

Experimentation has been done on Human Gender data. Features include :

Frequency – Mean, SD, Median, Q25, Q75, Centroid, Peak

Fundamental Freq – Mean, SD, Median, Max, Min

Dominant Freq – Mean, Min, Max, range

**Modulation Index** 

Skewness

Kurtosis

Label – Male or Female

### Scope of the project

- Run Kmeans on original data, features from Neural Networks and compare them.
- Run more experiments. They are as follows :
  - Create 10% of data as outlier for one feature. Rerun above step
  - Create 10% of data as outlier for all features. Rerun above step
  - Mislabel 10% of targets (Can Neural Network handle it?)
  - Mislabel 50% of targets

# **Primary Metrics**

#### Cluster Purity:

Sum of Maximum class in each cluster / Total Number of obs Value ranges from 0 – 1

#### > Improvement :

How well off are we as compared to original Purity. ((Current\_purity/Original\_purity)-1) \* 100 Value ranges from 0 – 100%

#### Outlook of data – This is what we are dealing with

#### PCA on the original features



Classes are overlapping. Not in a great position to cluster both classes properly.

**Cluster Purity : 0.65** 

#### Is there a way to seperate the data space?

#### > Here come Neural Networks!



Let's change the underlying structure of data using Neural Networks!

Features :

We extract the value out of the Activation Function for all the neurons in the network.

### **Neural Network and Kmeans settings**

- Since we have 2 classes we are assuming k should be 2
- >With a little trail and error, here are our hyper-parameters.
- <u>Learning Rate</u> 0.01
- **Activation Function** Tanh
- <u>Epochs</u> 15
- Hidden Layer and Neurons Variable

### **Changing Feature Space**



-0.050 -

-0.02

0.00 PC1 (84.89%)

0.02

0.04

-0.02

0.00

PC1 (84.02%)

0.02

## **Purity Comparison**



purity using original features vs NN features

The best purity is obtained by a Neural Network with 6, 2 as Hidden Layer setting with purity – 0.97

#### A significant increase from 0.65!

### **Result Summary**



Neural Network Improvement summarization

# Experiment – 1 10% of data into outlier for one feature



#### **Experiment1 – Purity Comparison**



Purity comparison. Outlier - 10% on single feature

The best purity is obtained by a Neural Network with 2, 1 as Hidden Layer setting with purity – 0.87

#### **Experiment1 – Result Summary**



NN Improvement summarization on 10% outliers on single feature

type

# Experiment – 2 10% of data into outlier for all features



#### **Experiment2 – Purity Comparison**



Purity comparison. Outlier - 10% on all features

The best purity is obtained by a Neural Network with 2, 1 as Hidden Layer setting with purity – 0.8

#### **Experiment2 – Result Summary**

Max Improvement-NN41 :43.09 40 -Mean Improvement: 0.9 20 improvement Min Improvement-NN11 :-15.2 NGKUNIZEUSISISISTEBASCOMUZUSHABASOOSINIGEUSHISISIDEZUSIZEUSISTEBASISISIGTEBASISISTEBASISISTEBASISISTEBASISISTE [63] 사람이 있는 것 같은 것을 알 수 있는 것 같은 것을 많이 있는 것을 알 수 있다.

NN Improvement summarization on 10% outliers on all features

type

# Experiment – 3 10% of mislabels



#### **Experiment3 – Purity Comparison**



The best purity is obtained by a Neural Network with 1, 2 as Hidden Layer setting with purity – 0.87

#### **Experiment3 – Result Summary**



type

# Experiment – 4 50% of mislabels



### **Experiment4 – Purity Comparison**



Neural Network doesn't do well when big number of targets are mismatched.

The best purity we obtained is 0.52 for a Neural Network setting 0,1

#### But it still leads by 0.01!



#### **Experiment4 – Result Summary**

NN Improvement summarization on 50% mislabels



# For better results apply Neural Networks somehow!



