Report on Clustering on the Bag of words

Soham Pyne/MDS202148 and **Bharath Ravilla/MDS202133**

CMI — May 23, 2022

Introduction

We are given with a dataset called Fashion MNIST. The goal is to apply the semisupervised learning method on the dataset for classification purposes.

1 Method

The idea of the code is to define a pipeline for the classification in the following way

- It defines k-means clustering for specified k
- It fit and transform data into the cluster distance space
- It finds the data point closest to each centroid
- It represents each cluster using that datapoint and its label
- It applies k-means to predict labels on the validation set
- It finds the accuracy scores

2 Applications

We have tried the algorithm on three different labelled instances of semi supervised learning.

2.1 25 clusters

This has comparatively smaller number of the labelled instances to start with. It has the accuracy level 0.5895.

2.2 45 clusters

This has comparatively smaller number of the labelled instances to start with. It has the accuracy level 0.6469.

2.3 85 clusters

This has comparatively smaller number of the labelled instances to start with. It has the accuracy level 0.699.

3 Conclusion

We see that increasing the number of semi supervised instances increases the accuracy level of the prediction. This is intuitive on the stand point that increase in number of labelled instances to start with makes it more of a supervised learning problem.