**Section A**

1. Linear Discriminant Analysis (LDA)
2. Linear Discriminant Analysis is a very popular technique used in Machine Learning, and Data Science for data classification and dimension reduction. In this case, we will use Linear Discriminant Analysis to help tackle dimension problems.

Discriminant Analysis is used when the data set has “two or more mutually exclusive groups” [3]. Discriminant Analysis is very helpful in speech recognition, to recognize a linear combination of features which differentiates two or more objects.

1. To perform Linear Discriminant Analysis, we need to follow few steps:

* Maximize the difference mean between classes
* Minimize variance within each projected class

**4a.** Numerical, categorical

**4b.** The assumption behinds LDA is that “Linear Discriminant Analysis easily handles the case where the within-class frequencies are unequal, and their performances has been examined on randomly generated test data” [1].

**4c.** The Linear Discriminant Analysis works very well when the data set has a normal distribution, however, it will not perform well when the data is not non-Gaussian.[2]

**5a.** In general dimension reductiondoes not solve the computational cost for a given problem, but it will help reduce the dimension to avoid overfitting the model (Linear Discriminant Analysis, Sebastian Raschka 2014) [4].

**5b.** LDA is assumed to be normally distributed, each class has same covariance matrices, and each feature is independent. This concept applies only to LDA as classifier. However, “LDA for dimensionality reduction can also work reasonably well if those assumptions are violated” (Sebastian Raschka 2014) [4].

**6a.** We import LinearDiscriminantAnalysis from sklearn

**6b.** After creating a discriminant Analysis object, we give the number of components.

We call also fit\_transform () and transform ().

**7a.** To evaluate the result, we look at the confusion matrix, and the accuracy of model

**7b.** High accuracy means that the model performs very well

**7c.** Functions uses are confusion\_matrix and accuracy\_score.

**Section B**

**8.** We use in this tutorial, the iris data set which is a well-known data set in scikit-learn

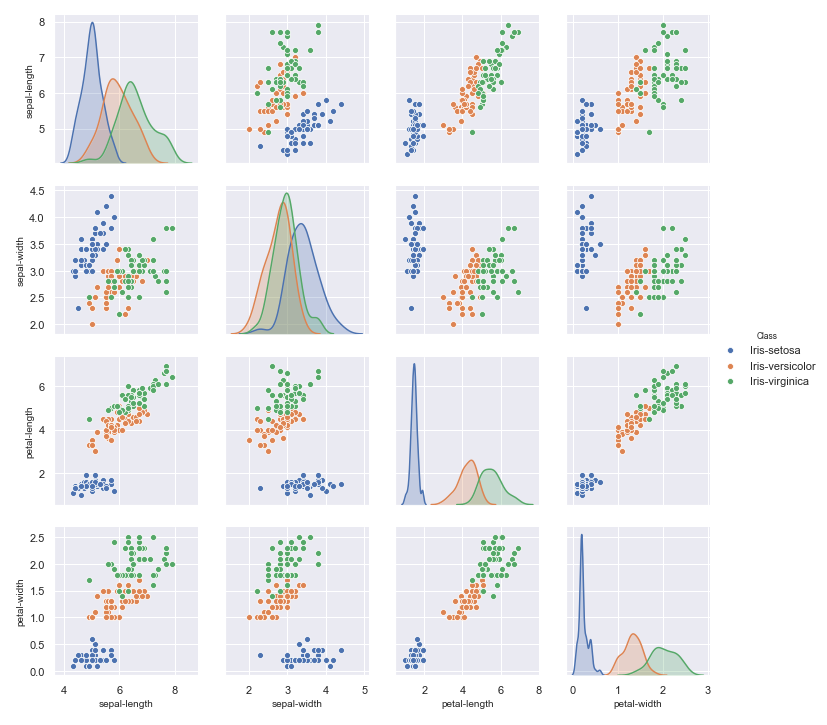
The data set has four features which are numerical values called sepal-length, sepal-width, petal-length and petal-width. Class is the target variable which has 3 groups (Iris-setosa, Iris-versicolor, and Iris-virginica).

This is the link to download the data set.

[**https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data**](https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data)

**9.** During this research we discover that Linear Discriminant Analysis is a technique that is used to do dimension reduction for supervised learning**.** We will find out also that Principal Component Analysis (PCA) is another type of dimension reduction for unsupervised learning. The only difference between the two is that LDA takes into consideration the target, and PCA does not. After researching and running the LDA, we used Random Forest and Decision Tree Classifier on the same data set to test the accuracy of the LDA

**10.**



This is a scatterplot for join relationship for univariate distribution

**References**

[1] Balakrishnama, Suresh, and Aravind Ganapathiraju. "Linear discriminant analysis-a brief tutorial." *Institute for Signal and information Processing* 18 (1998): 1-8.

[2] <https://towardsdatascience.com/is-lda-a-dimensionality-reduction-technique-or-a-classifier-algorithm-eeed4de9953a>

[3] <http://www.writeawriting.com/research/types-discriminant-analysis/>

[4] <https://sebastianraschka.com/Articles/2014_python_lda.html>