

Python/Numpy basics

Dataset download and learn about characteristics

1. Download wheat seeds Dataset from [here](#). Read the description of the dataset.
2. Read the dataset.
3. Keep first, second, and third dimensions only (area, perimeter, compactness). You should also keep the last column because it contains class label.
4. Plot the data points using the first three dimensions (area, perimeter, compactness) in 3d plots. Use three different shapes (triangle, square, circle) to plot data points for three different classes. You should use the class information from class label and use them when you decide on shapes. For shape you need only the class information.
5. Calculate the mean data point for each class and show them with similar shape with the larger size. For shape you need only the class information.
6. Plot histogram for each of the two dimensions (area, perimeter). One plot for each dimension.
7. Calculate standard deviation (sigma) for each of the two dimensions (area, perimeter). You can use library function to calculate standard deviation.
8. Draw two separate box plot for all data points for two dimensions (area, perimeter). X axis have dimensions such as area or perimeter. The Y axis will have the box plots. (See slide 7, lecture 1)
9. Now, draw box plots for each class separately. In each figure you have three boxes for each two dimensions (area, perimeter). but data will be from a particular class.
10. Make matrix plots for the entire dataset for 2 dimensions (see slide 21, lecture 1). So The whole figure will have 2x2 scatter matrices.
11. Draw parallel coordinates for the entire dataset for two dimensions (area, perimeter). You will have 2 equidistant axes that are parallel. Show the axes units and axes names. Visually observe the mean and std for each dimension and match with your previous calculation.

Some preliminary information about wheat seed dataset.

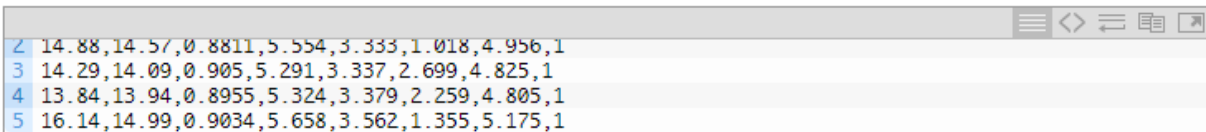
1. Link:
 - a. http://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt
2. More information:
 - a. <https://machinelearningmastery.com/standard-machine-learning-datasets/>

The Wheat Seeds Dataset involves the prediction of species given measurements of seeds from different varieties of wheat.

It is a multiclass (3-class) classification problem. The number of observations for each class is balanced. There are 210 observations with 7 input variables and 1 output variable. The variable names are as follows:

1. Area.
2. Perimeter.
3. Compactness
4. Length of kernel.
5. Width of kernel.
6. Asymmetry coefficient.
7. Length of kernel groove.
8. Class (1, 2, 3).

Snapshot of the dataset



2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
3	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
4	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
5	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1