SIT 720 - Machine Learning

Lecturer: Chandan Karmakar | karmakar@deakin.edu.au

School of Information Technology,
Deakin University, VIC 3125, Australia.

Assessment Task 2 (30 marks)

Submission Instruction

1. Student should insert Python code or text responses into the cell followed by the question.
2. For answers regarding discussion or explanation, **maximum five sentences are suggested**.
3. Rename this notebook file appending your student ID. For example, for student ID 1234, the submitted file name should be A2_1234.ipynb.
4. Insert your student ID and name in the following cell.

```
# Student ID:

# Student name:
```

Part 1: Clustering *(15 marks)*

Let's assume you want to design an environment to predict a class/category from a dataset based on specific features of that class. However, all the features are not strong enough or in other words features not that much variance/uniqueness across the classes. So, you have to design a clustering model by answering the following questions:

1. Download the attached clustering.csv file. Read the file and separate the class and feature matrix. **(2 marks)**

```
# INSERT your code (or comment) here
```

2. Determine the number of clusters from the dataset. Is this same as the actual number of classes in the dataset? **(1 marks)**

```
# INSERT your code (or comment) here
```

3. Perform K-Means clustering on the complete dataset and report purity score. **(2 marks)**

```
# INSERT your code (or comment) here
```

4. There are several distance metrics for K-Means such as euclidean, squared euclidean, Manhattan, Chebyshev, Minkowski. [**Hints:** See the pyclustering library for python.]
    - Your job is to compare the purity score of k-means clustering for different distance metrics. **(5 marks)**
    - Select the best distance metric and explain why this distance metric is best for the given dataset. **(2 marks)**

```
# INSERT your code (or comment) here
```

5. Use selection criteria (ANOVA, Chi-squared) to select best three features and use them for K-Means clustering. Based on the purity score which feature set are you going to recommend and why? **(3 marks)**

```
# INSERT your code (or comment) here
```

Part-2 (Dimensionality Reduction using PCA/SVD) *(15 marks)*

1. For the dataset (clustering.csv), perform PCA.
    - plot the captured variance with respect to increasing latent dimensionality. **(2.5 marks)**

    What is the minimum dimension that captures:
    - at least 89% variance? **(1.5 marks)**
    - at least 99% variance? **(1 marks)**

```
# INSERT your code (or comment) here
```

2. Determine the purity of clusters formed by the number of principal components which captured 89% and 99% variances respectively. Plot a line graph of the purity scores against the captured variances. Discuss your findings. **(7 marks)**

```
# INSERT your code (or comment) here
```

3. Let's assume you have two datasets one is linear and another is curved structural data.
    - Can we apply PCA on these datasets? Justify your answer. **(3 marks)**

```
# INSERT your code (or comment) here
```