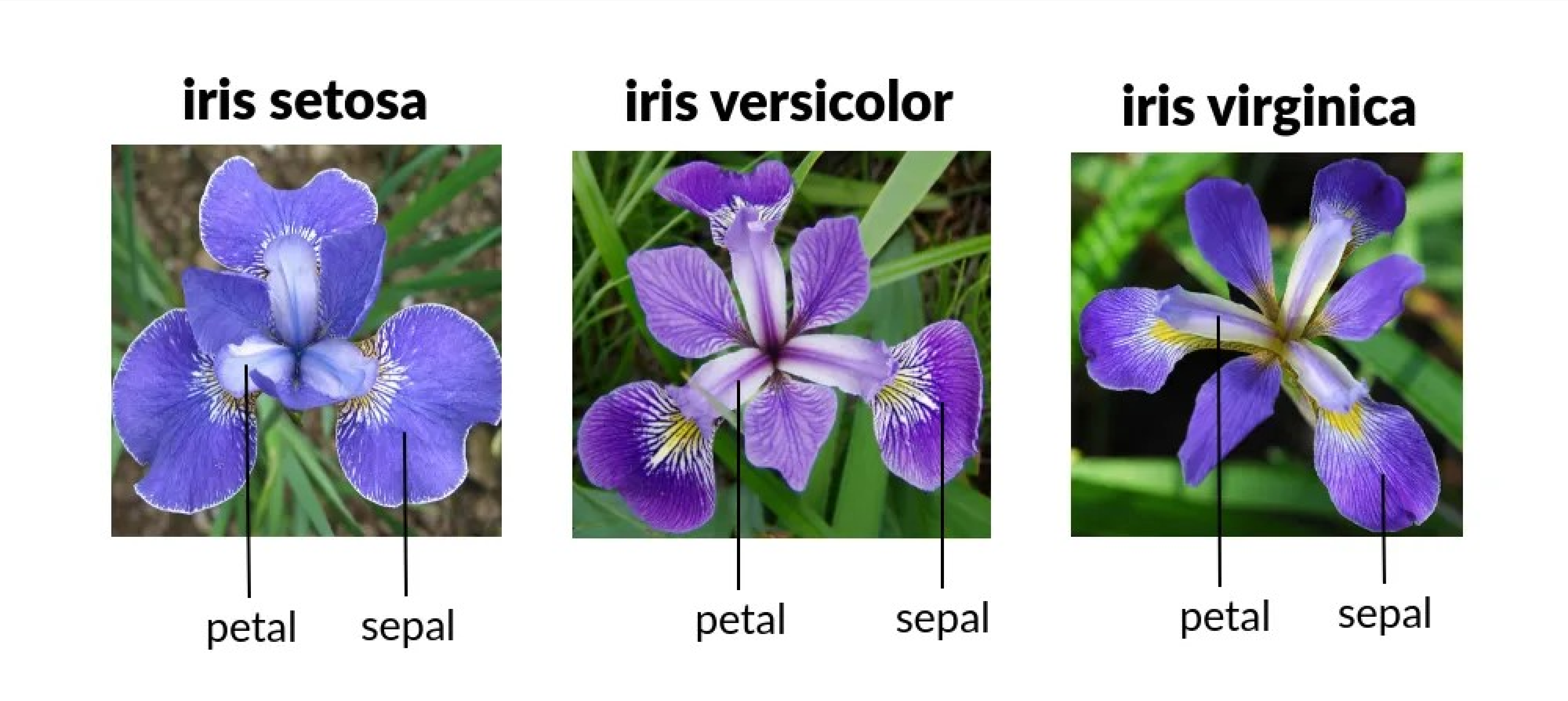
Exploratory Data Analysis: Iris dataset

Dr. Partha Pratim Sarangi
School of Computer Engineering

Iris Dataset



Data Summary:

- Number of Samples: 150
- Number of Features: 4
- Target Classes: Setosa, Versicolor, Virginica

Organization of Dataset

The data matrix refers to the array of numbers

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

where x_{ij} is the j-th variable collected from the i-th item (e.g., subject).

- items/subjects are rows
- variables are columns

$\mathbf{X}$ is a data matrix of order $n \times p$ (# items by # variables).

Collection of Column Vectors

We can view a data matrix as a collection of column vectors:

$$\mathbf{X} = \begin{pmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \\ | & | & & | \end{pmatrix}$$

where $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$ for $j \in \{1, \ldots, p\}$.

The $n \times 1$ vector $\mathbf{x}_j$ gives the j-th variable's scores for the n items.

Collection of row vectors

We can view a data matrix as a collection of row vectors:

$$\mathbf{X} = \begin{pmatrix} — & \mathbf{x}_1' & — \\ — & \mathbf{x}_2' & — \\ & \vdots & \\ — & \mathbf{x}_n' & — \end{pmatrix}$$

where $\mathbf{x}_i'$ is the i-th row of $\mathbf{X}$ for $i \in \{1, \ldots, n\}$.

The $1 \times p$ vector $\mathbf{x}_i'$ gives the i-th item's scores for the p variables.

Exploratory Data Analysis: Iris dataset

- To read Iris dataset (iris.csv) the easiest way is to use data frame of pandas library in python.
- Import following libraries of python:
  - import os
  - import pandas as pd
  - import numpy as np
  - from matplotlib import pyplot as plt
  - Import seaborn as sns
- Keep your python program and dataset (iris.csv) in a same folder.
- Load the Iris dataset from a local drive.
- data_path = os.getcwd()
- data = pd.read_csv(os.path.join(data_path, 'iris.csv'))

- #Add header in the dataset
- data.columns = [‘Sepal_Length’, ‘Sepal_Width’, ‘Petal_Length’, ‘Petal_Width’, ‘Species’]
- #Gaining information from data
- iris_df.info()
- #We need to know the overall statistical information of the dataset
- iris_df.describe()
- Checking the distribution of each species in the dataset
- iris_df[‘Species’].value_counts()
- sns.countplot(iris_df[‘Species’])
- plt.title(‘Species Count’)

Covariance of Dataset

The covariance matrix refers to the symmetric array of numbers

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1p} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2p} \\ s_{31} & s_{32} & s_3^2 & \cdots & s_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & s_{p3} & \cdots & s_p^2 \end{pmatrix}$$

where

- $s_j^2 = (1/n)\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ is the variance of the j-th variable
- $s_{jk} = (1/n)\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$ is the covariance between the j-th and k-th variables
- $\bar{x}_j = (1/n)\sum_{i=1}^n x_{ij}$ is the mean of the j-th variable

Correlation of Dataset

The correlation matrix refers to the symmetric array of numbers

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{pmatrix}$$

where

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

is the Pearson correlation coefficient between variables $\mathbf{x}_j$ and $\mathbf{x}_k$.