Foundations of Machine Learning - Lab 6

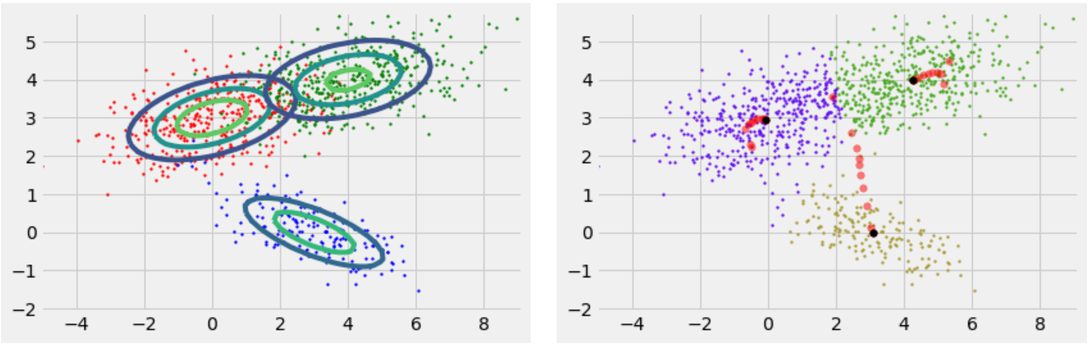
Johnny Joyce (jvjj1u19@soton.ac.uk)

December 5, 2019

The previous machine learning labs have been completed and the feedback from labs 1-3 have been taken into account. Feedback on lab 3 asked how the ROC curve for the Mahalanobis classifier was calculated. This was done by setting a threshold ranging from the minimum of the Mahalanobis distances of our data to the maximum of Mahalanobis distances. The threshold was varied over this interval as the threshold for classification in either class and the percentage of true positives and false positives was plotted

1 K-Means Clustering

1.1 Implementing K-means



(a)

(b)

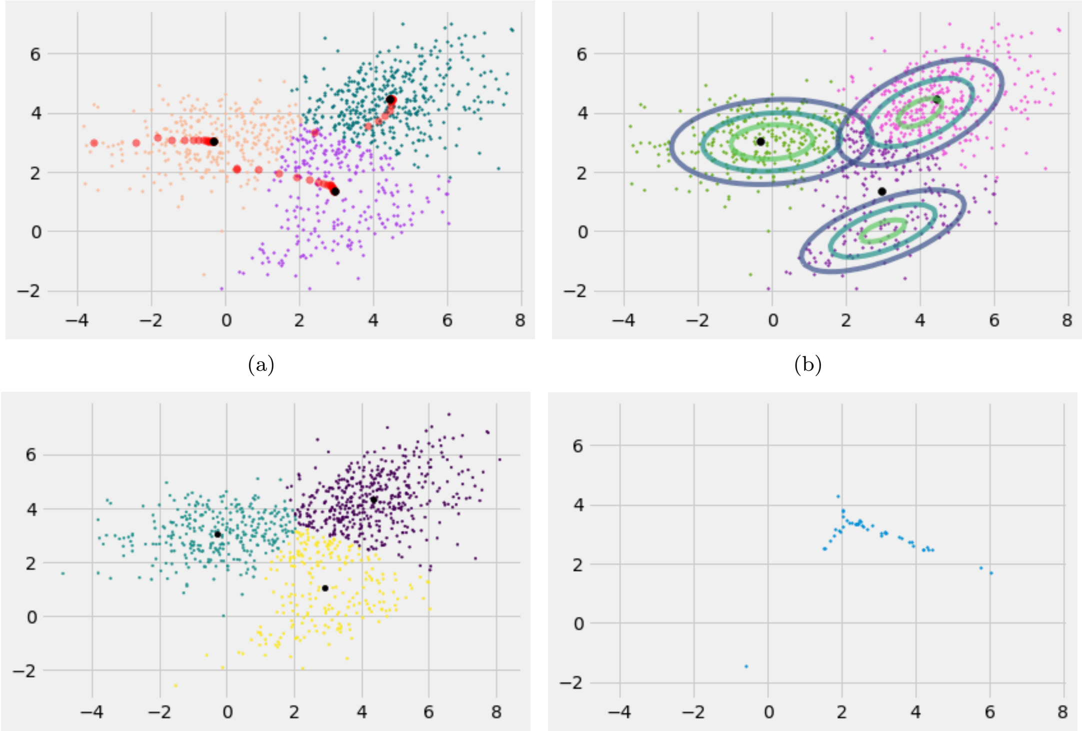
Figure 1: (a) Plot of three bivariate Gaussian distributions and their respective contours (b) The results of K-means clustering on said distributions. Black coordinates indicate cluster centres, red coordinates indicate previous cluster centres, and all other colours indicate clusters. 913 of 999 coordinates were correctly classified.

1.2 Comparing with contours and comparing with sklearn

Figures 2(a) and 2(c) both show fairly similar results. However, by examining Figure 2(d), we see that there are coordinates on the boundary lines of the clusters to which the two algorithms assigned different clusters. The sklearn implementation scored better than the self-implemented version, correctly classifying 801 of 998 coordinates compared to the self-implementation's 831 of 998.

1.3 Dependence on initial guesses

We can now investigate whether the performance of K-means algorithm depends on the initial guesses for cluster centres. To do so, let us use the distributions in Figure 3(a). We can see visually that the algorithm performed significantly worse in Figure **??**(b) (652 of 998) than in Figure 3(c) (906 of 998).



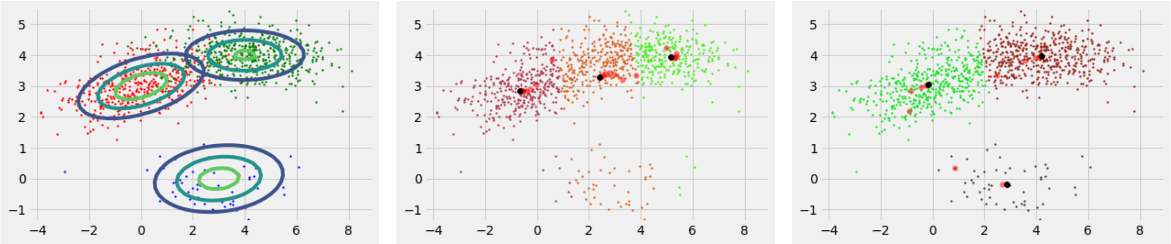
(a)

(b)

(c)

(d)

Figure 2: (a) The results of running the self-implemented K-means algorithm on a new set of bivariate Gaussian distributions (b) The results of the self-implemented K-means with the contours of the Gaussians (c) The results of sklearn's K-means algorithm (d) The coordinates where sent to different clusters by the self-implemented K-means compared to the sklearn's K-means.



(a)

(b)

(c)

Figure 3: (a) A new selection of Gaussian distributions and their contours (b)(c) The results of the self-implemented K-means algorithm with different starting guesses on the new distributions