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

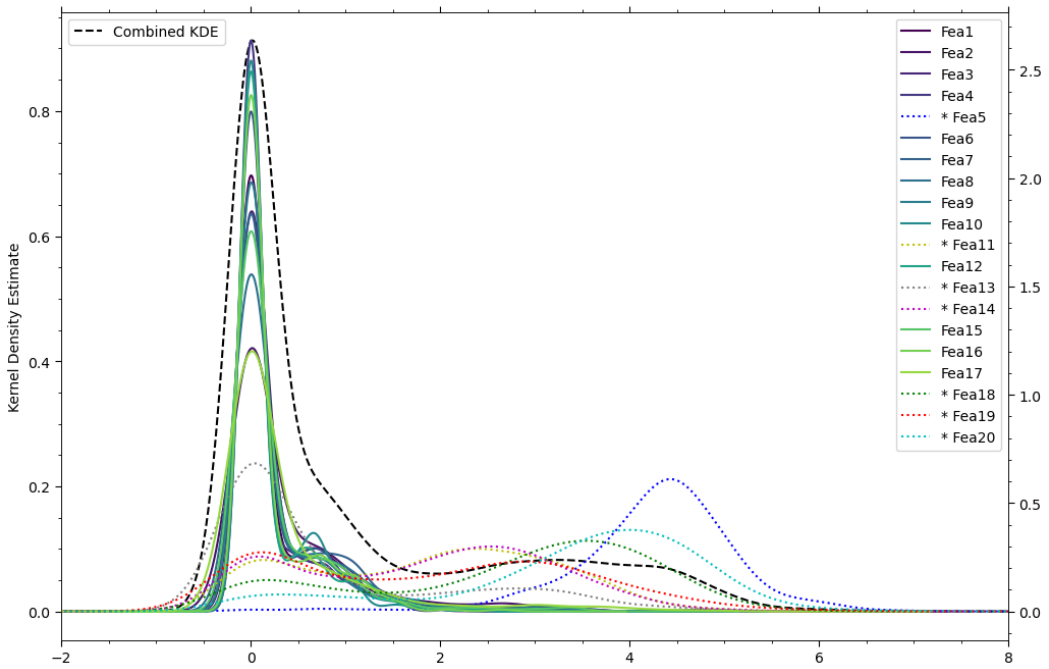
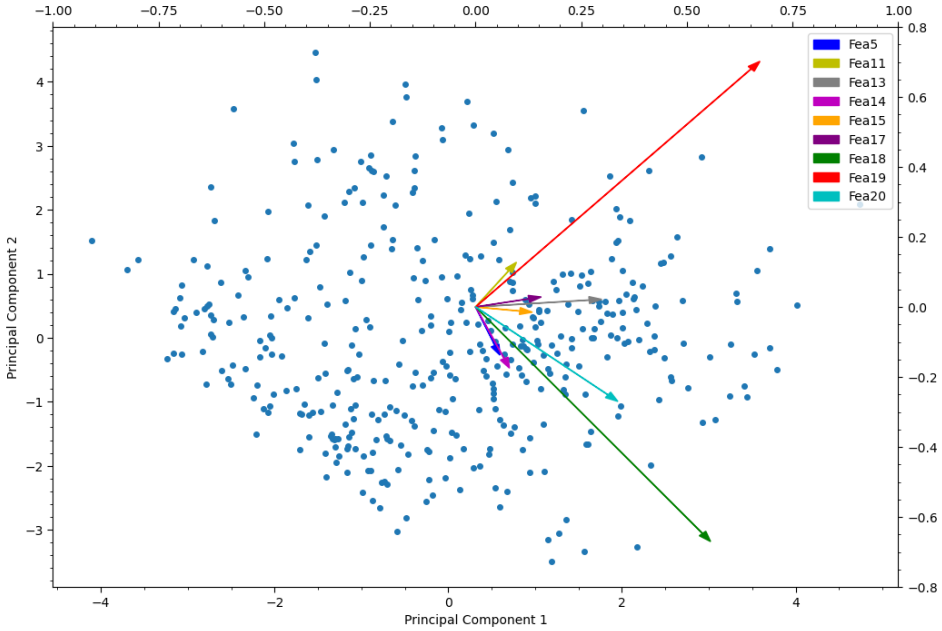


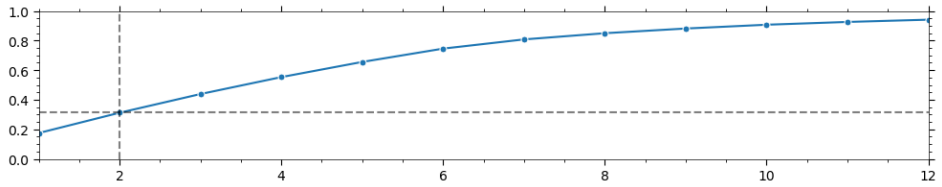
Figure 1: A kernel density estimate of the first 20 features of the `A_NoiseAdded.csv` dataset, which has 20 plots in total, with the legend and corresponding y-axis on the right. A combined KDE has also been plotted, with it's corresponding x-axis on the left. 7 of the 20 features have been highlighted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

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



(a) A biplot of the first two principal components in the dataset. The arrows indicate the first two principal component loading vectors for every feature which contributes to either principal component more than 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.



(b) A line plot of the explained variance ratio for the first 12 principal components.

Figure 2: Two plots illustrating the results of a PCA of the first 20 features of the `A_NoiseAdded.csv` dataset.

Fig(1) indicates that the features are all on a similar scale, so the features do not need to be scaled before applying PCA. The biplot in Fig(2a) shows a PCA of the first 20 features of the dataset, as predicted, the 7 features identified in the Fig(1) are the most discriminative, as they are amongst the 9 features in the biplot that contribute more than 1% to either principal component. Features 18 and 19 are the most discriminative, as they contribute the most to the principal components. However, Fig(2b) shows the explained variance ratio for

the first 12 principal components, and it can be seen that the first 2 principal components only explain 31.2% of the variance in the dataset.

1.2 Dataset A

1.3 Dataset A

2 Section B

References