# Analysis of 3-fold Cross Validation for KNN, K-means and Hierarchical Clustering implementations

Saketh Saxena

October 18, 2017

## 1 Introduction

As part of the CS256 Homework 3 I have implemented KNN algorithm, K-means and Hierarchical Clustering (single linkage clustering) for image data, and tested/validated them on images.

## 2 3-fold cross validation for KNN

3 fold cross validation for K nearest neighbour algorithm was done on the dataset, with tunable k value. For my experiments, I have plugged in k = 3 and also computed the accuracies for k = 1 to 10, the results are as depicted in the graphs and screenshot below

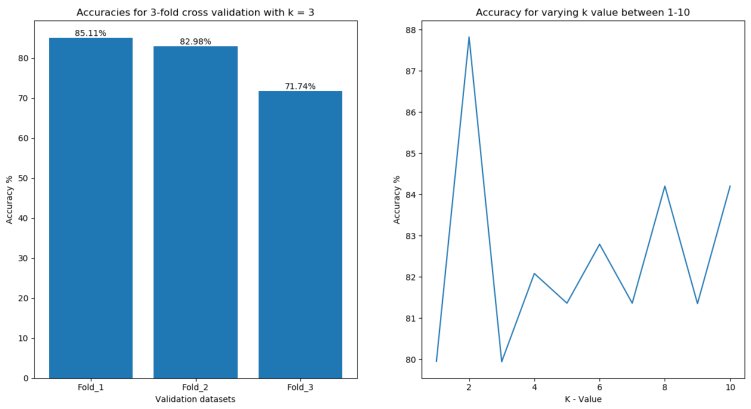
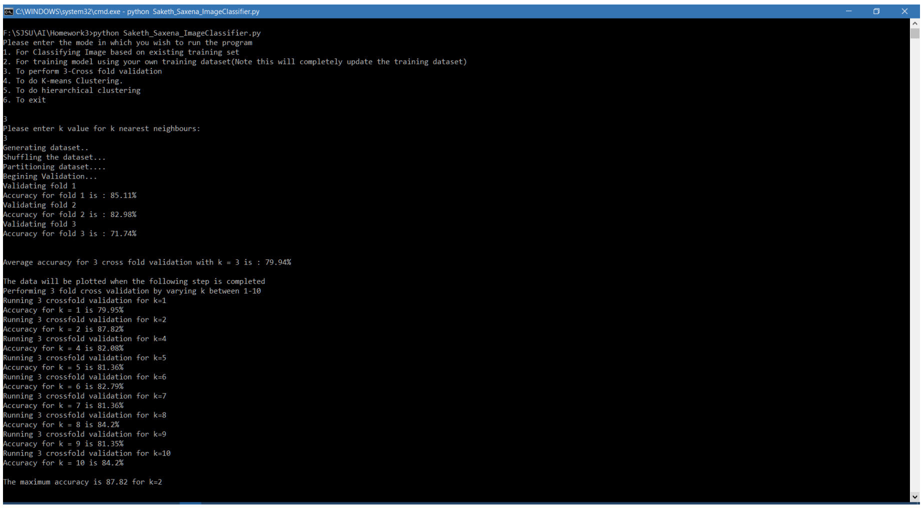


Figure 1: Bar graph showing accuracy for each fold when k =3 and line plot showing the variation of accuracy for k = 1 to 10

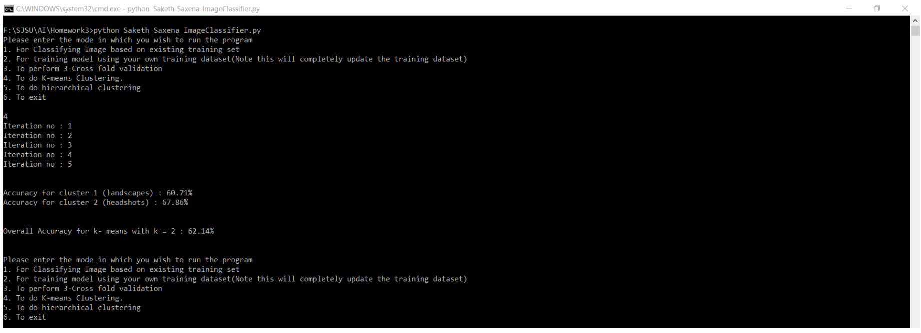

C:\WINDOWS\system32\cmd.exe - python  Saketh_Saxena_ImageClassifier.py

```
F:\SJSU\AI\Homework3>python Saketh_Saxena_ImageClassifier.py
Please enter the mode in which you wish to run the program
1. For Classifying Image based on existing training set
2. For training model using your own training dataset(Note this will completely update the training dataset)
3. To perform 3-Cross fold validation
4. To do K-means Clustering.
5. To do hierarchical clustering
6. To exit

3
Please enter k value for k nearest neighbours:
3
Generating dataset..
Shuffling the dataset...
Partitioning dataset....
Begining Validation...
Validating fold 1
Accuracy for fold 1 is : 85.11%
Validating fold 2
Accuracy for fold 2 is : 82.98%
Validating fold 3
Accuracy for fold 3 is : 71.74%


Average accuracy for 3 cross fold validation with k = 3 is : 79.94%

The data will be plotted when the following step is completed
Performing 3 fold cross validation by varying k between 1-10
Running 3 crossfold validation for k=1
Accuracy for k = 1 is 79.95%
Running 3 crossfold validation for k=2
Accuracy for k = 2 is 87.82%
Running 3 crossfold validation for k=4
Accuracy for k = 4 is 82.08%
Running 3 crossfold validation for k=5
Accuracy for k = 5 is 81.36%
Running 3 crossfold validation for k=6
Accuracy for k = 6 is 82.79%
Running 3 crossfold validation for k=7
Accuracy for k = 7 is 81.36%
Running 3 crossfold validation for k=8
Accuracy for k = 8 is 84.2%
Running 3 crossfold validation for k=9
Accuracy for k = 9 is 81.35%
Running 3 crossfold validation for k=10
Accuracy for k = 10 is 84.2%

The maximum accuracy is 87.82 for k=2
```

Figure 2: Screenshot depicting the output when k=3

# 3 k-means Clustering Algorithm

The K-means implementation was used to cluster about 140 images belonging to two classes - landscapes(77 in number) and headshots(63 in number). The implementation was tested with k = 2; and the percentage accuracy for each cluster i.e. number of correctly clustered images upon number of images in the cluster and the average percentage accuracy were computed, the results of which are displayed below:


C:\WINDOWS\system32\cmd.exe - python  Saketh_Saxena_ImageClassifier.py

```
F:\SJSU\AI\Homework3>python Saketh_Saxena_ImageClassifier.py
Please enter the mode in which you wish to run the program
1. For Classifying Image based on existing training set
2. For training model using your own training dataset(Note this will completely update the training dataset)
3. To perform 3-Cross fold validation
4. To do K-means Clustering.
5. To do hierarchical clustering
6. To exit

4
Iteration no : 1
Iteration no : 2
Iteration no : 3
Iteration no : 4
Iteration no : 5


Accuracy for cluster 1 (landscapes) : 60.71%
Accuracy for cluster 2 (headshots) : 67.86%


Overall Accuracy for k- means with k = 2 : 62.14%


Please enter the mode in which you wish to run the program
1. For Classifying Image based on existing training set
2. For training model using your own training dataset(Note this will completely update the training dataset)
3. To perform 3-Cross fold validation
4. To do K-means Clustering.
5. To do hierarchical clustering
6. To exit
```

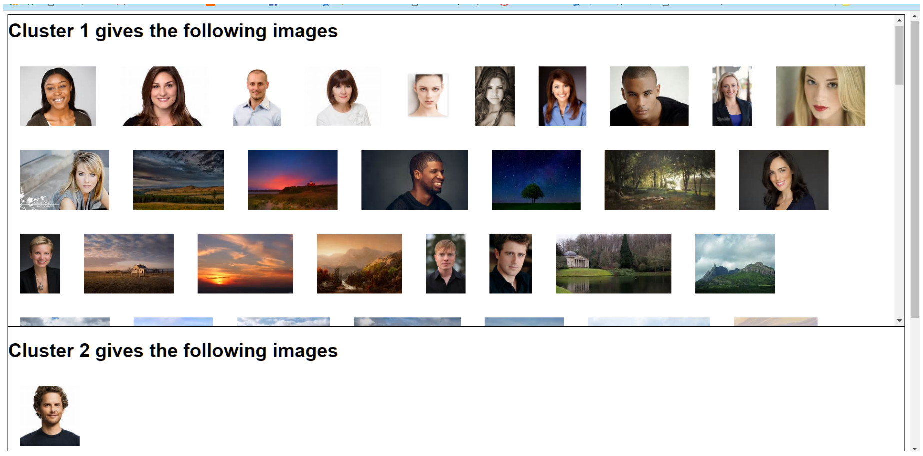
Figure 3: Accuracy of C1 = 60.71 % and C2 = 67.86 % and Average Accuracy = 62.14%

# 4 Hierarchical/Single linkage Clustering

The result of the Hierarchical cluster was partitioned into two clusters and the images displayed in the web browser. The resulting clusters were extremely skewed with cluster being large and one having 1-2 images in them, this could possibly be due to outliers or noise in the data, where some images are extremely different than others.

The implementation was used to cluster 1. landscape and headshot image data 2. 206 images of flags of nations.

## 4.1 Clusters for Landscape Vs Headshot

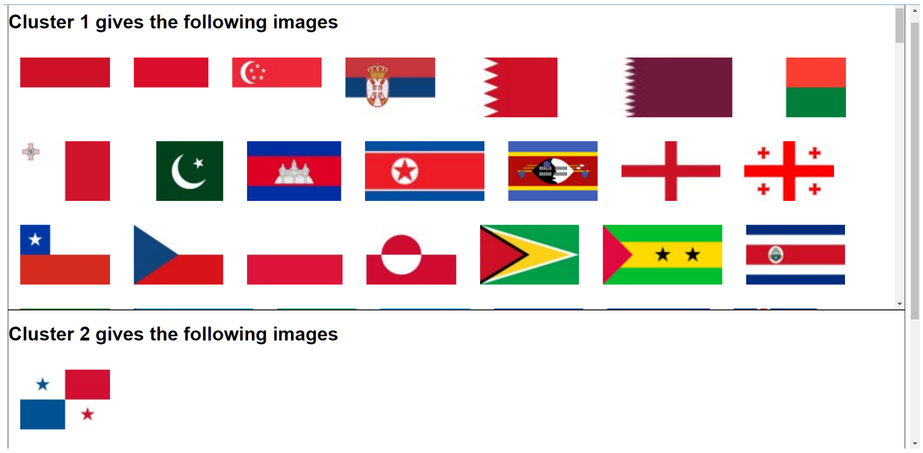


Cluster 1 gives the following images

Cluster 2 gives the following images

Figure 4: Screenshot of output showing images of headshots vs landscapes in 2 clusters

## 4.2 Clusters for flags



Cluster 1 gives the following images

Cluster 2 gives the following images

Figure 5: Screenshot of output showing images of flags in 2 clusters

## 4.3 Inference of results

From the results obtained by hierarchical clustering, it shows that this approach is not a very good approach to cluster these datasets, also hierarchical clustering is not particularly effective when the clusters need to be partitioned.