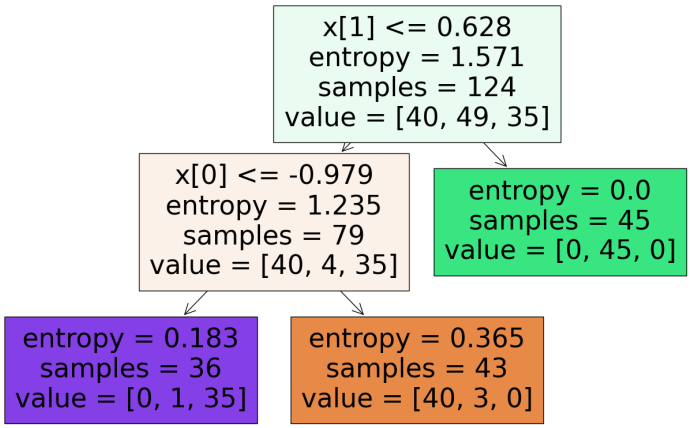

Рис. 5. Решающее дерево.

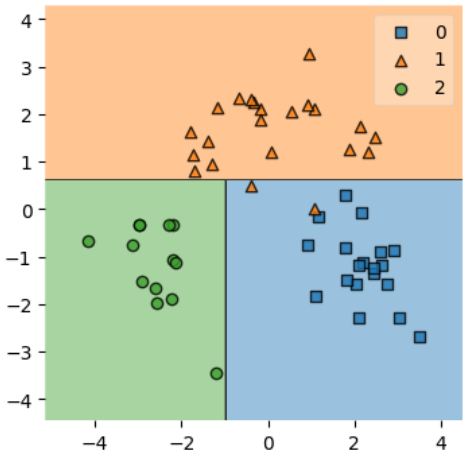

Рис. 6. Визуализация разделения по классам для решающего дерева.

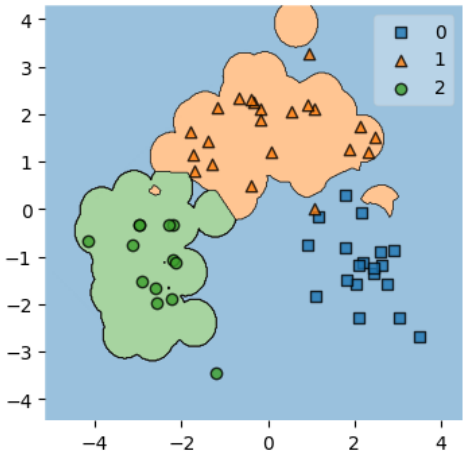

Рис. 7. Визуализация разделения по классам для парзеновского окна.

5. Выводы

По результатам данного иследования с решением поставленной задачи лучше справилась модель на основе решающего дерева.

Я считаю, что это связано, в первую очередь, с тем, что для данного датасета возможна классификация объектов по 2 признакам. Также возможными причинами могли стать:

- подбор параметра ширины окна эмпирически, а не с помощью более обоснованных методов;
- недостаточная оценка параметров для парзеновского окна (можно было исследовать другие параметры);
- случайность, связанная с маленьким размером выборки.