K-Means with PCA on Lung X-Rays

Wilson Tan, Girish Senthil, Sydney Hendricks, Kristin, Moengen, Erik Spone, and Andreas Bronderop

1.) Introduction

The purpose of this project is to compare and contrast the results and runtime of K-Means clustering with and without dimensionality reduction with principal component analysis (PCA) on lung x-rays. We used a dataset which consisted of chest x-rays, but more specifically, lung x-rays of normal and pneumonia-infected lungs. We tested on K-Means alone and K-means with PCA to compare how accurate were the x-rays clustered and how long K-Means took to cluster the original dataset and the reduced dataset.

2.) Data

The data used in this project came from "Kaggle" "https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia". A total of 5216 x-rays were used in this project with approx. 1800 x-rays of normal lungs and 3400 x-rays of pneumonia-infected lungs. Due to different sizes in images, all images were reshaped to 256x256 and color converted to gray. Each image was then labeled with their corresponding label (i.e., normal or pneumonia). These labels will be used to identify the images. We can see a visualization of the x-rays after image processing in the figure below.

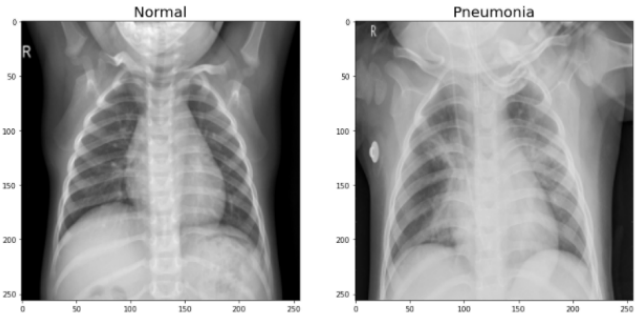

Figure 2.1

3.) K-Means

K-Means with two clusters showed promising results. Figure 3.1 shows the first 5 images in each cluster (10 total) where each row represents a different cluster. In cluster 1 (top row), 4 out of 5 images are pneumonia. In cluster 2 (bottom row), 4 out of 5 images are normal.

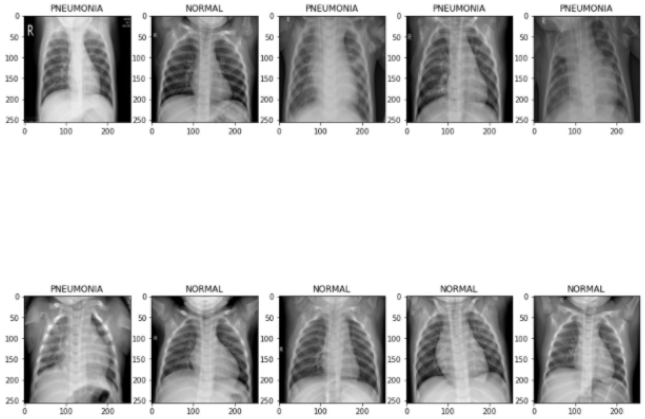

Figure 3.1

Because there are different types of pneumonia, K-Means with 4 clusters was also tested. Though because of the lack of labels, which means a lack of identification, distinguishing between these different clusters proves difficult. The results are depicted in Figure 3.2

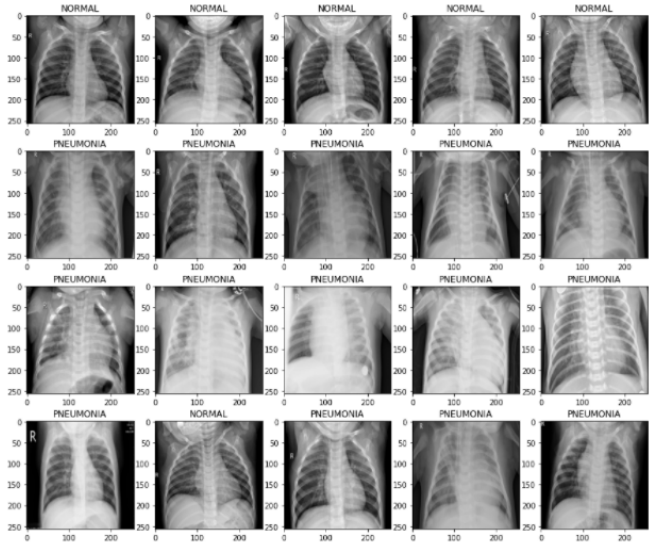

Figure 3.2

4.1) PCA

Using PCA, a lower dimensional representation of the images can be attained which are then used in K-Means. The first 625 principal components, which approximates to 96% of the variance, were used when combining K-Means with PCA.

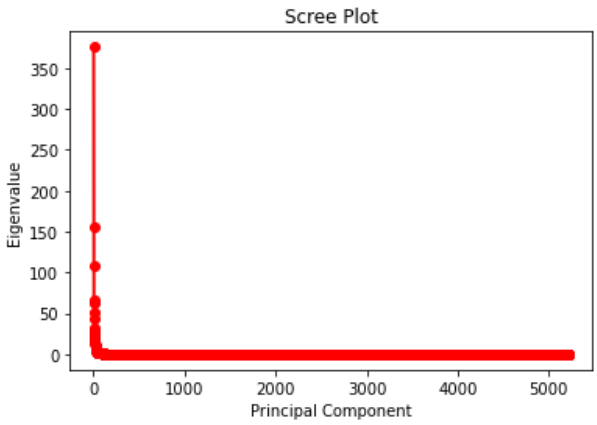

Figure 4.1

Figure 4.1 shows a scree plot which depicts the variance explained by each principal component. Although it's hard to see how much variance is explained by a single principal component, the graph clearly shows thousands of principal components that explain little variance. Therefore, dimensionality reduction would be useful in this case.

4.2.) K-Means with PCA

When running K-Means with PCA, we get the following results.

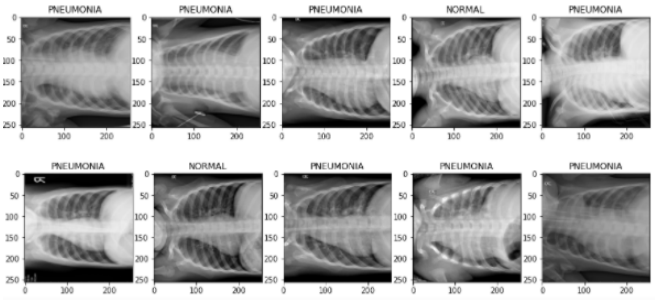

Figure 4.2: Note that the images are rotated due to taking the transpose, but labels are still correct

Figure 4.2 shows 4 out of 5 pneumonia lungs for both clusters. When reducing images using dimensionality reduction, K-Means was unable to distinguish between normal and pneumonia x-rays.

5.) Comparison

When comparing the results of K-Means with and without PCA (Figure 4.1 and 3.1 respectively), it seems as though K-Means with PCA was unable to find the underlying structure to the data whereas K-Means alone performed well in this category. However, this is not the case for runtime. K-Means alone took approximately 83s to complete for 2 clusters whereas with PCA, it took approximately 2s.

Figure 5.1.Note that only 625 images were used when testing K-Means alone. This was due shorten the runtime

From figure 5.1, even when running K-Means alone with a drastically reduced number of images (625 out of 5216), K-Means with dimensionality reduction is still faster.

6.) Conclusion/Discussion

From the results, we assume that K-Means alone outperformed K-Means with PCA. This is expected since having more information generally leads to better results. However, the results of K-Means with PCA is another story. We were able to account for approximately 96% of the variance by taking the first 625 principal components, but this does not seem to be enough. From figure 4.1, we see that there are a few thousand principal components that make up the missing 4% of variance. It's possible that including a couple thousand more principal components could drastically improve the results at the cost of resources, like time. However, with something as important as detecting pneumonia, should accuracy be traded off for anything else in the first place?