Subject: 21DS602

Lab Session: 03

Lab Session Date: 02/11/2022

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Report should be submitted to TurnItIn. Once done, submit your assignments in Teams.

Main Section (Mandatory):

Please use the data associated with your own project.

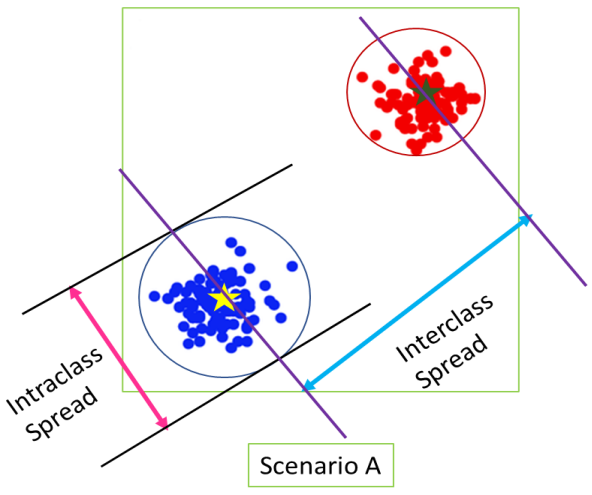
For dot product → use `numpy.dot()`
For Vector length → use `numpy.linalg.norm()`

Refer to lecture portions on k-NN. Also refer:

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

A1. Evaluate the intraclass spread and interclass distances between the classes in your dataset. If your data deals with multiple classes, you can take any two classes. Steps below (refer below diagram for understanding):

- Calculate the mean for each class (also called as *class centroid*)
  (*Suggestion: You may use* `numpy.mean()` *function for finding the average vector for all vectors in a given class. Please define the axis property appropriately to use this function. EX:* `feat_vecs.mean(axis=0)`)
- Calculate spread (standard deviation) for each class
  (*Suggestion: You may use* `numpy.std()` *function for finding the standard deviation vector for all vectors in a given class. Please define the axis property appropriately to use this function.*)
- Calculate the distance between mean vectors between classes
  (*Suggestion:* `numpy.linalg.norm(centroid1 – centroid2)` *gives the Euclidean distance between two centroids.*)



A2. Take any feature from your dataset. Observe the density pattern for that feature by plotting the histogram. Use buckets (data in ranges) for histogram generation and study. Calculate the mean and variance from the available data.

> (*Suggestion:* `numpy.histogram()` *gives the histogram data. Plot of histogram may be achieved with* `matplotlib.pyplot.hist()`)

A3. Take any two feature vectors from your dataset. Calculate the Minkwoski distance with r from 1 to 10. Make a plot of the distance and observe the nature of this graph.

A4. Divide dataset in your project into two parts – train & test set. To accomplish this, use the train-test_split() function available in SciKit. See below sample code for help:

```
>>> import numpy as np
>>> from sklearn.model_selection import train_test_split
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

X is the feature vector set for your project and y is the class levels for vectors present in X.

Note: Before set split, make sure you have only two classes. If your project deals with multi-class problem, take any two classes from them.

A5. Train a kNN classifier (k =3) using the training set obtained from above exercise. Following code for help:

```
>>> import numpy as np
>>> from sklearn.neighbors import KNeighborsClassifier
>>> neigh = KNeighborsClassifier(n_neighbors=3)
>>> neigh.fit(X, y)
```

A6. Test the accuracy of the kNN using the test set obtained from above exercise. Following code for help.

```
>>> neigh.score(X_test, y_test)
```

This code shall generate an accuracy report for you. Please study the report and understand it.

A7. Use the predict() function to study the prediction behavior of the classifier for test vectors.

```
>>> neigh.predict(X_test)
```

Perform classification for a given vector using neigh.predict(<<test_vect>>). This shall produce the class of the test vector (test_vect is any feature vector from your test set).

A8. Make k = 1 to implement NN classifier and compare the results with kNN (k = 3). Vary k from 1 to 11 and make an accuracy plot.

Optional Section:

O1. Create a normal distribution data, plot the graph and compare the normal distribution plot against the histogram plot.

O2. Use different distance metric for kNN classifier by tuning the metric parameters of KNeighborsClassifier(). Observe the behaviour with change in the distance for classification.

Report Assignment:

1. Do you think the classes you have in your dataset are well separated? Justify your answer. [1]
2. Do you think distance between class centroids (mean of vectors in a class) is a good enough measure to test for class separability? Justify your answers. Use diagrams to illustrate your arguments. [1.5]
3. Explain the behavior of the kNN classifier with increase in value of k. Explain the scenarios of over-fitting and under-fitting in kNN classifier. [1.5]