CS753 Assignment 5: Cluster Analysis

Due on Nov 27, 2018

Notes:

- This is an independent, individual assignment. Group work is prohibited.
- Copying others' work is considered cheating.
- The homework is worth *4 points* toward the final grade.
- The homework will be due at the beginning of the class on the due date. No late submission is accepted.
- Submit the printout of your homework. Do NOT submit it via email.
- Make sure you have your name on the printout.

Breakfast Cereals for Kids

The dataset cereals.xls includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

- Data preprocessing
  a. Remove all records with missing values. Use the Filter Examples operator and choose no_missing_attribute for condition class.
  b. Select only attributes with numerical values. Note that although Cups and Shelf also are numerical, they are quite irrelevant to the problem. So you should leave them out as well.
  c. Normalize the data (z-transformation).

- Clustering
  a. Apply hierarchical clustering using Euclidian distance to the normalized data. Use Single Linkage and Complete Linkage respectively. Copy and paste the dendrograms. Which method produces more meaningful clusters? Why?
  b. How many clusters do you recommend? How many cereals are there in each cluster?
  c. Use the number of clusters you have recommended and apply the k-means method instead. Copy and paste the centeroid table. How many cereals are there in each cluster? How would you describe each cluster in terms of their "healthiness"? (Hint: unhealthy food usually is high in calories, fat, sodium and sugar, but low in protein).
  d. The elementary schools in Waltham would like to choose a set of cereals to include in their daily menu. Every day a different cereal is offered, but all cereals should support a healthy diet. Which cluster would you recommend as "healthy cereals" to the kids? Do you think the kids will like them? Why or why not?