Задача дискретной классификации для Iris flower data set.

Баталов Семен

18.02.2021

1. Iris flower data set

Это стандартный набор данных, интегрированнный в модуль «**sklearn**» языка «**Python**» (Рис. 1). Набор данных состоит из 150 образцов каждого из трех видов ириса (Iris setosa, Iris virginica и Iris versicolor). Для каждого образца были измерены четыре характеристики: длина и ширина чашелистиков и лепестков в сантиметрах. Основываясь на комбинации этих четырех характеристик, можно разработать несложный классификатор, чтобы отличать виды друг от друга.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target	name
0	5.1	3.5	1.4	0.2	0	setosa
1	4.9	3.0	1.4	0.2	0	setosa
2	4.7	3.2	1.3	0.2	0	setosa
3	4.6	3.1	1.5	0.2	0	setosa
4	5.0	3.6	1.4	0.2	0	setosa
...	...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2	virginica
146	6.3	2.5	5.0	1.9	2	virginica
147	6.5	3.0	5.2	2.0	2	virginica
148	6.2	3.4	5.4	2.3	2	virginica
149	5.9	3.0	5.1	1.8	2	virginica

150 rows × 6 columns

Рис. 1. Набор данных трех видов ириса.

2. Классификатор

Классификатор был написан на языке «**Python**». Подробнее о программе можно узнать в папке «**source**» проекта.

При построении дерева (Рис. 2) использовался стандартный классификатор «**DecisionTreeClassifier**» модуля «**sklearn**». Обучающая выборка составила 50% от всех полей (переставленных в случайном порядке) в наборе.

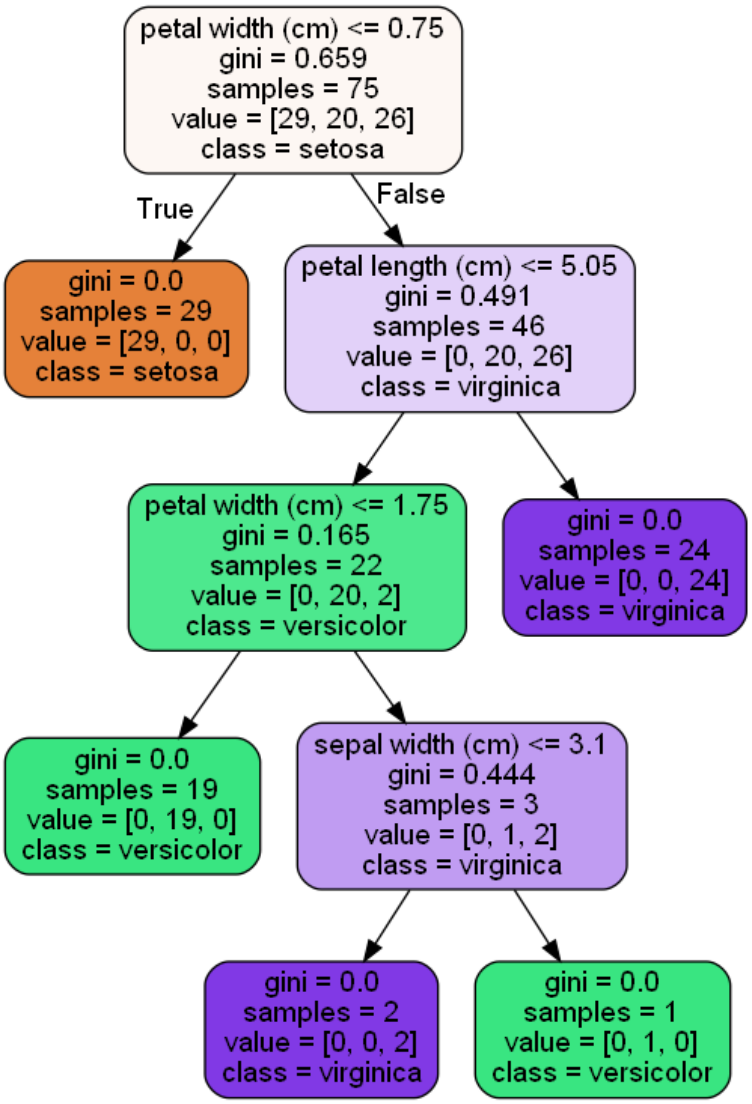

Рис. 2. Дерево классификации.

После была произведена проверка работоспособности классификатора на оставшихся в датасете примерах. Результат проверки изображен на рисунке (Рис. 3).

- **accuracy** – это главная метрика, которая показывает долю правильных ответов модели. Ее значение равно отношению числа правильных ответов, которые дала модель, к числу всех объектов. Но она не полностью отражает качество модели. Поэтому вводятся **precision** и **recall**.

- **precision** – эта метрика показывает, насколько мы можем доверять модели, другими словами, какое у нас количество «ложных срабатываний». Значение метрики равно отношению числа ответов, которые модель считает правильными, и они действительно были правильными (это число обозначается «true positives») к сумме «true positives» и числа объектов которые модель посчитала правильными, а на самом деле они были неправильные (это число обозначается «false positives»). В виде формулы: precision = «true positives» / («true positives» + «false positives»).

- **recall** – эта метрика показывает насколько модель может вообще обнаруживать правильные ответы, другими словами, какое у нас количество «ложных пропусков». Ее численное значение равно отношению ответов, которые модель считает правильными, и они действительно были правильными к числу всех правильных ответов в выборке. В виде формулы: recall = «true positives» / «all positives».
- **f1-score** – это объединение **precision** и **recall**.
- **support** – это просто число найденных объектов в классе.

Accuracy: 0.96

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	21
versicolor	0.94	0.97	0.95	30
virginica	0.96	0.92	0.94	24
accuracy			0.96	75
macro avg	0.96	0.96	0.96	75
weighted avg	0.96	0.96	0.96	75

Рис. 3. Оценка работы классификатора.