

## HW1 CSC 767 – Neural Networks & Deep Learning

### Descriptive Statistics and Analysis Using Python and Python Libraries

#### Description of Iris flower data set

- Attributes are numeric (float values) and they are:  
sepal - length  
sepal - width  
petal - length  
petal - width
  - The classes are: Iris Setosa, Iris Versicolor and Iris Virginica
  - The number of instances of each class is 50
- All the attribute values are in the same units (cm) and same scale as well.

#### Listing 1a: Load libraries

```
In [1]: 1 # Load libraries
2 from pandas import read_csv
3 from pandas.plotting import scatter_matrix
4 from matplotlib import pyplot
5 from sklearn.model_selection import train_test_split
6 from sklearn.model_selection import KFold
7 from sklearn.model_selection import cross_val_score
8 from sklearn.metrics import classification_report
9 from sklearn.metrics import confusion_matrix
10 from sklearn.metrics import accuracy_score
11 from sklearn.linear_model import LogisticRegression
12 from sklearn.tree import DecisionTreeClassifier
13 from sklearn.neighbors import KNeighborsClassifier
14 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
15 from sklearn.naive_bayes import GaussianNB
16 from sklearn.svm import SVC
```

```
In [ ]: 1
```

#### Listing 1.b: Load the Iris dataset

```
In [4]: 1 # Load dataset
2 filename = 'iris.data.csv'
3 names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
4 dataset = read_csv(filename, names=names, delimiter = '\t')
```

```
In [ ]: 1
```

**Listing 2** – Dimensions of the dataset. Peek at the data itself. Statistical summary of all attributes. Breakdown of the data by the class variable.

**a) Print the shape of the data-set.**

```
In [6]: 1 # Summarize Data
        2
        3 # Descriptive statistics
        4 # shape
        5 print(dataset.shape)
```

(150, 5)

```
In [ ]: 1
```

**b) Print the first few rows of the data-set.**

```
In [2]: 1 # head
        2 print(dataset.head(20))
```

	sepal-length	sepal-width	petal-length	petal-width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa

**c) Print the statistical descriptions of the data-set.**

In [3]: `print(dataset.describe())`

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

d) Print the class distribution in the data-set.

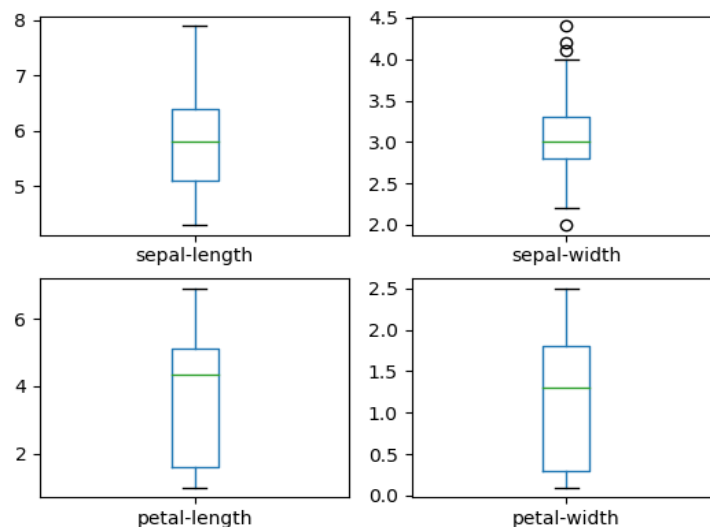
In [4]: `# class distribution`  
`print(dataset.groupby('class').size())`

```
class
Iris-setosa      50
Iris-versicolor  50
Iris-virginica   50
dtype: int64
```

**Listing 3.** Univariate plots to better understand each attribute. Multivariate plots to better understand the relationships between attributes.

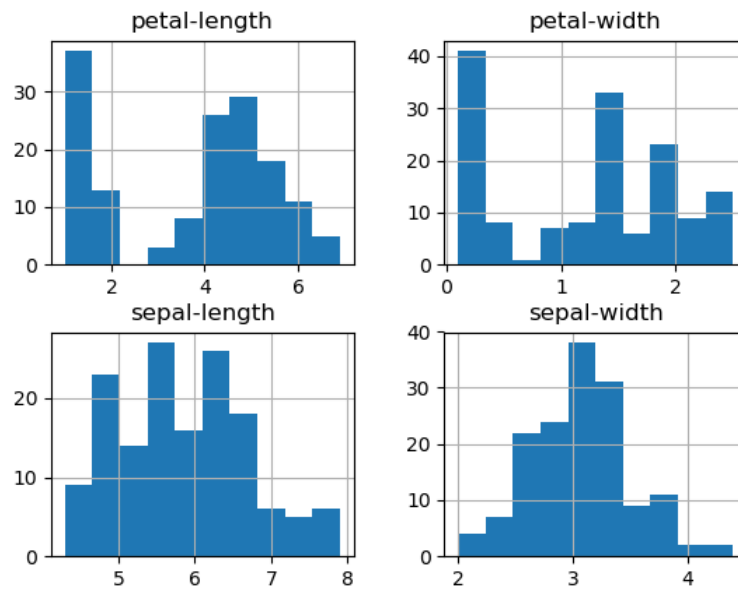
a) Univariate plot.

In [5]: `# box and whisker plots`  
`dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)`  
`pyplot.show()`



b) Visualize the data-set using histogram plots.

```
In [6]: 1 # histograms
        2 dataset.hist()
        3 pyplot.show()
```



c) Visualize the dataset using scatter plots.

```
In [7]: 1 # scatter plot matrix
        2 scatter_matrix(dataset)
        3 pyplot.show()
```

