Homework 5 & 6

Andrew Lei
Arizona State University

December 1, 2016

1. Weights for the first two components of PCA can be found in 'data.txt'. Variance explained by component one was about 36.2%. Variance explained by component two about 19.2%. Total of first two about 55.4%.

   The true categories for the wine look like this:

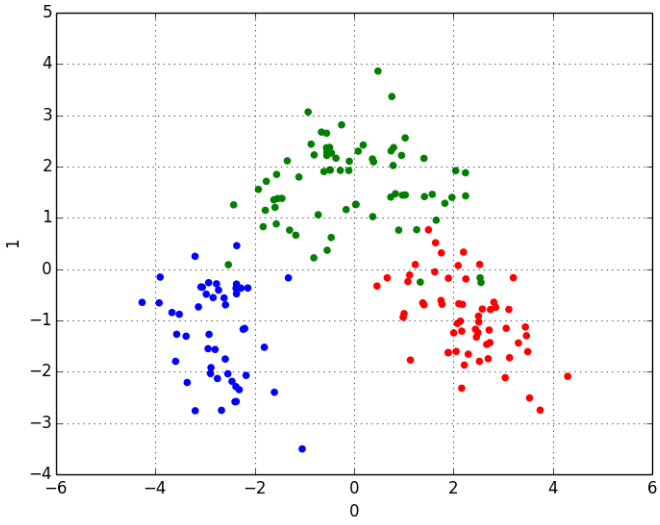


Figure 1: First two principal components of wine data, colored by type.

2. Here is the result of running kmeans once:

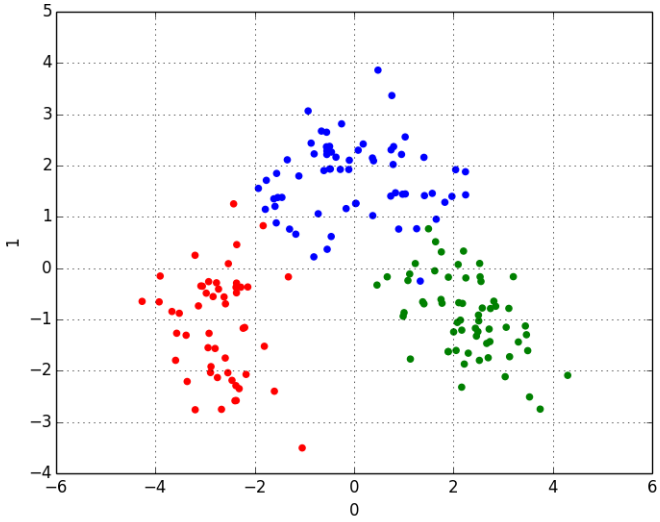


Figure 2: First two principal components of wine data, clustered using K-means.

Apart from the different colours, this is quite similar to the above. The points that are different were extracted:

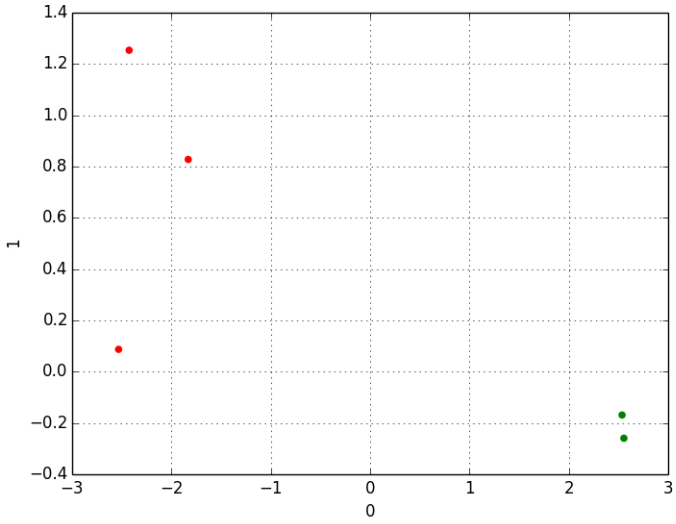


Figure 3: First two principal components of wine data, difference between points clustered using K-means and true clusters.

Images from the ten runs of K-means can be found with the titles of the form 'hw5pt2no*.png'. Here is the run with the highest BSS (equivalent to lowest within-cluster-distance):

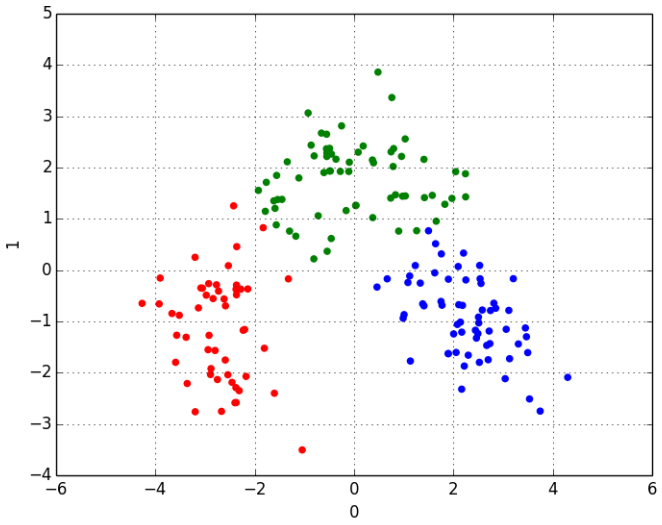


Figure 4: First two principal components of wine data, clustered using K-means; best of 10.

And its errors:

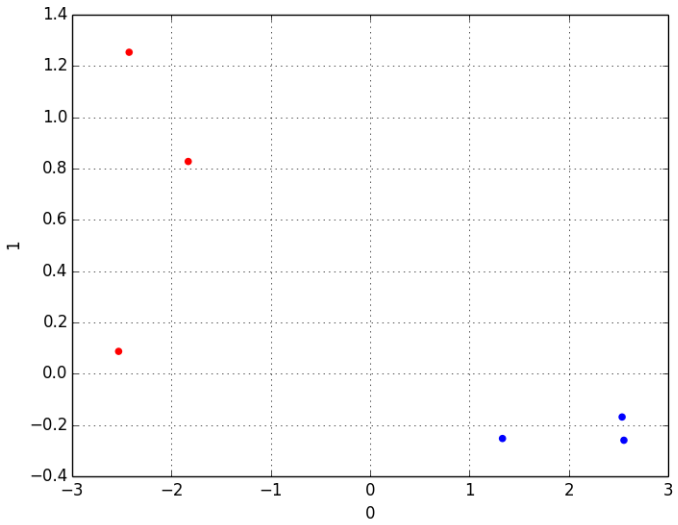


Figure 5: First two principal components of wine data, difference between points clustered using K-means and true clusters; best of 10.

3. I decided to run these with the same initial starting points as in part 2. Note that in 'data.txt' the BSS for these trials are identical to the trials in part 2; it seems Sci-Kit Learn implements K-means in a manner that results in an identical answer. So while the images are available with the names 'hw5pt3no*.png', they don't need much discussion since they are identical to those of the previous section.

   Note that I have checked the list the keeps track of the initial centroid locations before and after running K-means on it in section 2 to make sure nothing was changed by reference there, so that the initial values used in this section were indeed the same as in section 2.

4. Single-link was quite bad at clustering the data; most of the data ended up in one large cluster:

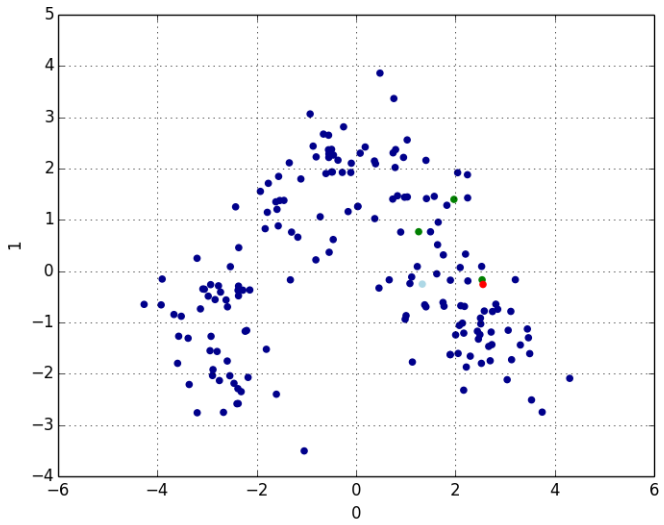
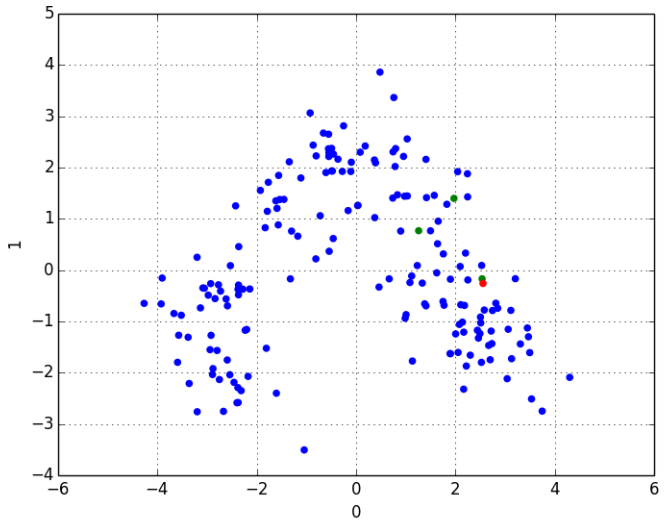


Figure 6: Above: Single-link clustering with four clusters. Below: Single-link with three.



By comparison, complete-link got us closer to the true values:

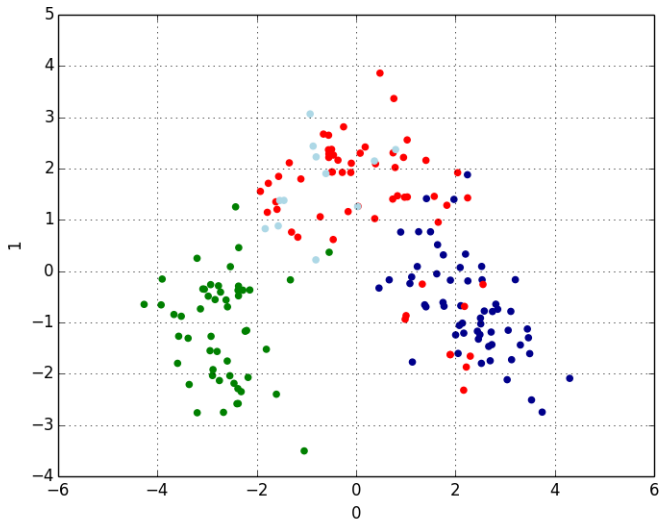
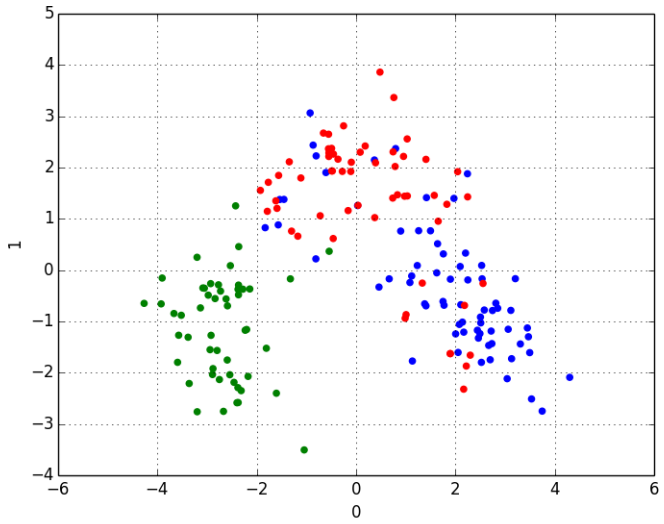


Figure 7: Above: Complete-link clustering with four clusters. Below: Complete-link with three.



However, if you note the substantial overlap between the reds and the blues here, it is immediately obvious that this is somewhat poorer than K-means.