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 Question 1c

Fig(3) shows the contingency table for two k-means clusterings of the dataset. The two clusterings were formed by training two k-means models on half of the dataset, and then the other half were mapped onto the learned clusters. K-means works computing a centroid for each cluster such that the sum of the squared distances between each centroid and the observations in the cluster is minimised, there are several algorithms for computing centroid positions. Predictions are then made by assigning new observations to the cluster whose centroid is closest to the observation [1].

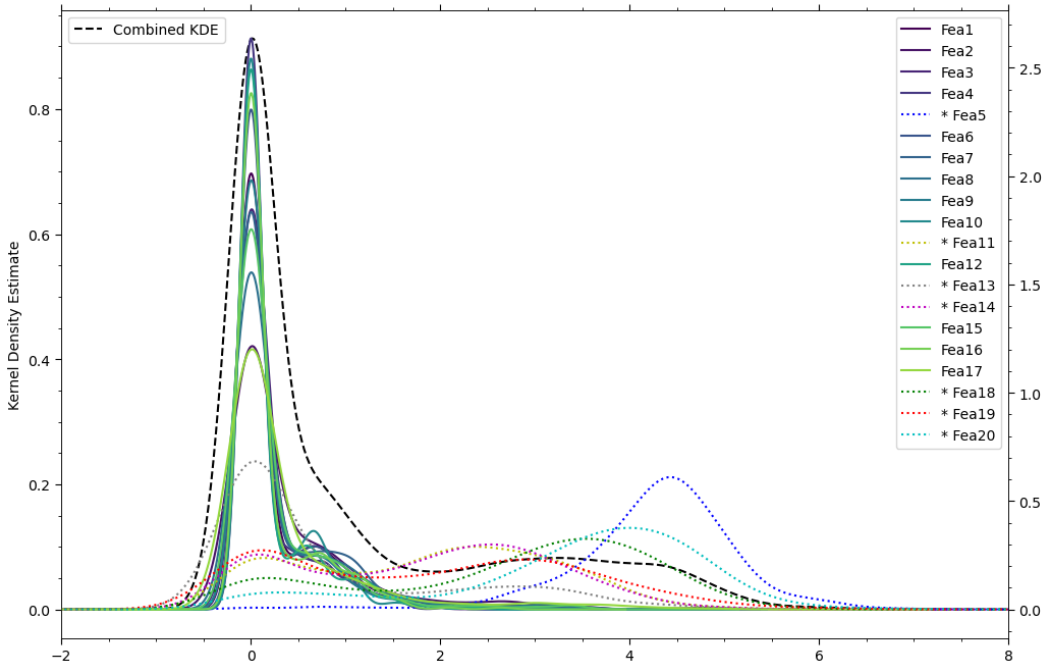

Figure 1: A kernel density estimate of the first 20 features of the `A_NoiseAdded.csv` dataset, which has 20 plots in total, with the legend and corresponding y-axis on the right. A combined KDE has also been plotted, with it's corresponding x-axis on the left. 7 of the 20 features have been highlighted 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.

The contingency table indicates that the two clusterings are very dissimilar as the majority of the observations lie outside the leading diagonal. This indicates that $k = 8$ is a bad choice for this dataset - the dataset is labelled and as such it is known that $k = 3$ for this dataset, so a poor clustering as indicated by the contingency table is expected. The contingency table also indicates that the clustering is not very stable, as the majority of the observations

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].

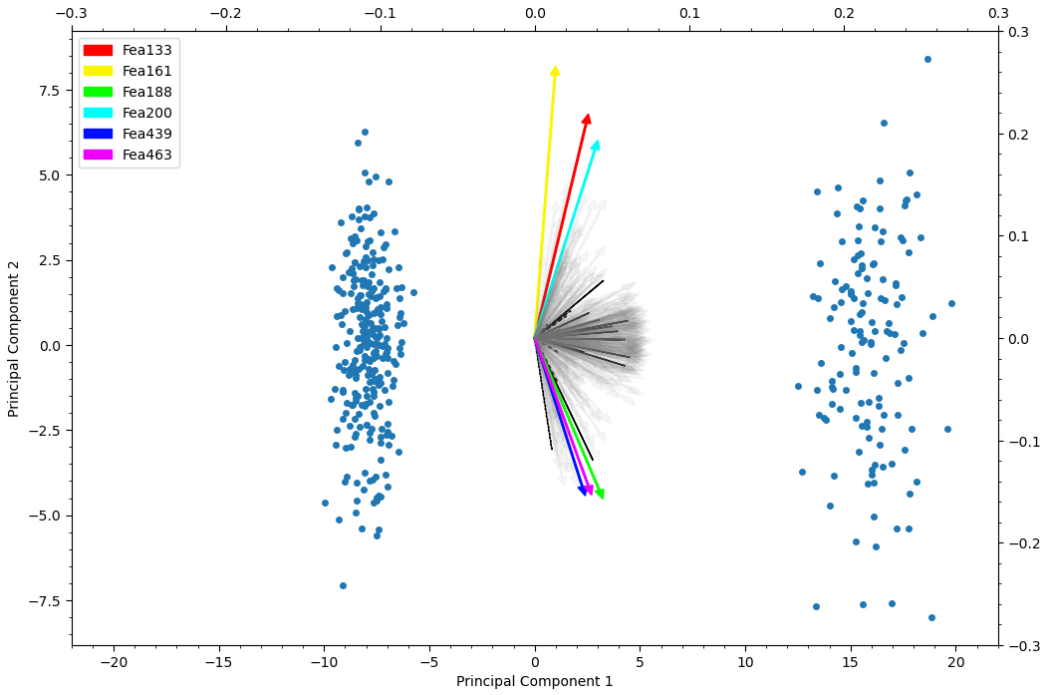

Figure 2: A biplot of the first two principal components in the `A_NoiseAdded.csv` dataset. The coloured arrows indicate the first two principal component loading vectors for every feature which contributes to either principal component more than 2%, these are the most discriminative features. Every 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.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster 1	**0**	8	0	4	1	43	9	14
Cluster 2	185	**0**	0	0	0	0	0	0
Cluster 3	87	0	**0**	0	0	0	0	0
Cluster 4	0	5	0	**3**	0	12	2	6
Cluster 5	0	0	0	0	**0**	1	0	0
Cluster 6	0	0	0	0	0	**6**	0	0
Cluster 7	0	0	1	1	0	12	**0**	7
Cluster 8	0	0	0	0	0	1	0	**0**

Figure 3: A contingency table for two k-means clusterings of the `A_NoiseAdded.csv` dataset, with $k = 8$, the default `scikit-learn` value.