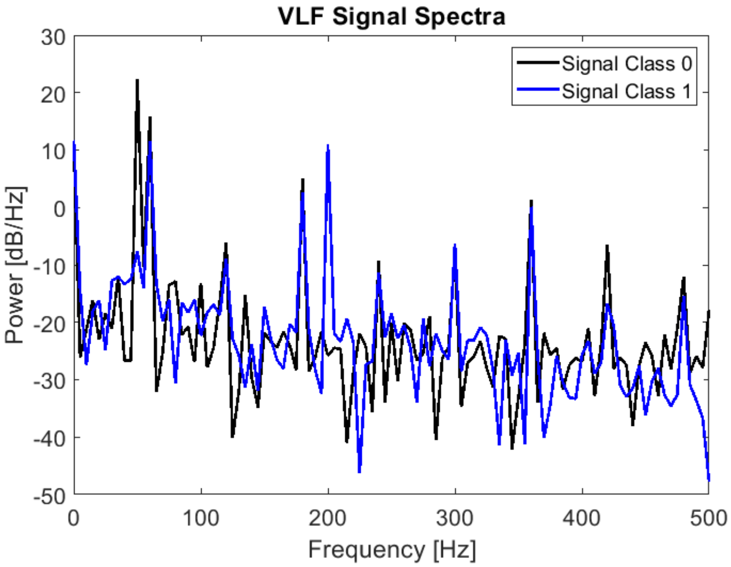
Homework #5 – Extra Credit

Assigned: 04/15/2021
Due: 05/04/2021 (11:59 PM on CANVAS)

1. **VLF Signal Detection with PCA:** In this problem you will repeat the analysis from Exam 1, but after PCA decomposition. As a reminder for this problem, two classes of signal spectra from a low frequency loop transmitter have been collected between 0 and 500 Hz. The figure below shows a picture of the transmitter along with sample cases of the two different signal classes.

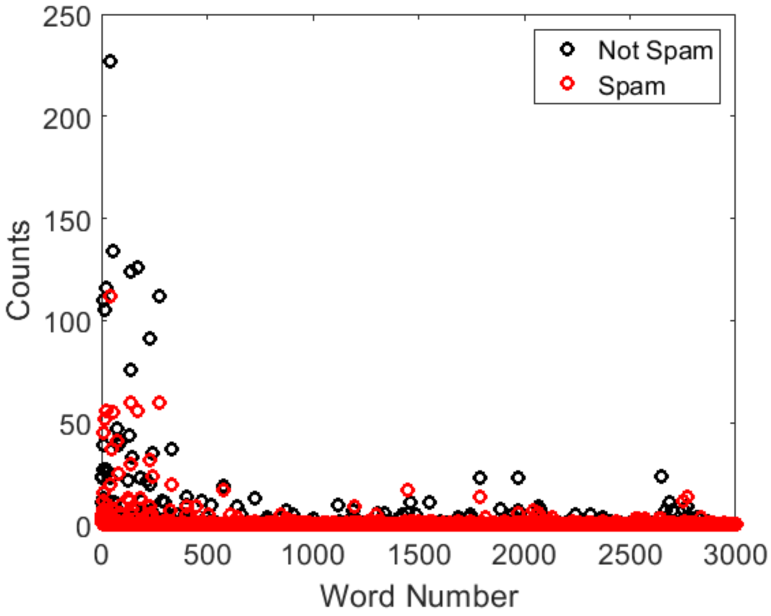


The input data consists of 1000 signal collections with 101 frequency bins (input features). The two signal classes are labeled '0' and '1' respectively. The input file is 'Signal_input_data.csv' and corresponding class labels are in 'Signal_class_data.csv'.

a) Split the data using an 80-20 split of training to test data. Using the training data build a perceptron classifier and show a confusion matrix on both the test and training data (you already needed to do this on Exam 1, you should observe 100% accuracy with this model).
b) Using the training data, decompose the input using PCA. Plot the sum of the explained variance as a function of PCA component number for all 101 components.
c) Plot the first PCA component as a function of frequency index. What are the dominant frequencies present in the first PCA component? **Hint: The input features are equally space in frequency between 0 and 500 Hz (inclusive).*
d) Show that a perceptron classifier using only 1 PCA dimension is just as accurate as the original perceptron classifier in part (a).

e) What do your answers in (c) and (d) imply about the frequency spectrum of the two signal classes?

2. **Spam Email Classifier with PCA:** For this problem you will once again design a spam classifier, but this time with PCA. As a reminder, you can find the processed data for 5172 emails using *spam_data.csv*. For each email (row) the word count is shown for 3000 different possible words (columns). The last column contains the class ("1" for spam and "0" for not spam). To visualize this, the figure below shows the word count vs word number for a spam and a non-spam email.



Split the data using an 80-20 split of training to testing data.

a) Using the training data, decompose the input using PCA. Plot the sum of the explained variance as a function of PCA component number for all 3000 components

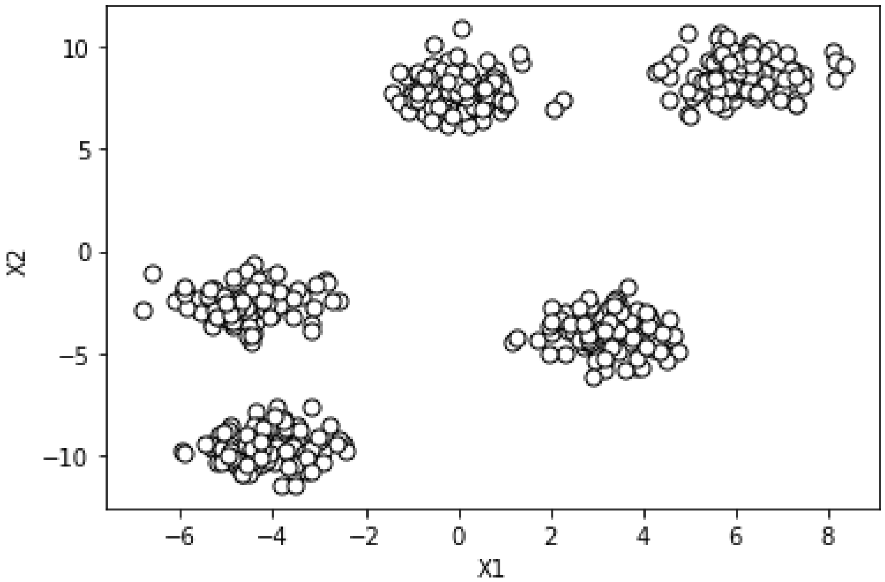
b) Build a logistic regression classifier that utilizes only the first *N* PCA components.

c) Using your model in (b), plot the accuracy of the classifier as a function of PCA component number up to *N = 50.*

d) Using your model in (b), plot the training time of the classifier as a function of PCA component number up to N = 50.

e) Based on your answers in (c) and (d), how many dimensions would you reduce the initial data to? **Note: There isn't just one answer to this, just make sure to explain your reasoning.*

3. **Clustering Models Comparison:** For this problem you will utilize clustering algorithms to find clusters in a 2-dimensional dataset. The cluster data is saved as a Numpy variable in *'Cluster_data.py'* as a 2D array *X*. The array is size 512x2 which correspond to the 512 data samples and 2 features. A scatterplot of the data is shown below:



a) Using K-means clustering, determine the centroids of each cluster in the dataset. Make a scatterplot of the data with each cluster shown in a different color. *Hint*: You can determine K in any way you want, but it should be obvious from the figure above.*

b) Using a Gaussian Mixture Model, determine the means of each Gaussian in the dataset (*Note*: This information is available in the model attributes, "means_"*). How do these values compare to the centroids from K-means in part (a)?