Classification in the Iris data set
A test example for `Jupyter nbconvert`

Dylan P. Tweed

See Linkedin profile for affiliation

February 8, 2017

Introduction

The iris dataset is common test example for machine learning and can be found in the `datasets` packages of `R` or as in this instance the `sklearn` package in `python`. This data set was first published in [Fisher, 1936], in was further use for the purpose of testing machine learning classification algorithm such as in [Ro and Pe, 1973], [Dasarathy, 1980].

Data Set Characteristics

1. Number of Instances: 150 (50 in each of three classes)
2. Number of Attributes: 4 numeric, predictive attributes and the class
    - sepal length in cm
    - sepal width in cm
    - petal length in cm
    - petal width in cm
3. class:
    - Iris-Setosa
    - Iris-Versicolour
    - Iris-Virginica

Classification targets

```
>>> # This should appear everywhere
... Counter(target)
Counter({0: 50, 1: 50, 2: 50})
```

```
>>> # This should appear everywhere
... list(target_names)
['setosa', 'versicolor', 'virginica']
```

Classification targets

Beware

The 3 class are indicated in the data as integers 0, 1 and 2:

```
>>> # This should appear everywhere
... Counter(target)
Counter({0: 50, 1: 50, 2: 50})
```

```
>>> # This should appear everywhere
... list(target_names)
['setosa', 'versicolor', 'virginica']
```

Classification targets

Beware

The 3 class are indicated in the data as integers 0, 1 and 2:

```
>>> # This should appear everywhere
... Counter(target)
Counter({0: 50, 1: 50, 2: 50})
```

But

With the corresponding class names:

```
>>> # This should appear everywhere
... list(target_names)
['setosa', 'versicolor', 'virginica']
```

In the next slides

We explore the first few element of the iris data set for each class:

- setosa encoded as 0 (see Table silde 9),
- versicolor encoded as 1 (see Table silde 10)
- virginica encoded as 2 (see Table silde 12).

In the next slides

We explore the first few element of the iris data set for each class:

- setosa encoded as 0 (see Table silde 9),
- versicolor encoded as 1 (see Table silde 10)
- virginica encoded as 2 (see Table silde 12).

We note that the row are ordered by class. This is not important here, since we try to test reference to some tables but for machine learning tasks it is advised to shuffle the row both in the data and the target.

	sepal length	sepal width	petal length	petal width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4
6	4.6	3.4	1.4	0.3
7	5.0	3.4	1.5	0.2
8	4.4	2.9	1.4	0.2
9	4.9	3.1	1.5	0.1

Table: First ten rows corredsponding to the Setosa class

	sepal length	sepal width	petal length	petal width
50	7.0	3.2	4.7	1.4
51	6.4	3.2	4.5	1.5
52	6.9	3.1	4.9	1.5
53	5.5	2.3	4.0	1.3
54	6.5	2.8	4.6	1.5
55	5.7	2.8	4.5	1.3
56	6.3	3.3	4.7	1.6
57	4.9	2.4	3.3	1.0
58	6.6	2.9	4.6	1.3
59	5.2	2.7	3.9	1.4

Table: First ten rows corresponding to the Versicolor class

This text is a Lorem Ipsum, it should not appear in the documentation template, and should add a lorem ipsum in the chapter and article template. It should also appear in the beamer, to test the animation on the table above.

	sepal length	sepal width	petal length	petal width
100	6.3	3.3	6.0	2.5
101	5.8	2.7	5.1	1.9
102	7.1	3.0	5.9	2.1
103	6.3	2.9	5.6	1.8
104	6.5	3.0	5.8	2.2
105	7.6	3.0	6.6	2.1
106	4.9	2.5	4.5	1.7
107	7.3	2.9	6.3	1.8
108	6.7	2.5	5.8	1.8
109	7.2	3.6	6.1	2.5

Table: First ten rows corresponding to the Virginica class

This text is a Lorem Ipsum, it should not appear in the documentation template, and should add a lorem ipsum in the chapter and article template. It should also appear in the beamer, to test the animation on the table above.

Distribution of the different classes

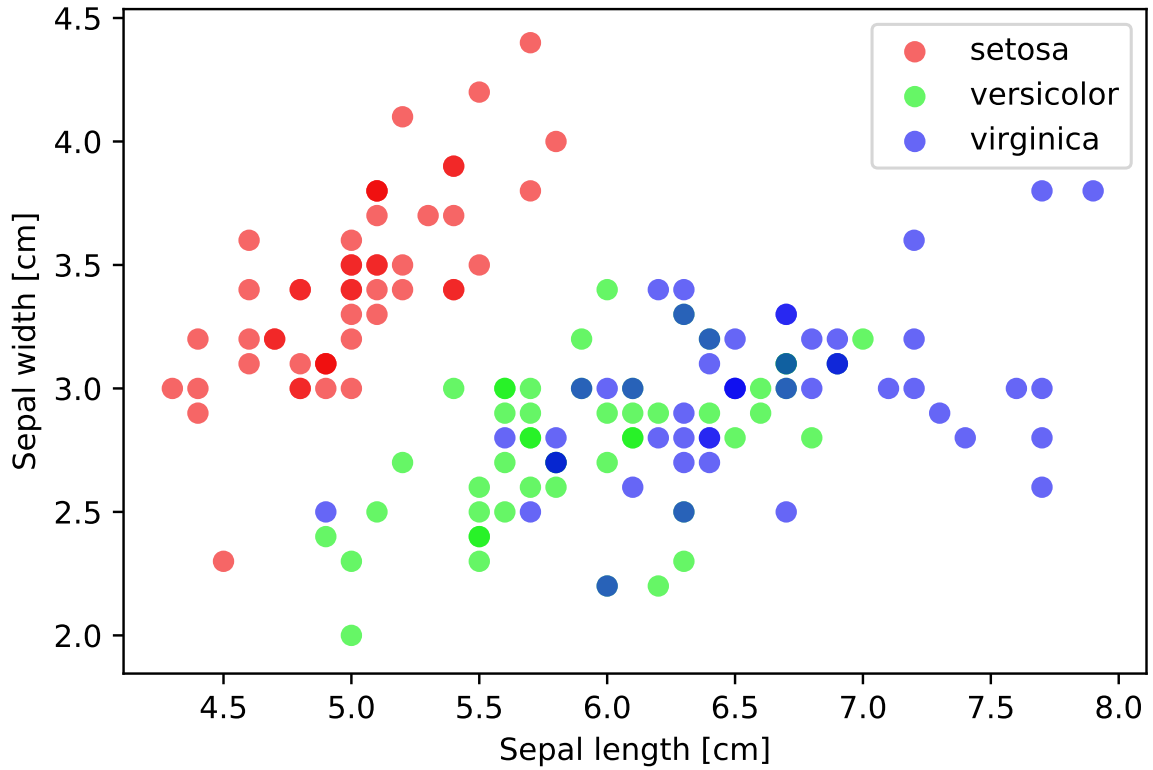


Figure: Scatter plot sepal width as a function of the sepal lenght for the iris dataset. As the legend indicates, the color code corresponds to the class.

Support Vector Classification models

For fun were testing different classification models for the iris dataset using the Support Vector Classification (SVC) method. This exemple is taken from the`sklearn` documantation. We test the SVC methods with:

- a linear kernel (see Figure silde 16)
- a Radial Basis Function kernel (RBF, see Figure silde 17)
- a degree 3 polynomial kernel (see Figure silde 18)

This text is a Lorem Ipsum, it should not appear in the documentation template, and should add a lorem ipsum in the chapter and article template. It should also appear in the beamer, to test the animation on the table above.

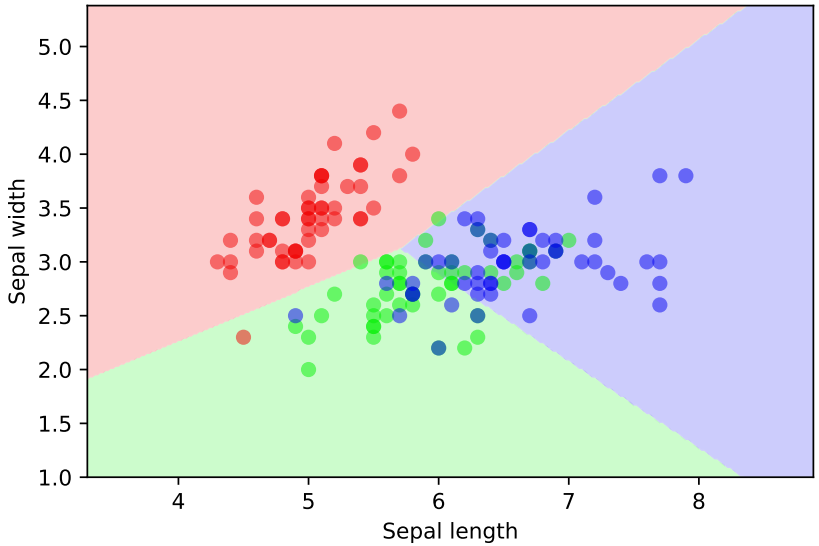


Figure: Same as Figure silde 13. The shaded region correspond to the predictions of Linear SVC model.

This text is a Lorem Ipsum, it should not appear in the documentation template, and should add a lorem ipsum in the chapter and article template. It should also appear in the beamer, to test the animation on the table above.

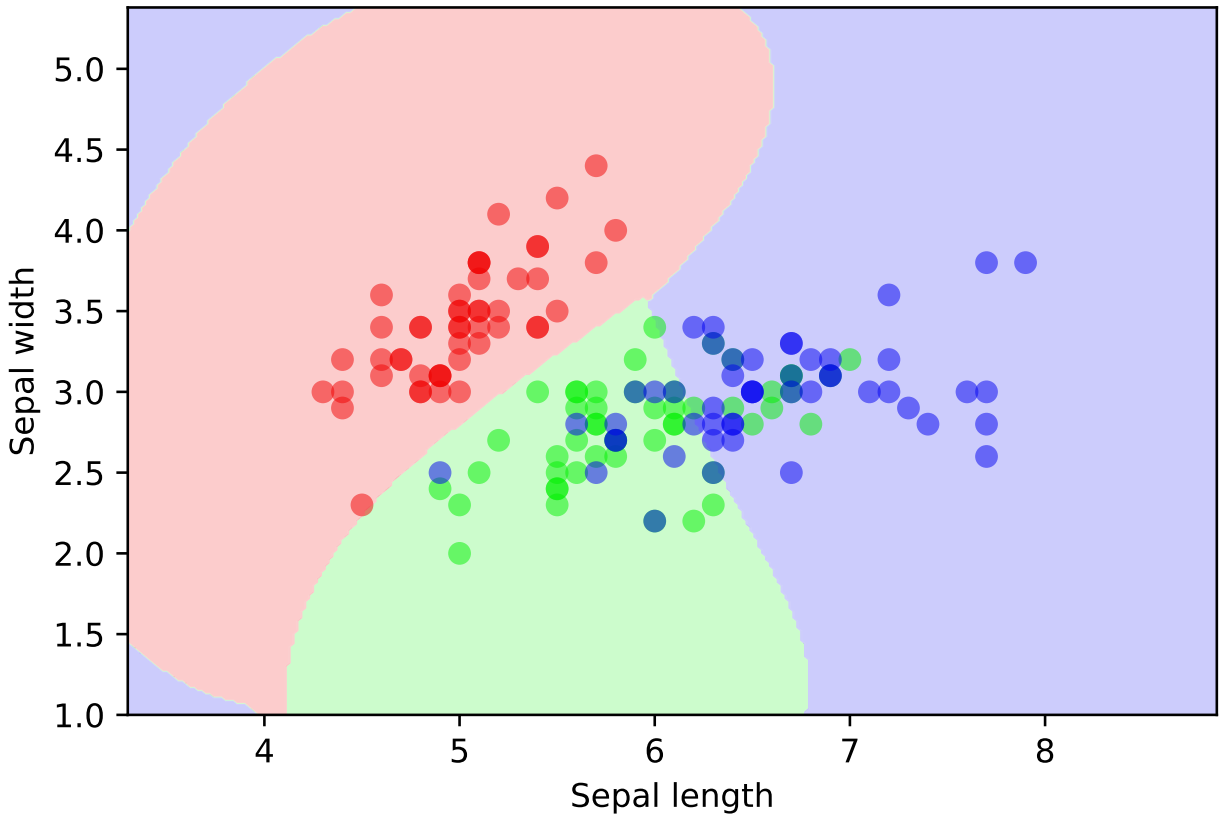
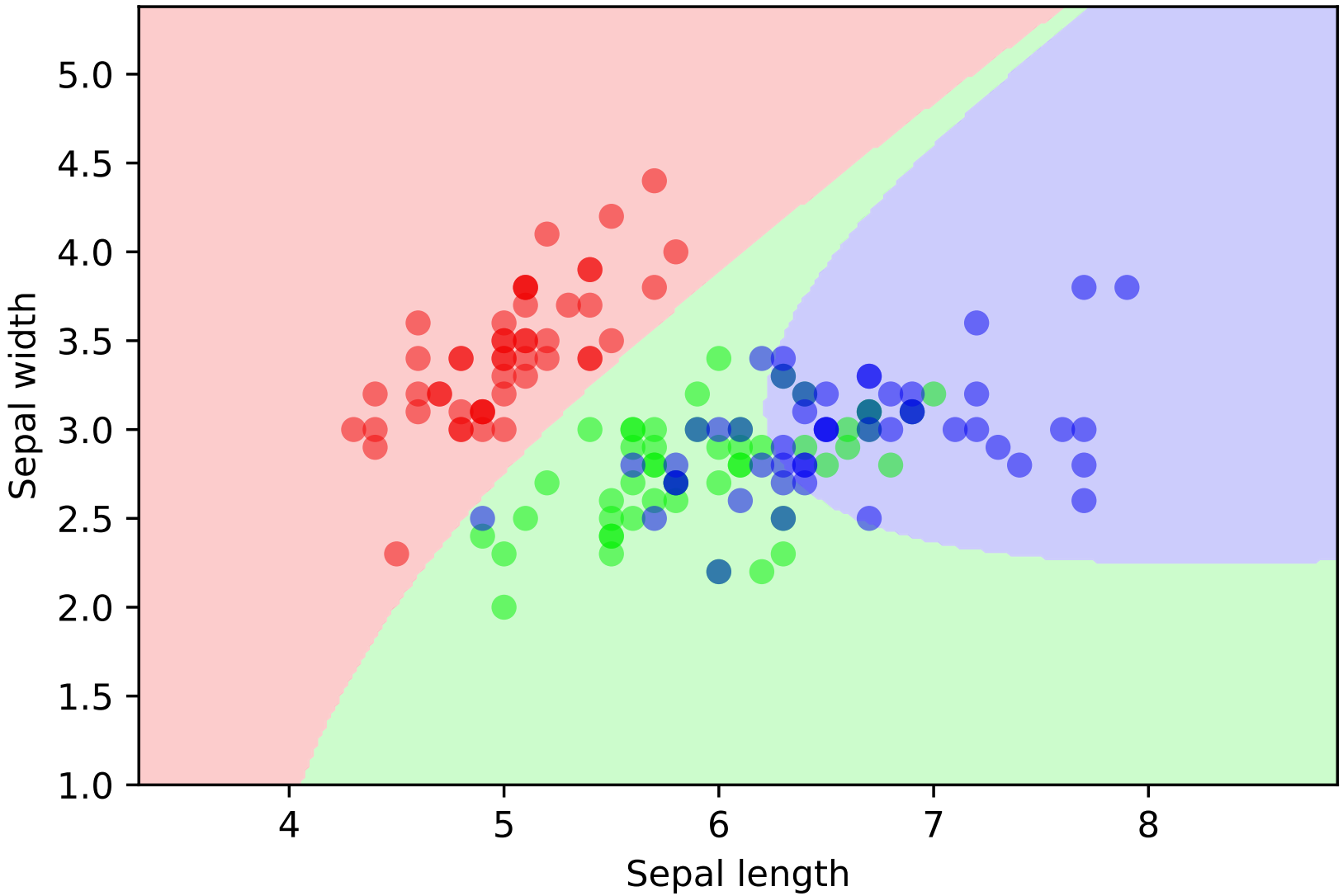


Figure: Same as Figure silde 13. The shaded region correspond to the predictions of SVC RBF model.


Sepal width
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8
Sepal length

bibliography


Dasarathy, B. V. (1980).
Nosing around the neighborhood: A new system structure and classification rule for recognition in partially exposed environments.
IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-2(1):67–71.


Fisher, R. A. (1936).
The use of multiple measurements in taxonomic problems.
Annals of Eugenics, 7(2):179–188.


Ro, D. and Pe, H. (1973).
Pattern Classification and Scene Analysis.
Wiley.