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

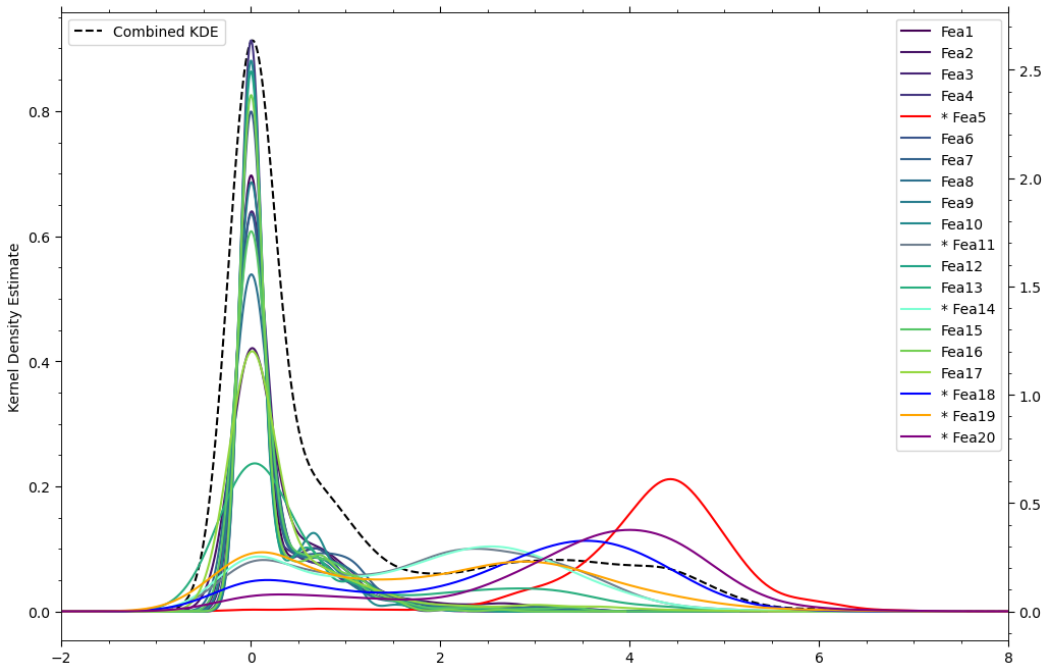

Figure 1: A kernel density estimate of the first 20 features of the dataset, which has 20 plots in total, with the legend and corresponding y-axis on the right. A combined KDE has also been plotted, with it's corresponding x-axis on the left. 6 of the 20 features have been highlighted in the legend with an asterisk, and have been coloured in slightly more contrasting colours to highlight them. These features contain more density in the non-zero region and are 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 the exception of features 5, 11, 14, 18, 19 and 20. These features are likely to be more discriminative, as they contain more variability. Feature 5 is concentrated around a value of 4.5, whilst the other 5 of these discriminative features are more widely spread around the range of values, which gives them more potential to be more discriminative in classification algorithms.

1.1.2 Question 1b

Fig(1) indicates that the features are all on a similar scale, so the features do not need to be scaled before applying PCA.

1.2 Dataset A

1.3 Dataset A

2 Section B

References