Clustering Lab/Homework (Lab/Homework 8)

About the Data

We will analyze data on the sales of orthopedic devices to hospitals. The data set is hospital.csv on Blackboard.

This file contains information compiled by a company that sells orthopedic devices to hospitals. The observations are 4703 different U.S. hospitals, and the variables are listed in the table below. The company would like to better understand its client base, for instance by grouping the client hospitals based on their operational and financial characteristics and understanding what attributes define those groups.

The data set is a comma-delimited text file. Each row represents a hospital and each column a particular demographic or response variable. Here are the meanings of the variables:

Variable Name	Description
ZIP	US postal code
HID	Hospital ID
CITY	City name
STATE	State
BEDS	Number of hospital beds
RBEDS	Number of rehab beds
OUTV	Number of outpatient visits
ADM	Administrative cost (thousands of $'s per year)
SIR	Revenue from inpatient
SALESY	Sales of rehab. equipment since Jan. 1
SALES12	Sales of rehab. equip. for last 12 months
HIP95	Number of hip operations for 1995
KNEE95	Number of knee operations for 1995
TH	Indicator of teaching hospital
TRAUMA	Indicator of having a trauma unit
REHAB	Indicator of having a rehab unit
HIP96	Number of hip operations for 1996
KNEE96	Number of knee operations for 1996
FEMUR96	Number of femur operations for 1996

The Task

Read the data set into R using the read.table command. Check that the data have been read in correctly.

We will not use any categorical variables or SALESY or SALES12 for clustering. K-means cannot handle categorical variables, and the SALES variables are measurements of the sales to the hospital rather than the hospital itself. Remove these variables from the data set (make sure to remove "Hospital ID", "ZIP", and all indicator variables), and call the resulting data frame "orthoClust".

Take the log transformation, like we did in class for this dataset: **orthoClust = log(orthoClust + 1)**. Then standardize the variables in the "orthoClust" object. Having done this, create pairwise scatterplots for all of the standardized variables. Are there clusters visible in the scatterplots?

Apply k-means using the "kmeans" function. Check the help file for "kmeans" to get more information about how to use it. Specify that you want 3 clusters, and let R randomly select the cluster centers.

Find the value of the loss function $W(C)$ for this encoder C, by calling **sum(orthoKmeans$withinss)** where **orthoKmeans** is the object returned by the "kmeans" function. Try calling the "kmeans" function fifteen or twenty times, and each time evaluate $W(C)$. Usually it will be the same but you should find that sometimes you get a different value of $W(C)$. This is because the kmeans function uses random initial values, and can converge to different local minima of the loss function depending on the initialization.

What is the smallest value of $W(C)$ that you obtain? This corresponds to the "best" encoder C. Keep the **orthoKmeans** object that has the lowest value of $W(C)$. Then interpret the three clusters, by looking at the locations of the cluster centers as we did in class. To get the cluster centers, call **orthoKmeans$centers**. **How would you characterize cluster 1, in terms of the values of the variables? How would you characterize the other clusters?**

Check the sizes of the clusters by calling **orthoKmeans$size**. Now, call the "kmeans" function again several times, until you get the smallest value of $W(C)$ again. Check the sizes of the clusters again. Has the ordering of the clusters changed? This happens because the cluster labels "1", "2", and "3" have no particular meaning; the objective function does not take into account the labels of the groups, so the different possible labelings are all equally valid and have the same value of $W(C)$.

The "orthoKmeans$cluster" object is a vector containing the cluster assignments (integer values 1, 2, or 3) for the hospitals. Create the pairwise scatterplots again, but this time color the points in the plots according to their cluster assignment. This can be done by specifying **col = orthoKmeans$cluster** when calling the "pairs" function. Remember that the "col" argument specifies the color of the points. The resulting plot gives you a visual understanding of what the clusters represent; compare to the cluster descriptions that you obtained earlier.

We specified $K = 3$ clusters when we called "kmeans", but this was an arbitrary choice. Try different values for the number of clusters, and try the above exercises again. See how your results change as you change K.

Clustering Homework (HW 8)

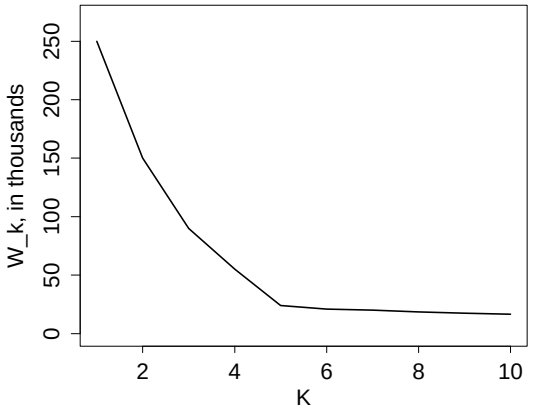
1. Report your answers to the questions in bold in Lab 8.

2. (From the Suh Text) Consider the following table on average weekly household alcohol and tobacco spending in pounds for different regions of Great Britain.

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92

Cluster these regions using the two variables: alcohol and tobacco spending, by applying the k-means algorithm by hand. Initialize the clusters by assigning North, Yorkshire, and Northeast to cluster 1 and the other regions to cluster 2. Show your work for each iteration of the algorithm, and report the final cluster assignments, cluster centroids, and value of the objective function $W(C)$.

3. Below is a plot of the within-cluster variability W_K as a function of K for the clusters obtained using K-means clustering (up to $K = 10$ clusters were evaluated). Based on this information, what number of clusters would you choose and why?



4. We pointed out in class that for n training observations and K clusters, the number of encoders C is K^n, if we allow empty clusters. However, two different encoders are equivalent if they have exactly the same groups but the labels associated with the groups are different;

for instance, if cluster 1 in the first encoder corresponds to cluster 2 in the second encoder and vice versa. **For $K = 2$ and $n = 4$, exactly how many unique (non-equivalent) encoders C are there (allow empty clusters)? Justify your answer.**

5. Suppose that we have performed K-means clustering with $P = 2$ variables and $K = 2$ clusters. The estimated cluster means are $m_1 = (0.50, -0.12)$ and $m_2 = (-0.27, 0.69)$. **What is the formula of the line that separates cluster 1 from cluster 2? Will a new point at $(-0.01, 0.11)$ be predicted to be in cluster 1 or cluster 2?**

6.

(b) True or false: the clusters resulting from application of K-means with $K = 5$ are always nested within the clusters resulting from application of K-means with $K = 4$, meaning the following. One of the four clusters from the $K = 4$ case splits into two in order to obtain the clusters for the $K = 5$ case.

(c) True or false: the dissimilarity measure D used in K-means clustering is symmetric, so that $D(w, v) = D(v, w)$ for $w, v \in \mathbb{R}^P$.

7. Is K-means clustering translation invariant, in the sense that if we add a constant amount b to a particular variable before application of K-means, it will not affect our final cluster assignment? Why or why not? Assume that k-means is initialized at the same value of C regardless of the translation.