CS7641 HW3 -- Unsupervised Learning

Yi Qin

# 1 Objective

The objective of this assignment is to explore unsupervised learning such as clustering and dimensionality reduction algorithms by applying different datasets. The goal is to learn the difference, similarity and interactions with other machine learning algorithms. In the first part, I will apply two clustering algorithm which are K-Means and Expectation Maximization; in the second part, 4 dimensionality reduction methods will be used on same datasets and run clustering again based on the dimension reduced results; in the third part, I will run neural network using dimension production data; in final, I will use clustering as feature reduction algorithm for running neural network model.

# 2 Dataset

## 2.1 Movie Review

This is the dataset which I used in assignment 1. The dataset has 25000 movie reviews; each review will be transfer into a vector with 300 dimensions using word embedding. This dataset is interesting for this clustering and dimension reduction topic, because it has high dimensions, also, the vector itself shows the contextual meaning for a single word, I’m very curious about if dimension reduction will break the relationship and scale factor within a vector.

## 2.2 Phishing Websites

Since both datasets from my assignment 1 are NLP related datasets, it will be more meaningful to reproduce the process in a different dataset.

Phishing website become a serious security problem right now, firewall has its limitations, build a model to detect it is of practical importance for cyber security. I will use this dataset to predict if it’s a phishing website with 30 categorical features and 11055 instances. The target has 2 classes indicating whether the website is phishing or legitimate. The dataset was downloaded from https://www.openml.org/d/4534