K-Means Clustering

Clustering and Classification

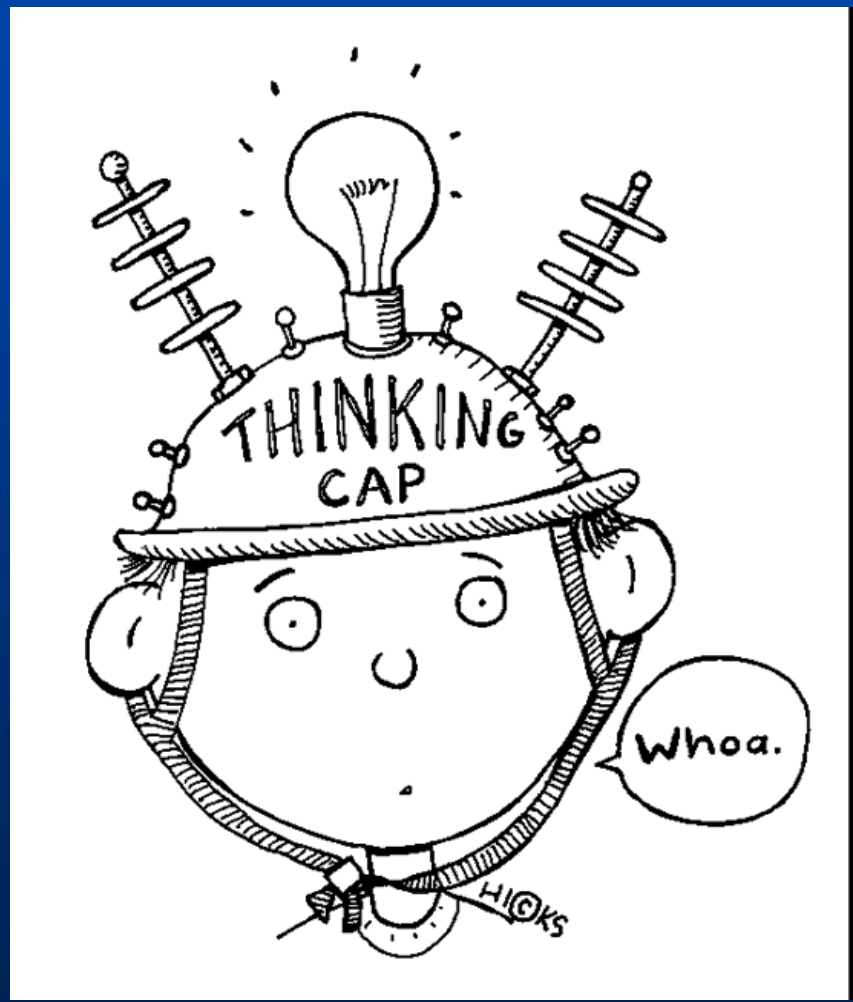
Lecture 8

Today's Class

- K-means clustering:
  - What it is
  - How it works
  - What it assumes
  - Pitfalls of the method (locally optimal results)

From Last Time…

- If you recall the last few minutes of our last class.
- I implored you to consider how to take data from a single variable, and do an ANOVA
  - Just in this case, you had no idea what the grouping variable was.
- How would you assign the observations to groups?



ANOVA Model (Really, the GLM)

- The One-Way ANOVA model:

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- The model can be re-written as:

$$Y_{ij} = \beta_0 + \beta_{1j} X_{ij} + \epsilon_{ij}$$

  - Here X_{ij} is the indicator that observation i is a member of group j (for dummy coding).

Imagine If…

- Imagine I set the following parameters for the data:
  - $\beta_0 = 64$
  - $\beta_1 = 5$
  - $\sigma_\varepsilon^2 = 9$
- I will then use R to get some data…
- But I will not let you see what X happens to be.

Ok, So You Tell Me…

- How would you find what X was for each person?
- How would you find what the means for each group were?
- Note that the data we picked had group means of 64 and 69.
  - These numbers were selected to mimic the distribution of heights for American males and females.
  - If given a set of heights, can you give me:
    - The gender an individual height came from?
    - The mean for both genders?

K-means Clustering Algorithm

K-Means Clustering Algorithm

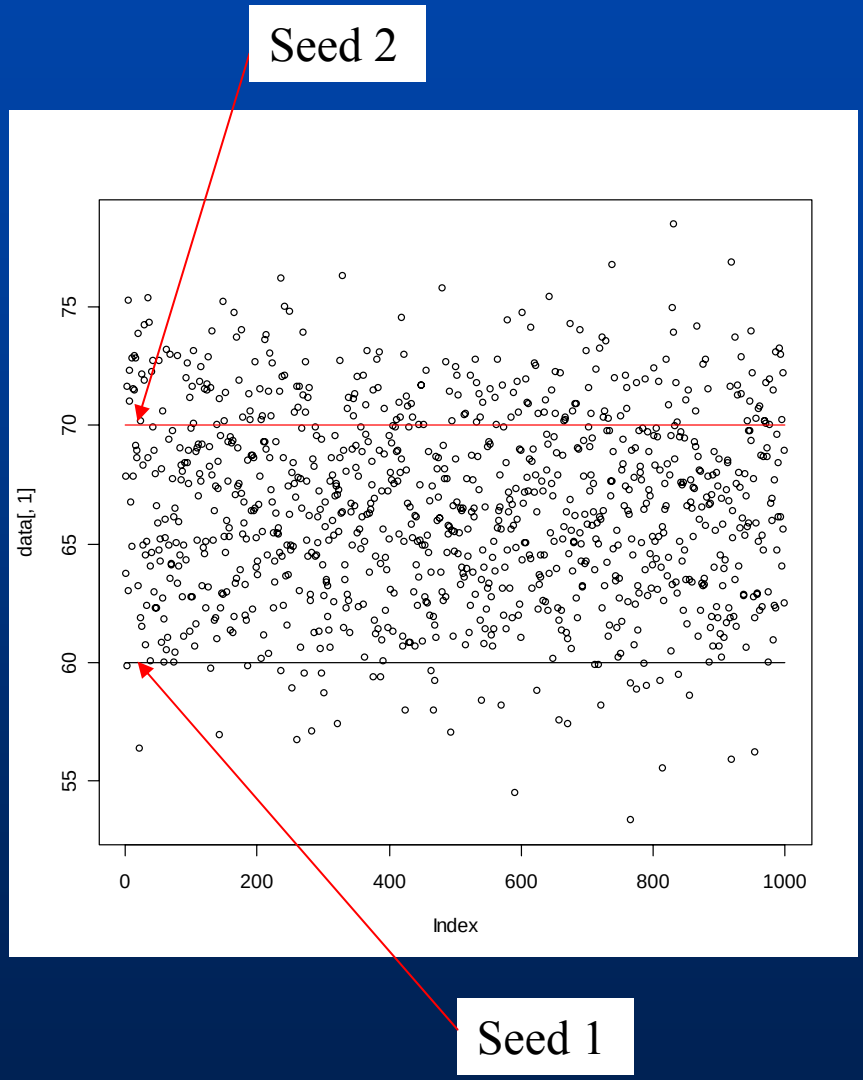
- The K-means clustering algorithm is a simple method for estimating the mean (vectors) of a set of K-groups.

- The simplicity of the algorithm also can lead to some bad solutions.
  - Sub-optimal clusterings of objects

K-means Algorithm Step #1

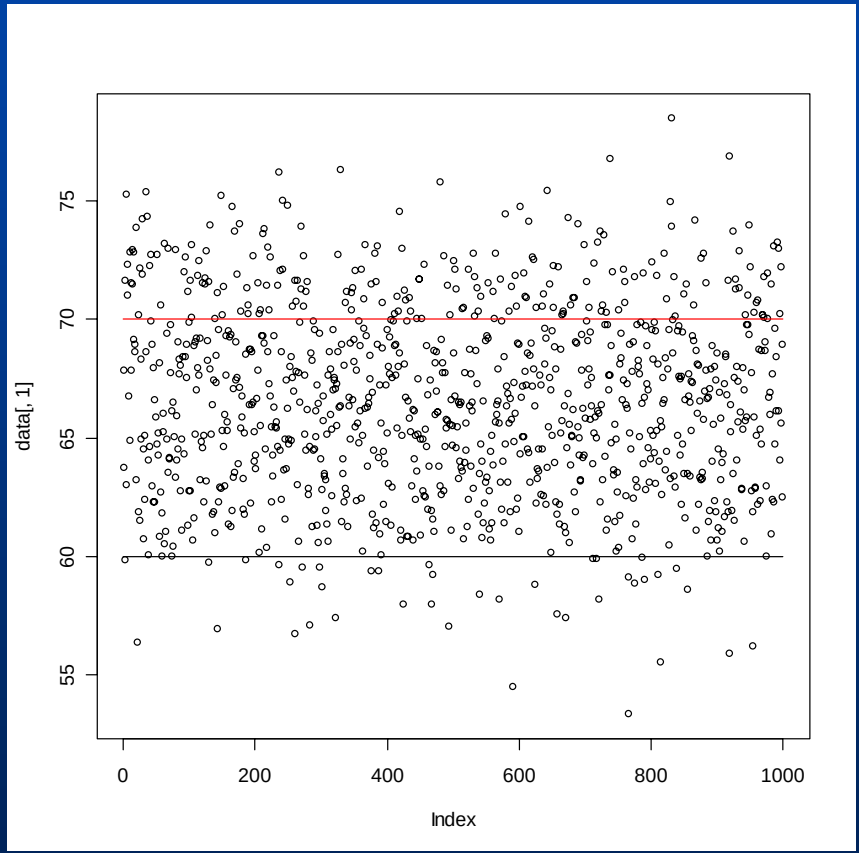
- A typical version of the K-means algorithm runs in the following steps:

1. Initial cluster seeds are chosen (at random).
   - These represent the "temporary" means of the clusters.
   - Imagine our random numbers were 60 for group 1 and 70 for group 2.



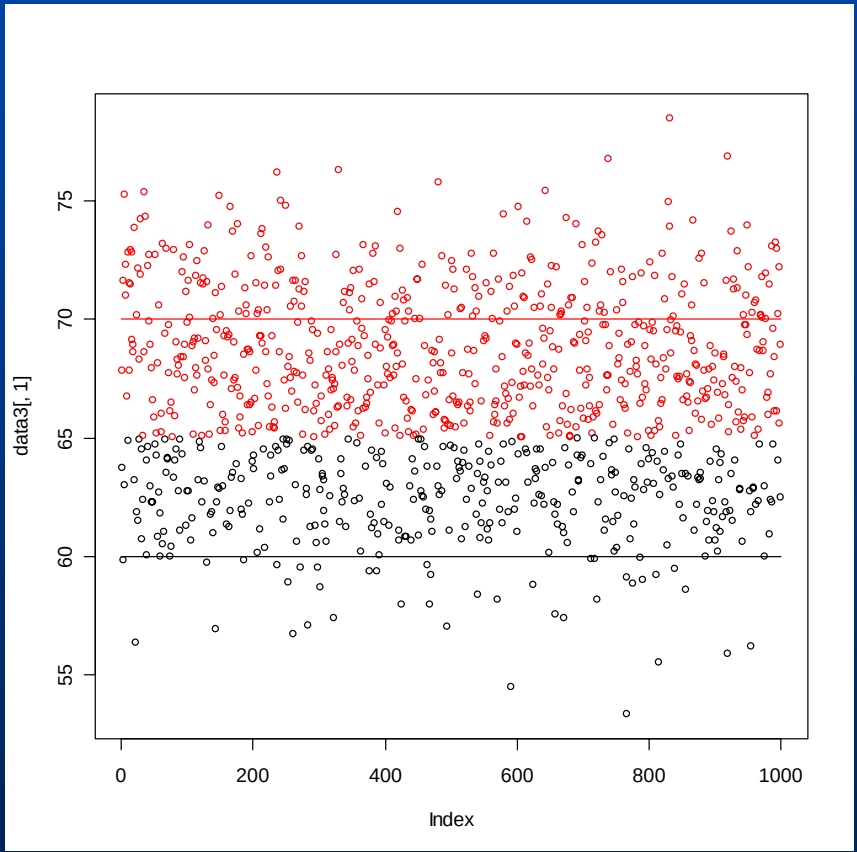
K-means Algorithm Step #2

2. The squared Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster.



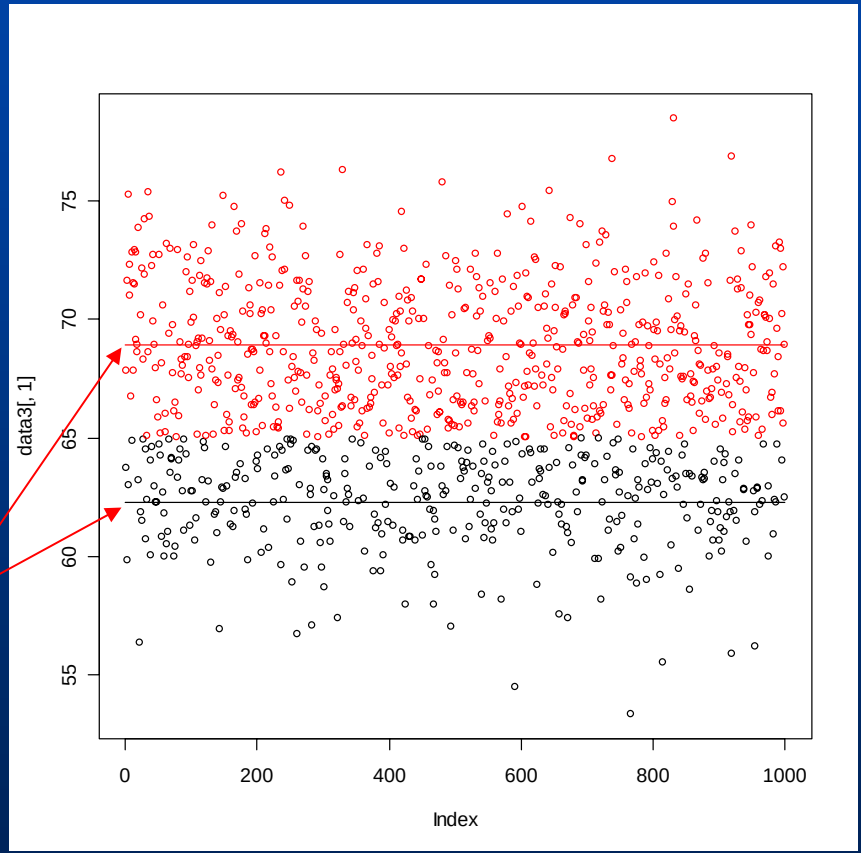
K-means Algorithm Step #2

2. The squared Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster.



K-means Algorithm Step #3

3. For each cluster, the new centroid is computed – and each seed value is now replaced by the respective cluster centroid.
   - The new mean for cluster 1 is 62.3
   - The new mean for cluster 2 is 68.9



K-means Algorithm Step #4 – #6

4. The squared Euclidean distance from an object to each cluster is computed, and the object is assigned to the cluster with the smallest squared Euclidean distance.

5. The cluster centroids are recalculated based on the new membership assignment.

6. Steps 4 and 5 are repeated until no object moves clusters.

K-means Assumptions

- K-means uses the squared Euclidean distance to allocate objects to clusters.
  - There is the implicit assumption that the data should have roughly the same scale to use such distances.

- Because the squared Euclidean distance is used, the K-means algorithm proceeds to try to find a minimum for the Error Sum of Squares (SSE).
  - The weak assumption is each group has roughly the same SSE – meaning that the variance/covariance matrix between objects within a group is equal.

Error Sum of Squares

- For the entire set of objects, the Error Sum of Squares is calculated by:

$$SSE = \sum_{j=1}^{p} \sum_{t=1}^{K} \sum_{i \in C_t} \left(x_{ij} - \bar{x}_j^{(t)} \right)^2$$

- The goal of K-means is to find a solution such that there are no other solutions with lower SSE.

- This commonly does not happen in a single run of the algorithm.
  - In our example, because we used a single variable and had two groups that were relatively non-overlapping, it will.

Wrapping Up

- K-means is a nice method to quickly sort your data into clusters.
  - All you need to know are the number of clusters you seek to find.

- Local optima in K-means can derail your results, if you are not careful.
  - Run the process many times with differing starting values.

Local Optima

- As discussed in the article by Steinley (2003), the solutions for K-means are not guaranteed to be globally optimal.

- Globally optimal = solution where SSE is smallest compared to ***any*** other possible solution.

- Solutions typically found in K-means are locally optimal – they have found the peak of a function in a small part of the space.

Local Optima Example

- Using code from Steinley (2003), we will now demonstrate the optimality problem in K-means using MATLAB.
- Steinley's code takes an input data set and runs K-means a number of times, recording the SSE at each time.
- The lowest SSE should be the one used.
- To demonstrate, I will use our data – with zero local optima, and the data set described on the next page.

Hartigan Data

- The data come from Hartigan (1975, p. 88) – part of your readings.
- The nutritional content of eight foods are listed with respect to energy, protein, and calcium.
- We will try putting these objects into three clusters.

K-means Estimation Process

- Part of the process of estimation of K-means is determining how many means are necessary.

- Hartigan (and others) suggest running (M)ANOVA following the solution to tell if the groups are significantly different.
  - If so, more clusters may need to be extracted – try with K+1
  - If not, too many clusters may have been extracted.

- This approach, however is at best a crude heuristic for determining the number of clusters.
  - There will be bias in the hypothesis test.

Summarizing Your Result

- To summarize your result, you can take the cluster means and describe the clusters.
- You can also examine the objects within each cluster to determine cluster meaning.
- Just note that if you find a local optima, both of these entities will change if you run the analysis again.

Wrapping Up

- K-means is a simple procedure for extracting clusters from data.
  - Sometimes simplicity is not good – if assumptions are violated.
- The K-means procedure will serve as an entry point into similar procedures in the future.
  - We will find that by using statistical distributions, we can achieve similar results with more flexible assumptions.

Next Time

- How to do K-means in R.
- Discussion of class project specifics.
- Presentation and discussion of an empirical research article featuring K-means.