9.3 Clustering

Brandon Sams

2/9/2020

Overview

These assignments are here to provide you with an introduction to the "Data Science" use for these tools. This is your future. It may seem confusing and weird right now but it hopefully seems far less so than earlier in the semester. Attempt these homework assignments. You will not be graded on your answer but on your approach. This should be a, "Where am I on learning this stuff" check. If you can't get it done, please explain why.

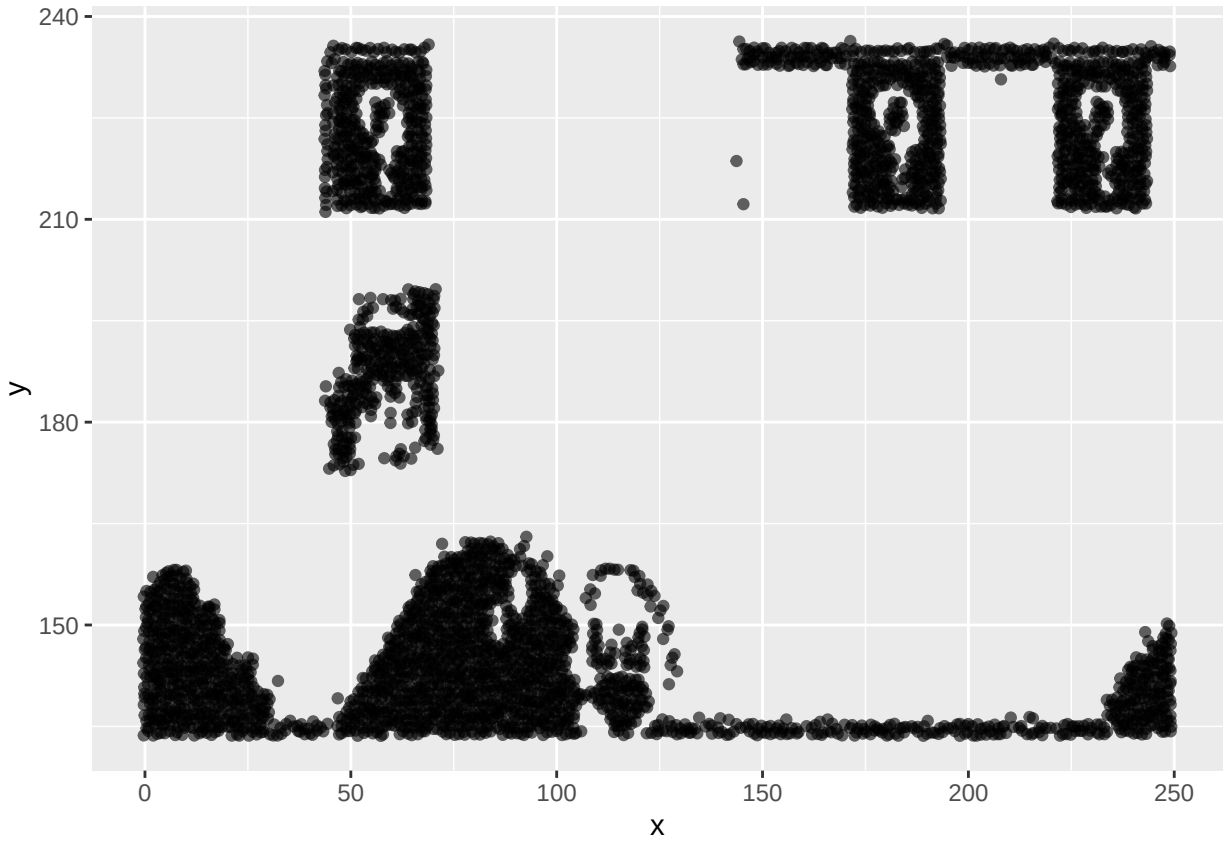
Remember to submit this assignment in an R Markdown report.

Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data. In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv.

```
cluster_data <- read.csv(url("http://content.bellevue.edu/cst/dsc/520/id/resources/clustering-data.csv")
```

(a) Plot the dataset using a scatter plot.

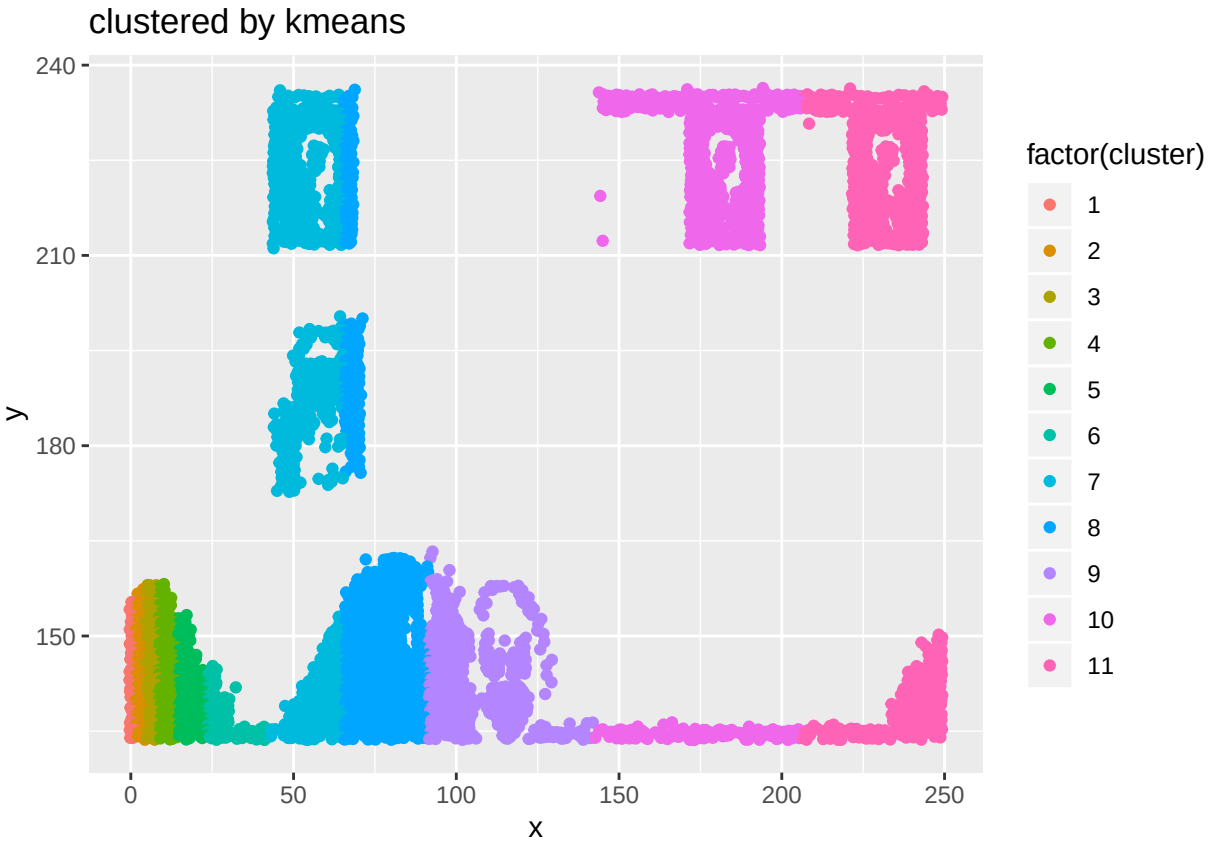
```
ggplot(cluster_data, aes(x=x,y=y)) + geom_jitter(alpha=0.6)
```



(b) Fit the dataset using the k-means algorithm from $k = 2$ to $k = 12$. Create a scatter plot of the resultant clusters for each value of k.

```
km <- kmeans(cluster_data$x,2:12)
#summary(km)
clustered_data <- data.frame(x=cluster_data$x,y=cluster_data$y,cluster=km$cluster)

ggplot(clustered_data,aes(x=x,y=y,color=factor(cluster))) + geom_jitter() + ggtitle("clustered by kmeans
```



(c) As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.

No idea how to approach this one.

Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.

No idea how to do the previous part.

(d) One way of determining the "right" number of clusters is to look at the graph of k versus average distance and finding the "elbow point". Looking at the graph you generated in the previous example, what is the elbow point for this dataset?

Im lost on this one too.