

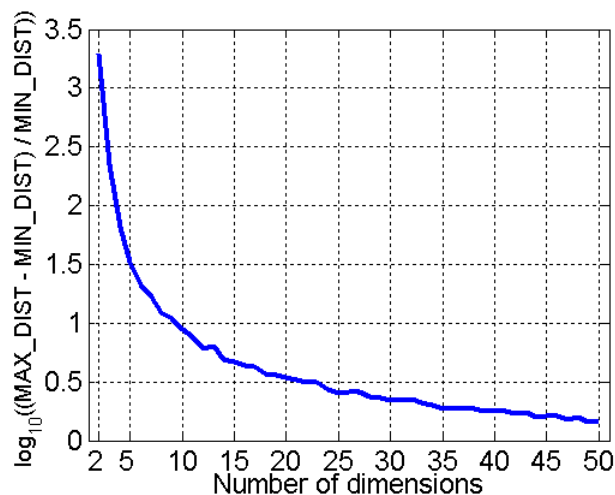
Due Date: See Webcampus
How to submit: Webcampus

General Guidelines:

- Please prepare a **typed** report that describes what you did. The report should be as concise as possible while providing all necessary information required to replicate your plots.
- For each problem, please provide, at the end of your report, a commented version of your python code files. **Python Notebook files are preferred. You may put the codes for all the problems in a SINGLE ipynb file with necessary texts to separate each problem.**

P1-1. Curse of Dimensionality.

Reproduce a figure similar to the figure in slide 37 in Chapter 2, i.e.,



(a) Generate 1000 points following a uniform distribution under a given dimension, and then compute difference between max and min distance between any pair of points. *Hint: Refer to the tutorial “Introduction to Numpy and Pandas” on how to generate random points.*

(b) Repeat (a) for different dimensions from 2 to 50.

Plot $\log_{10} \frac{\text{max}-\text{min}}{\text{min}}$ under different number of dimensions.

P1-2. The Iris Dataset (https://en.wikipedia.org/wiki/Iris_flower_data_set)

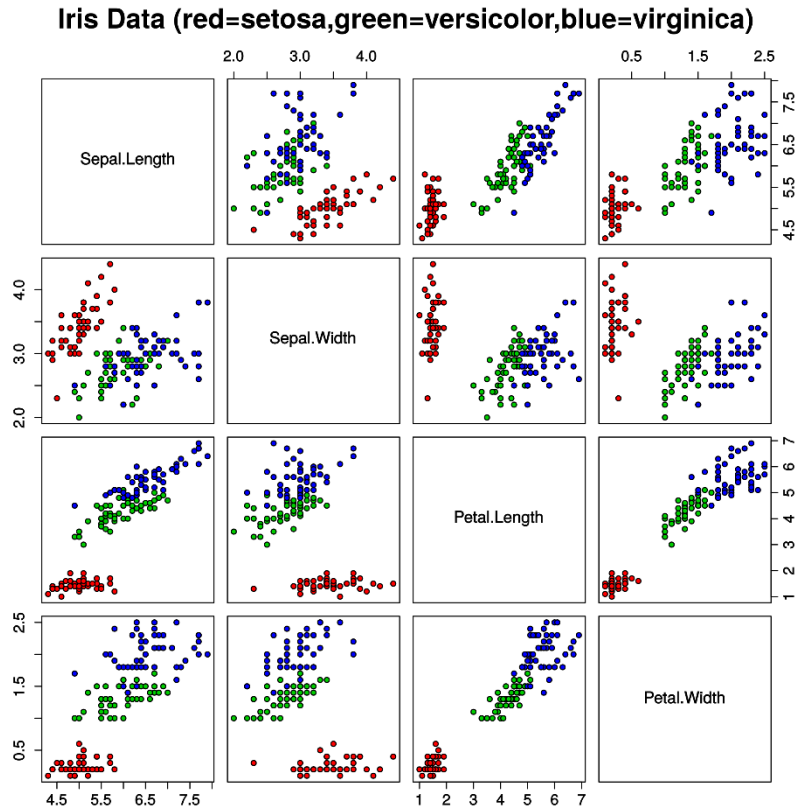
The Iris dataset is embedded in scikit-learn. You can install scikit-learn by following the instructions (<https://scikit-learn.org/stable/install.html>). Then you can load the Iris dataset using the following codes:

```
from sklearn import datasets
iris = datasets.load_iris()
```

The Iris dataset consists of 3 different types of irises’ (Setosa, Versicolour, and Virginica) petal and sepal length, stored in a 150x4 numpy.ndarray.

Tasks:

a) Data Visualization. Duplicate the following figure using scatter plot.



b) Find the best discretization for the petal length that can best separate the Iris data and plot a figure similar to the figure in slide 54 in Chapter 2. For each flower type, list in a table how many data samples are correctly separated and how many are not correctly separated.

P1-3. Principal Component Analysis for The Iris Dataset

You can use PCA embedded in scikit-learn by the following code:

```
from sklearn.decomposition import PCA
```

Tasks:

a) Use the Iris dataset and plot all the samples in a figure using Sepal Length and Sepal Width, i.e., `xlabel('Sepal length')` and `ylabel('Sepal width')`.

b) The Iris dataset has 4 attributes (sepal length, sepal width, petal length, and petal width). Use PCA to reduce the dimension of the dataset from 4 to 2. Plot all the samples after the dimensionality reduction in a 2D figure. Compare this figure with the figure in (a) and discuss whether you can better separate the data samples after the dimensionality reduction.