Homework 5: kMeans

CS412

Released: April 23rd

Due: April 30th, 11:30pm on Gradescope

1 kMeans

For this assignment, you will use sklearn.cluster.kMeans on the digits data. For this, you will use all ten digits and you should the entire dataset (training + testing). This is the 256D data, but should not have any of the actual digit labels attached to it. You will need to alter this data on your own.
Complete the following operations using kMeans and the following parameters. Let `init` be "random" for all values here.

a) Plot a graph where the x axis is n clusters as the numbers between 2 and 20. Let the y axis be `inertia` which is a property of the kMeans object after the fit. **Label this Figure 5.1**.

b) Based on your experimental data, which value for n clusters is the best application for the data? Does this match your expectations or not?

c) Now repeat the graph from a) where n init = 1 and max iter = 1. **Label this Figure 5.2**

d) Explain the differences, if any, between these two graphs.

e) Let n init and max iter be default again. Plot a graph where the x axis is n clusters from 2-20 and the y-axis is n iter. **Label this figure 5.3**

f) How do we expect the number of iterations to be affected by our number of clusters? Does the data match your expectations, why or why not?

g) **Graduate Student Question** Plot the data points as classified by the kMeans clusterer fit predict onto your 2D features from HW1. Now that you are considering data beyond only ones and fives, do the two features you initially selected in HW1 still seem to effectively separate the data? Explain why or why not.

Extra Credit

Using the 10 kMeans clustering approach from the portion above, try to match the clusters to an appropriate digit. Compare the accuracy of this unsupervised learning to the accuracy of supervised approaches and report your results. Then try different values for k to see find the value that gives the highest unsupervised learning accuracy.