Assignment 3

Louis Le Breuilly, Nils Jansen, Alec Knapp

April 4, 2017

1 Introduction

The tasks stipulated by the exercise 3 report specification were performed and are detailed in this document.

2 Method

Ex. 3.1: K-Means Clustering

Ex. 3.1.1: K-means clustering was applied to each cipher individually in the training set in order to represent the data as a set of centroids, thereby sparcifying the datasets and attempting to improve computational efficiency.

Ex. 3.1.2: KNN was then performed using the centroids of these clusters, and the performance found to be dependant on cluster size. Figure 1 shows the observed relationships. The time relationships were found to be somewhat unexpected, chaotic and inconsistent. The authors attribute this to the fact that the runs were performed on an Amazon computer cluster system, which may dedicate variable amounts of computational resources throughout a given computational run.

Accuracy performance was observed to be largely independent of cluster sizes, behaving reasonably consistently throughout computational runs. It is observed that accuracy is broadly quite low, in a potentially anomalous fashion for all values of k.

Ex. 3.1.3: This process was found to be excessively computationally intensive to be able to complete within the requisite time frame. Computation for the below parameters was initiated at 12:42hrs on Friday 31st March, using a computer with 2x 3GHz CPUs and 8GB RAM:

- Number of k values: 17
- Number of cluster sizes: 5
- Groups: 1-10
- Data type: person independent

As of 2000hrs on Tuesday 4th April this has failed to complete and therefore is not included here due to temporal constraints.

Ex. 3.2: Hierarchical clustering

Ex. 3.2.1: Figure 2 shows the appropriate low-level KNN dendrogram for two person data of five instances per digit before the application of k-means clustering.

Ex. 3.2.2: Figure 2 shows the appropriate low-level KNN dendrogram for two person data of five instances per digit after the application of k-means clustering.

Ex. 3.2.3: Comparison of the subfigures of figure 2 reveals that preprocessing with k-means clustering results in clusters being formed in the KNN process much later (relatively, throughout the full proportion of the clustering process) than compared to the case where no k-means clustering preprocessing has been applied. The preprocessing also results in the last cluster being formed at a much lower height than without preprocessing.

Ex. 3.3: K-Means Clustering

Ex. 3.3.1: Precision-recall curves for $1 \leq k \leq 13$ have been plotted in figure 3.

Ex. 3.3.2: The maximum F1 values for each k value have been plotted in 3.

Ex. 3.3.3: F_1 score is defined in equation 1.

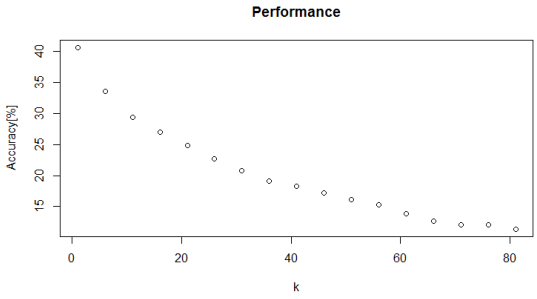
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{1}$$

Inspection of figures 3 and 3 show that the maximum F1 score generally decreased for higher k values in this instance. In general, a higher F1 score is ideal, and as such, lower k values might be more appropriate (though the difference of $\approx \pm 1.5$ is not an exceedingly dramatic difference). For certain optical character recognition applications, such as for highly important documents (e.g. paycheques, legal documents etc), an extremely high recall factor is likely to be essential - this therefore may necessitate human intervention if a suitably high recall factor cannot be obtained using such methods.

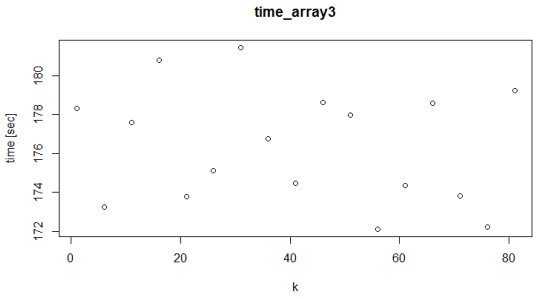
3 Executive Summary

The tasks stipulated in the exercise 3 project specification have been performed as required. Certain unexpected results have been encountered, such as the seemingly anomalous time and accuracy efficiency ratings in figure 1. Temporal restrictions have prevented the completion of Ex 3.1.3. Otherwise, the processes have proceeded in a nominal fashion.

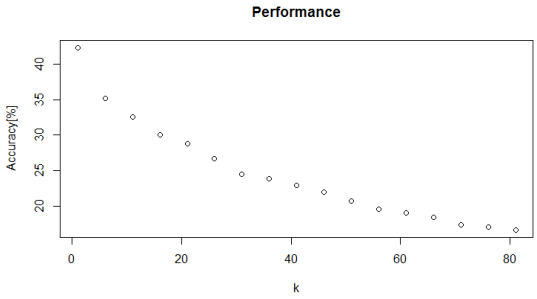
4 Annexe



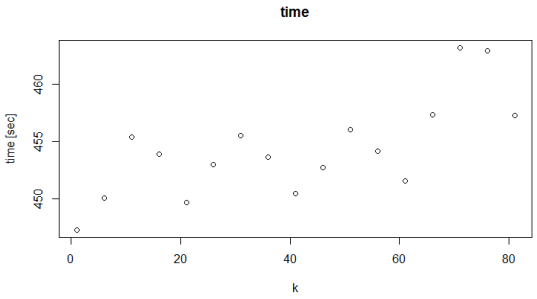
(a) C25 performance



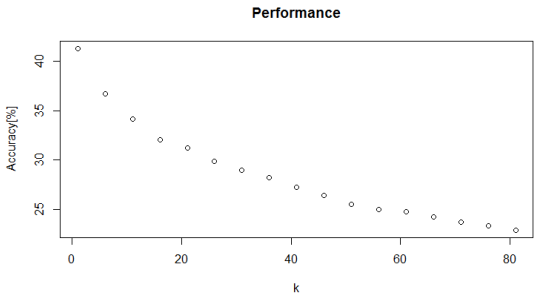
(b) C50 time



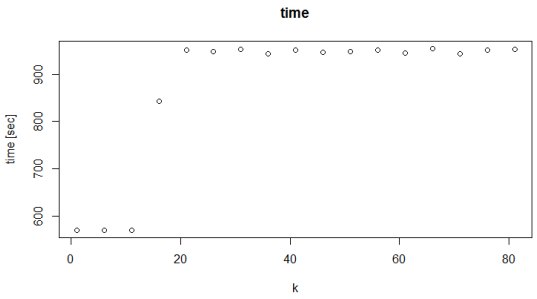
(c) C50 performance



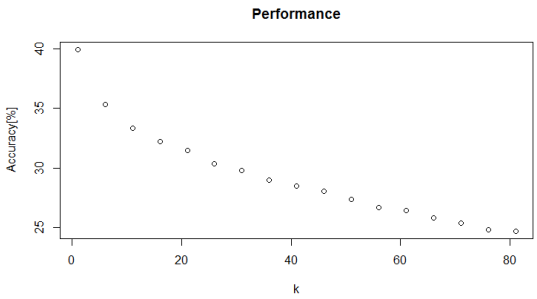
(d) C50 time



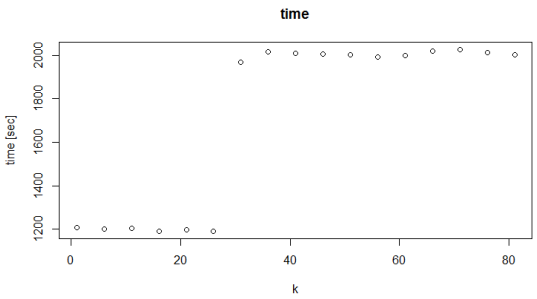
(e) C100 performance



(f) C100 time

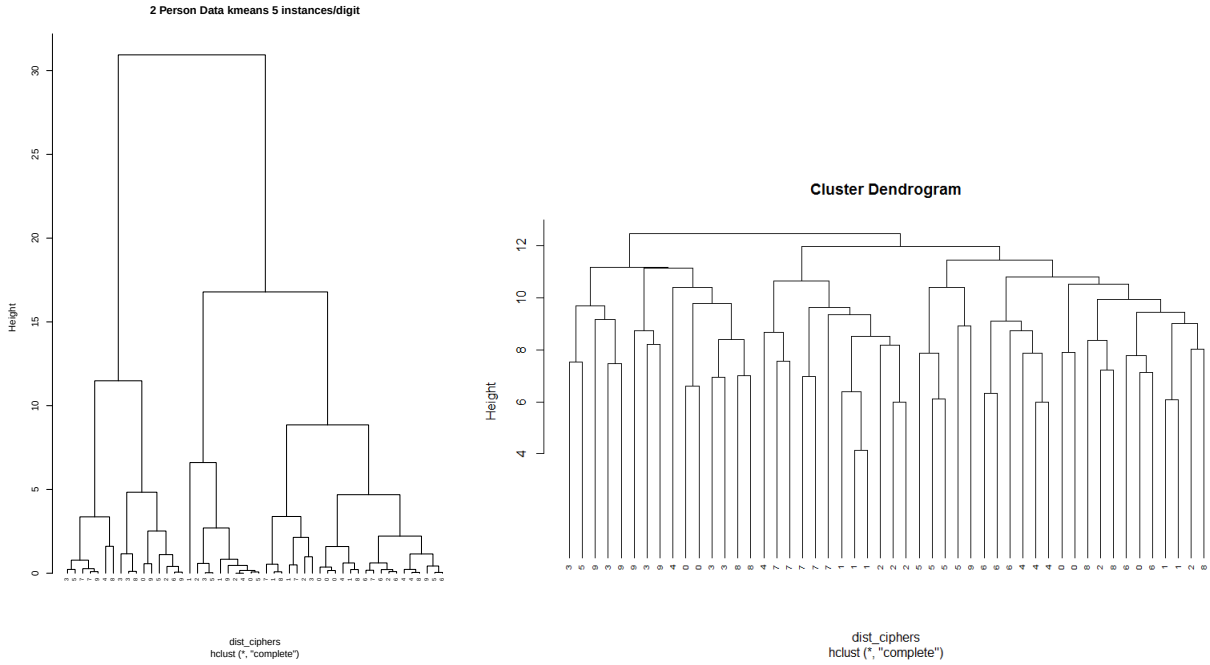


(g) C200 performance



(h) C200 time

Figure 1: Performance (accuracy and temporal) for varying k and cluster sizes



(a) KNN before k-means clustering for two person data

(b) KNN after k-means clustering for two person data

Figure 2: Dendrograms of KNN before and after k-means clustering for two person data

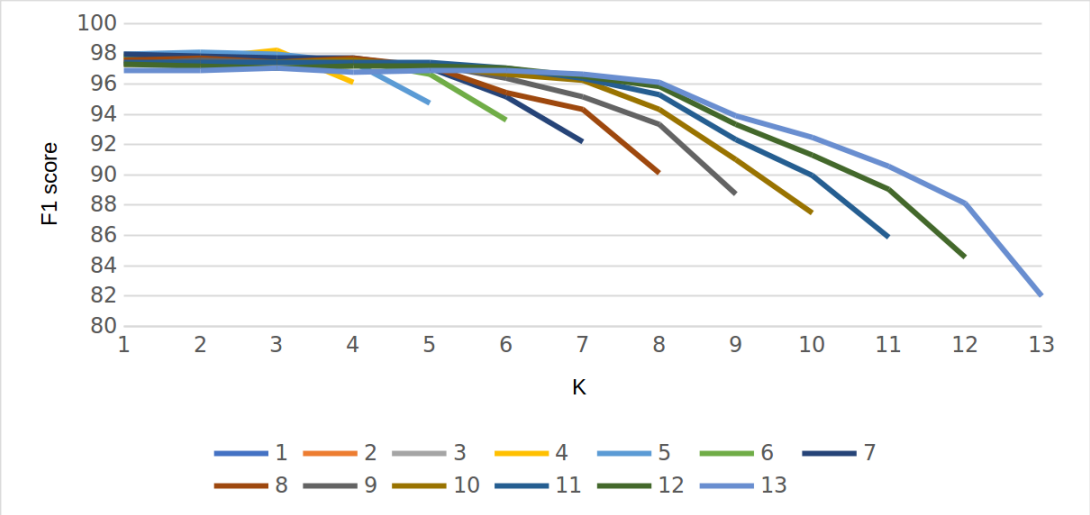


Figure 3: Precision-recall curve for $1 \leq k \leq 13$

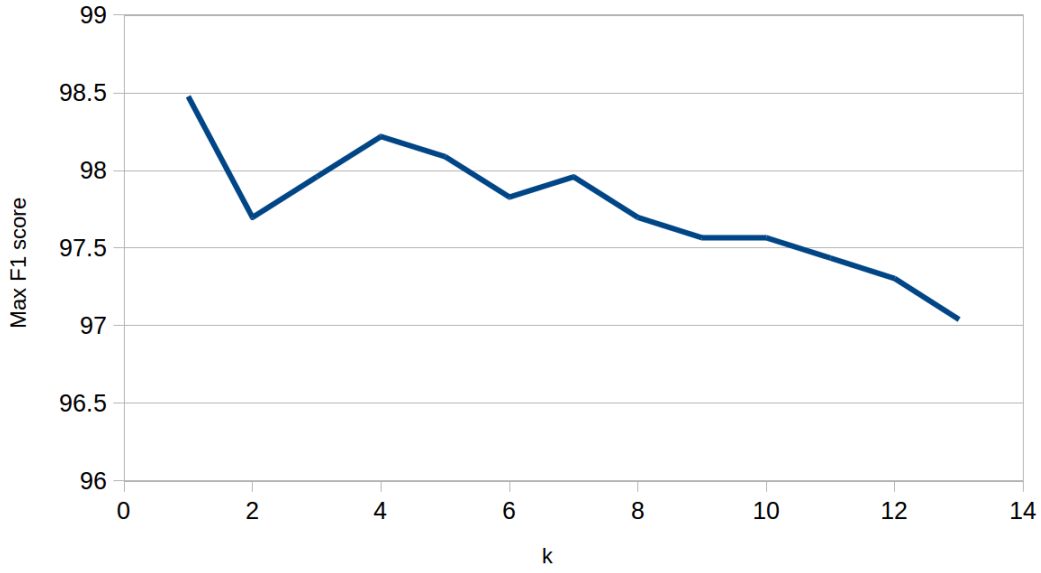


Figure 4: Maximum F1 score for $1 \leq k \leq 13$