STAT 240: Introduction to data science

Lab 08: kmeans

Lloyd T. Elliott

Question 1

kmedians is like kmeans, except with centres chosen according to medians.
You didn't see kmedians in class: it is explained in this question!

- In this question, you will implement kmedians with L1 distances.
- Write an *R* function *kmedians* that takes a numeric data frame *x* (with one row per data item, and one column per dimension *D*), and a positive integer *K* (for the number of clusters), and a positive integer *iters* for the number of iterations.
- The function should be a modification of the kmeans algorithm, except:
  - 1) In the step where the cluster locations are updated, set the cluster locations to be the medians of the data items assigned to the cluster (by setting each dimension of the cluster location to the median of the values for that dimension of the data items assigned to that cluster).
  - 2) In the step where data items are assigned to the nearest cluster, assign them to the nearest cluster based on their L1 distance to the cluster locations (instead of their Euclidean distance).
- Your function should return a list with an element named "locations" providing a *K x D* table showing the cluster locations, and "assignments" providing a vector of length *N* showing the cluster assignments at the last iteration. Provide your code.

(5 marks)

Question 2

- Further modify your code so that instead of initializing the cluster assignments of the data items randomly, instead assign data item *i* to cluster *((i-1)%K)+1*. (i.e., if *K = 3* and there are 7 data items, the initialization assigns the data items in order to 1, 2, 3, 1, 2, 3, 1).

- Run this further modified version of kmedians on the *parkinsons* dataset with *K = 3*, and *iters = 1000*. Report the locations of the 3 clusters.

(5 marks)

Question 3

- Modify your code to match the usual kmeans algorithm (and rename the function to *mykmeans*).
- Run your *mykmeans* algorithm with $K = 3$ on the *parkinsons* data.
- Run *R*'s built in kmeans function on the *parkinsons* data, with $K = 3$.
- Explore the differences between the two implementations with some figures or examinations, and explain why these differences may arise.

(5 marks)

Question 4

- List three benefits of kmeans and three drawbacks.

(6 marks)

Question 5

- Consider the following two points:
  - The kmeans algorithm is guaranteed to converge after a finite number of iterations (that is, if *iters* is large enough, eventually the two steps of the kmeans iterations will not change the cluster locations or the cluster assignments of the data items).
  - If the kmeans algorithm is run twice with two different random initializations, then the solutions the two runs converge to may be different.
- Create a (small?) dataset that demonstrates the phenomenon whereby two runs with two different random initializations converge to two different solutions. You may create the dataset any way you want, and fix the random initializations any way you want. Provide the dataset and the initializations and code and plots demonstrating the difference.

(2 marks)