A01706648 Wenguang Hu

For this dataset, I tried to classification the variety of the Iris based on the Sepal Length/Width and Petal Length/Width.

For the Jupyter Notebook.

1, I load the Iris data set

We have cover the original dataset to an csv data set, to be easier to load.

2, I check the data set, review the columns and data type of the dataset

In order to see what column we have, and what columns is x, what columns is y

3, I make the Sepal Length/Width and Petal Length/Width in to a Feature Vector

Because it is a model to predict the class of Iris, so we use the variety as y, and all the Sepal and Petal size as x, we treat x as Feature.

After make them as the Feature-Label Vector, it will be easier for Spark to handle.

4, I make index of the label of Variety

To turn the String Class to the number index, it will be easier to handled by Spark. Because all will be numbers.

5, I combine the Feature Vector with the indexed Label, as a dataset for Training and Testing

For the dataset, we shell use a Vector style for Spark to handle.

And to split the dataset to Training and Testing, to isolate the Training and Testing dataset. Do not let the Testing dataset to affect the training.

6, I divided the Dataset as 75% for training and 25% for testing

Take major part of dataset for training, in order to make module more robuster.

7, I use the Logical Regression to train the model with the training dataset

Since we are doing the classification model, so the Logical Regression will be better than linear regression.

8, using K folder method to do the cross validation, k = 3

Use cross validation to make the model more general. Using K folder to divided the training and validation data, to make the model with lower bias, and make the model less overfit.

Since we do not have a big data set, so we use k = 3 to make the training and validation data set.

9, I test the Test Dataset with the model, and get the accuracy value.

A01706648 Wenguang Hu

In the end, we need to use the trained model to predict the test dataset, to check how goo our model perform.

For the Dashboard

I used Tableau to perform the work.

1, I load the data to Tableau

2, I create a WorkSheet to display the Variety relationship between Sepal Length and Sepal Width.

- 2.1 I put the Petal Length as dimension in column part
- 2.2 I put the Petal Width as dimension in row part
- 2.3 I put the Variety in Mark part, and choose the color for marking
- 2.4 I choose ShowMe as scatter plots

3, I create a WorkSheet to display the Variety relationship between Petal Length and Petal Width.

- 3.1 I put the Petal Length as dimension in column part
- 3.2 I put the Petal Width as dimension in row part
- 3.3 I put the Variety in Mark part, and choose the color for marking
- 3.4 I choose ShowMe as scatter plots

4, I create a WorkSheet to display the counter of each Variety, to check if the dataset has some bias or not.

- 4.1 I put the Count of Variety in Column Part
- 4.2 I put Variety in Row Part
- 4.3 I put the Variety in Mark part, and choose the color for marking
- 4.4 I choose ShowMe as Horizontal Bar

5, Finally, I create a Dash board to Combined all the above WorkSheets

- 5.1 I add Horizontal Widget to the Dash Board first
- 5.2 I choose the Petal and Sepal sheet, and added them to the Horizontal Widget
- 5.3 I add Vertical Widget to the Dash Board then
- 5.4 I add the Count Variety sheet to the Vertical Widget
- 5.5 I add another Horizontal Widget to the Dash Board

5.6 I add the selection of the variety to the Horizontal Widget

5.7 I add a Vertical Widget to the Dash Board

5.8 I add a Text to the Vertical Widget, and enter the Title of the Dash Board.

Done