Prepared for submission to University of Cambridge

M1: Applied Data Science - Coursework Assignment

Adnan Siddiquei

University of Cambridge

E-mail: as3438@cam.ac.uk

Contents

1 Section A

This section contains the answers to the questions in Section A of the coursework assignment. Wherever the phrase 'dataset' is used in this section, it refers to `A_NoiseAdded.csv`.

1.1 Dataset A

1.1.1 Question 1a

Fig(1) shows the combined and separated kernal density estimates of the first 20 features of the dataset. Several conclusions can be drawn from this plot. For most of the features, the KDE is concentrated around the zero value with very little variance, with the exception of 7 features: 5, 11, 13, 14, 18, 19 and 20. These features are likely to be more discriminative within classification algorithms, as they contain more variability.

1.1.2 Question 1b

The biplot in Fig(2) shows a PCA of the entire dataset, after standardisation has been applied to each feature. The coloured arrows indicate the loadings of the features which contribute more than 2% to either principal component, which implies that none of the first 20 features are particularly discriminative. An interesting observation is that the most discriminative features are very disciminative with respect to the second principal component, whereas the less discriminative features (of which there are many, many more of) are more discriminative with respect to the first principal component. As a result, the scores on the biplot for the observations w.r.t. the first two principal components separate the observations more along the first principal component, rather than the second.

1.1.3 Questions 1c and 1d

Fig(3a) shows the contingency table for two k-means clusterings of the dataset with $k = 8$ and $k = 3$. The two clusterings (`kmeans_1` and `kmeans_2`) were formed by training two k-means models on half of the dataset, and then the other half were mapped onto the learned clusters. Standardisation was applied to each feature before clustering as k-means is sensitive to feature scaling, this is because k-means works by computing a centroid for each cluster such that the sum of the squared Euclidean distances between each centroid and the observations in the cluster is minimised. Therefore, without standardisation, features with larger variances and scale will dominate the clustering. Predictions are then made by assigning new observations

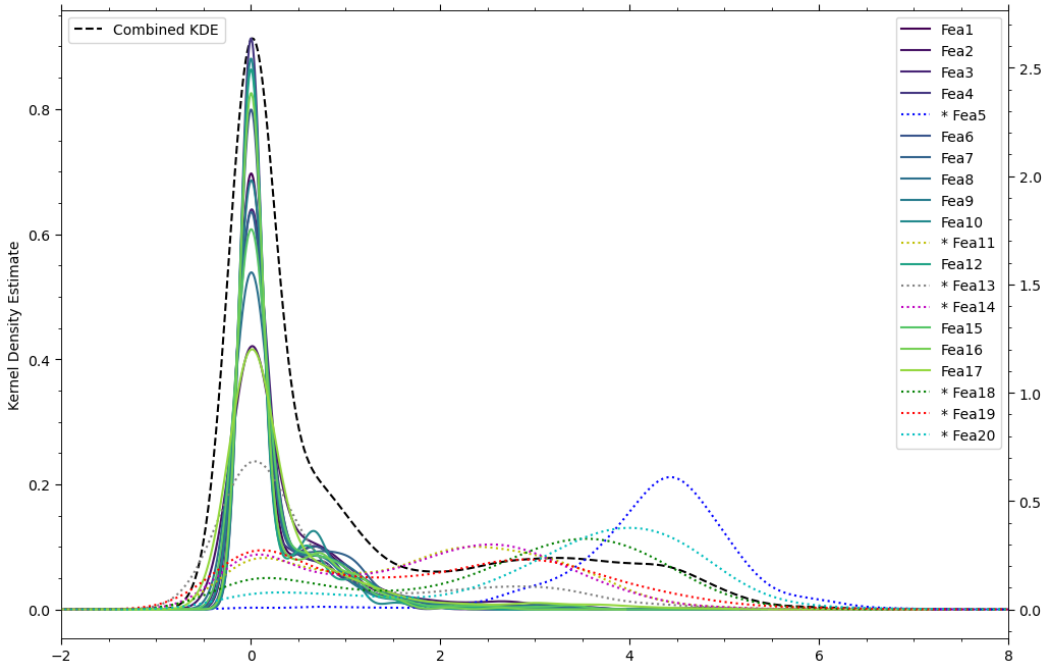

Figure 1: A kernel density estimate of the first 20 features of the `A_NoiseAdded.csv` dataset, which has 20 plots in total, with the legend and corresponding y-axis on the right. A combined KDE has also been plotted, with it's corresponding x-axis on the left. 7 of the 20 features have been highlighted in the legend with an asterisk, and have been coloured in slightly more contrasting colours and plotted with a dotted line. These features have a larger variance in their density, and are therefore more likely to be more discriminative.

to the cluster whose centroid is closest to the observation [1]. Crucially, the labels assigned to each cluster from each k-means computation are arbitrary, therefore, the labels in `kmeans_2` were mapped back to the labels in `kmeans_1` by finding which centroids in each clustering were closest to each other. This step was crucial, otherwise the contingency table would be meaningless.

For $k = 8$, 59% of the observations lie on the leading diagonal, which indicates that the two clusterings were not very similar or stable, as a large proportion of the observations were assigned to different clusters in each clustering. `kmeans_1` clustered 85% of the observations into clusters 1, 2 and 3. `kmeans_2` clustered 86% of the observations into clusters 1 and 2. This indicates that both clusterings identified most of the data lives within a small set of clusters, which was expected as the dataset is labelled and so it is known that there are only 3 clusters - both clusterings identified at least 2 similar clusters (1 and 2). However, the presence of very small clusters indicates that there may be outliers present in the dataset or smaller clusters that the large $k = 8$ is overfitting to.

Fig(3b) shows the contingency table for two k-means clusterings of the dataset with $k = 3$. A larger proportion of the observations lie on the leading diagonal compared to $k = 8$, which is expected because the number of clusters is both smaller and equal to the actual number of clusters in the dataset. `kmeans_2` identified 3 distinct clusters whereas `kmeans_1` only identified 2 distinct clusters with a few remnant observations in assigned to cluster 2.

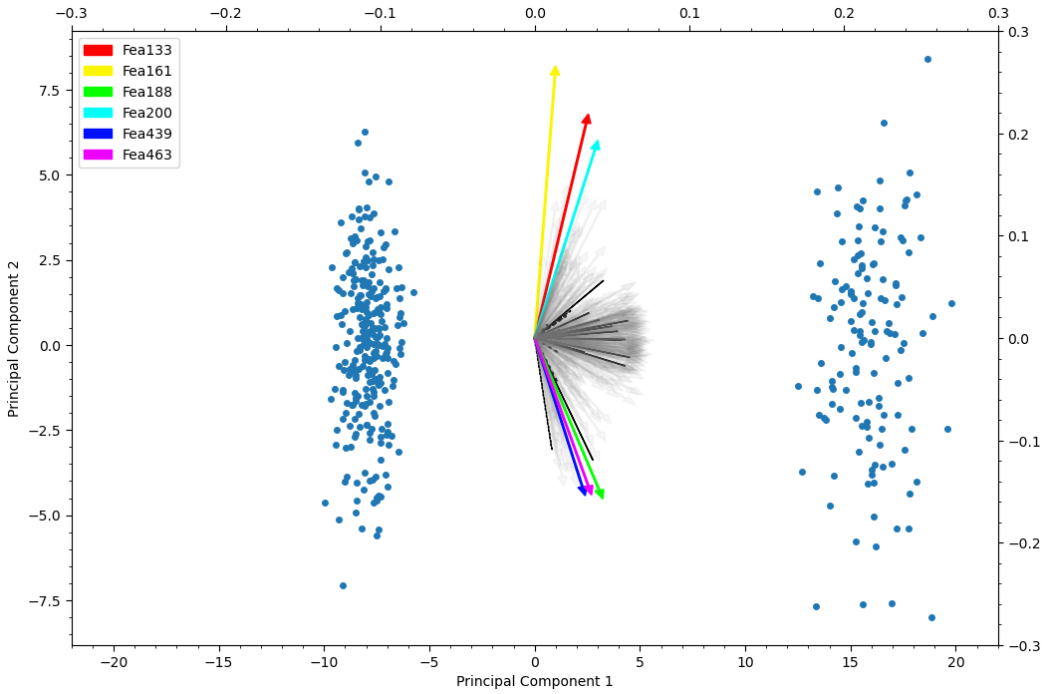

Figure 2: A biplot of the first two principal components in the `A_NoiseAdded.csv` dataset. The coloured arrows indicate the first two principal component loading vectors for every feature which contributes to either principal component more than 2%, these are the most discriminative features. Every other loading vector has been plotted in very light grey so their general directions and magnitudes are visible. The loading vectors for the first 20 features have also been plotted in darker black lines, with the dotted black lines corresponding to the dotted KDEs in Fig(1). The loading vectors use the right and top axis of the plot. The blue dots indicate the scores for each observation in the dataset for the first two principal components. The scores use the left and bottom axis of the plot. All four axis are symmetric about the origin for ease of comparison.

1.2 Dataset A

1.3 Dataset A

2 Section B

References

[1] scikit-learn developers, *`scikit-learn` k-means documentation*. Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means [Accessed: 13-Dec-2023].

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Total
Cluster 1	**43**	0	9	8	0	4	14	1	**79**
Cluster 2	0	**185**	0	0	0	0	0	0	**185**
Cluster 3	0	87	**0**	0	0	0	0	0	**87**
Cluster 4	12	0	2	**5**	0	3	6	0	**28**
Cluster 5	1	0	0	0	**0**	0	0	0	**1**
Cluster 6	6	0	0	0	0	**0**	0	0	**6**
Cluster 7	12	0	0	0	1	1	**7**	0	**21**
Cluster 8	1	0	0	0	0	0	0	**0**	**1**
Total	**75**	**272**	**11**	**13**	**1**	**8**	**27**	**1**	

(a) A contingency table for two k-means clusterings of the `A_NoiseAdded.csv` dataset, with $k = 8$, the default `scikit-learn` value. 240 of the 408 observations lie on the leading diagonal.

	Cluster 1	Cluster 2	Cluster 3	Total
Cluster 1	**155**	117	0	**272**
Cluster 2	0	**0**	10	**10**
Cluster 3	0	0	**126**	**126**
Total	**155**	**117**	**136**	

(b) A contingency table for two k-means clusterings of the `A_NoiseAdded.csv` dataset, with $k = 3$. 281 of the 408 observations lie on the leading diagonal.

Figure 3: Contingency tables for two k-means clusterings of the `A_NoiseAdded.csv` dataset with number of clusters $k = 3$ and $k = 8$. Each feature in the dataset was standardised before clustering. `kmeans_1` totals are on the right and `kmeans_2` totals are on the bottom.