Lecture 03. Clustering (Sections 12.1 and 12.4)

Ping Yu

HKU Business School
The University of Hong Kong

Introduction

- Most of this course focuses on supervised learning methods such as regression and classification.
- In that setting we observe both a set of features $X_1, X_2, \cdots, X_p$ for each object, as well as a response or outcome variable Y. The goal is then to predict Y using $X_1, X_2, \cdots, X_p$.
- Here we instead focus on unsupervised learning, we where observe only the features $X_1, X_2, \cdots, X_p$. We are not interested in prediction, because we do not have an associated response variable Y.
- The goal is to discover interesting things about the measurements: is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
- We focus on two methods:
  - Clustering, a broad class of methods for discovering unknown subgroups in data, will be discussed in this lecture. This is like the scenario where the group labels are missing in classification.
  - Principal Components Analysis (PCA), a tool used for data visualization, data preprocessing before supervised techniques are applied such as principal components regression (PCR), and data imputation, will be discussed in Lecture 5.

The Challenge of Unsupervised Learning

(Section 12.1)

The Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
  - It is often performed as part of an exploratory data analysis (EDA).
- It is hard to assess the results from unsupervised learning because we do not know the true answer – the problem is unsupervised.
- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories (e.g., Amazon),
  - movies grouped by the ratings assigned by movie viewers (e.g., Netflix).
- Another advantage: it is often easier to obtain unlabeled data – from a lab instrument or a computer - than labeled data, which can require human intervention.
  - For example, it is difficult to automatically assess the overall sentiment of a movie review: is it favorable or not?

Clustering Methods

(Section 12.4)

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
- To make this concrete, we must define what it means for two or more observations to be similar or different.
- This is often a domain-specific consideration, i.e., depending on the data being studied.
- Consider the application of clustering in market segmentation .
- Suppose we have access to a large number of measurement (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
- Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product. This amounts to clustering the people in the data set.

PCA versus Clustering

- PCA looks for a low-dimensional representation of the observations that explains a good fraction of the variance.
  - Reduce the number of columns.
- Clustering looks for homogeneous subgroups among the observations.
  - Divide rows into groups, and the number of columns does not change.[1]

[1] Of course, we can also cluster columns by transposing **X** below.

Two Best-Known Clustering Methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

History of K-Means Clustering

- The standard algorithm was first proposed by Stuart Lloyd of Bell Labs in 1957, but not published as a journal article until 1982.
- Edward W. Forgy published essentially the same method in 1965.
- So K-means clustering is sometimes referred to as the Lloyd–Forgy algorithm.
- Unfortunately, I cannot find photos for these two persons.

K-Means Clustering

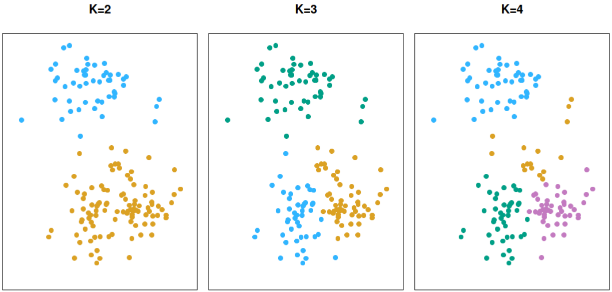


FIGURE 12.7. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.*

Details of K-Means Clustering

- Let $C_1, ..., C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:
  1. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \cdots, n\}$. In other words, each observation belongs to at least one of the K clusters.
  2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

  - Each observation belongs to one and only one cluster.
- For instance, if the ith observation is in the kth cluster, then $i \in C_k$.
- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster C_k is a measure $W(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\min_{C_1,\cdots,C_K} \sum_{k=1}^{K} W(C_k). \qquad (1)$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

How to Define Within-Cluster Variation?

- Typically we use (average squared) Euclidean distance

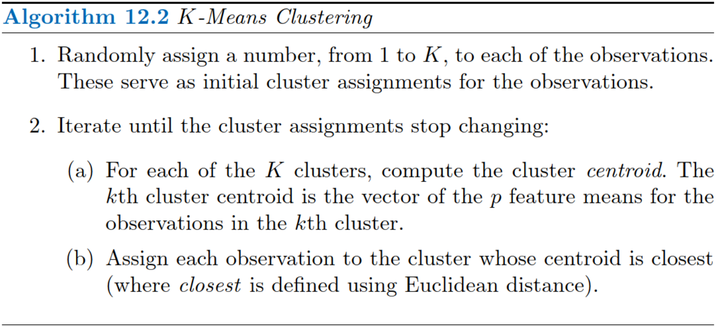
$$W\left(C_k\right)=\frac{1}{\left|C_k\right|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2=:\frac{1}{\left|C_k\right|}\sum_{i,i'\in C_k}\left\|x_i-x_{i'}\right\|^2, \tag{2}$$

  where $\left|C_k\right|$ denotes the number of observations in the kth cluster.
  - Note that there are $C_{\left|C_k\right|}^2$ pairs in the summation of (2).

- Combining (1) and (2) gives the optimization problem that defines K-means clustering,

$$\min_{C_1,\cdots,C_K}\sum_{k=1}^{K}\frac{1}{\left|C_k\right|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2=:\min_{C_1,\cdots,C_K}\sum_{k=1}^{K}\frac{1}{\left|C_k\right|}\sum_{i,i'\in C_k}\left\|x_i-x_{i'}\right\|^2. \tag{12.17}$$

- This is a very difficult problem – there are almost K^n ways to partition n observations into K clusters, i.e., an NP (nondeterministic polynomial) hard problem.
- Fortunately, the following algorithm provides a local optimum.



Algorithm 12.2 *K-Means Clustering*

```
1. Randomly assign a number, from 1 to $K$, to each of the observations.
   These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the $K$ clusters, compute the cluster *centroid*. The
       $k$th cluster centroid is the vector of the $p$ feature means for the
       observations in the $k$th cluster.

   (b) Assign each observation to the cluster whose centroid is closest
       (where *closest* is defined using Euclidean distance).
```

Properties of the Algorithm

- This algorithm is guaranteed to decrease the value of the objective (12.17) at each step.
  - Why? Note that [Exercise]

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2=2\sum_{j=1}^{p}\sum_{i\in C_k}\left(x_{ij}-\bar{x}_{kj}\right)^2, \qquad (3)$$

  where $\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$ is the mean for feature j in cluster C_k.
  - The refitting step, Step 2(a), finds $\bar{x}_{kj}$ that solves $\min_a\sum_{i\in C_k}\left(x_{ij}-a\right)^2$ for each feature j (which is why the name "K-means").
  - The assignment step, Step 2(b), reallocates the observations to improve (3). [see Figure 12.8]

- However, it is not guaranteed to give the global minimum.
  - Why not? The ultimate clustering results depend on the initial (random) cluster assignment. [see Figure 12.9]
  - It is suggested to run the algorithm multiple times from different random initial configurations, and then select the best solution.

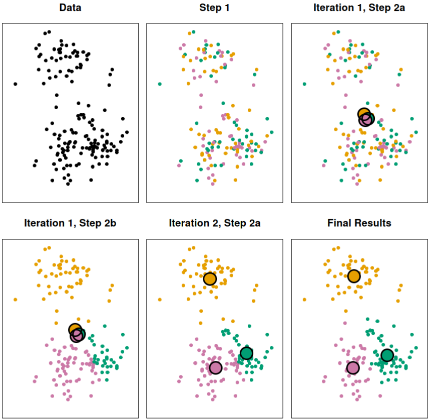


FIGURE 12.8. *The progress of the K-means algorithm on the example of Figure 12.7 with $K=3$.* Top left: *the observations are shown.* Top center: *in Step 1 of the algorithm, each observation is randomly assigned to a cluster.* Top right: *in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.* Bottom left: *in Step 2(b), each observation is assigned to the nearest centroid.* Bottom center: *Step 2(a) is once again performed, leading to new cluster centroids.* Bottom right: *the results obtained after ten iterations.*

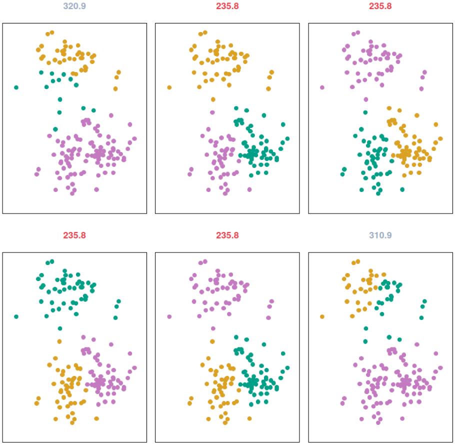


FIGURE 12.9. *K-means clustering performed six times on the data from Figure 12.7 with $K = 3$, each time with a different random assignment of the observations in Step 1 of the K-means algorithm. Above each plot is the value of the objective (12.17). Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters. Those labeled in red all achieved the same best solution, with an objective value of 235.8.*

(**) Technicalities

- Why do we use (2) or equivalently, (3), to measure the within-cluster variation?
- Essentially, we are trying to find K clusters to maximize the likelihood function:

$$\prod_{k=1}^{K}\prod_{i\in C_k} f_k\left(x_i\right),$$

  where $f_k\left(\cdot\right)$ is the density function of the kth cluster.
- If f_k is the normal density with a mean μ_k and a (common in k) variance matrix $\sigma^2\mathbf{I}$, then maximizing the likelihood function is equivalent to

$$\min_{\left(\mu_1,\cdots,\mu_K\right)}\sum_{k=1}^{K}\sum_{i\in C_k}\left\|x_i-\mu_k\right\|^2.$$

- Another way of formulating the likelihood function is to augment the data $\left\{X_i\right\}_{i=1}^n$ by the "cluster label" random vector $\left\{L_i\right\}_{i=1}^n$, where $\left\{\left(L_i,X_i\right)\right\}_{i=1}^n$ are assumed to be i.i.d. with $P\left(L_i=k\right)=p_k$, and estimate the unknown parameters $\left\{\left(\mu_k,p_k\right)\right\}_{k=1}^K$ by the EM algorithm.
- The EM algorithm will generate an estimate of $p_{ik}:=P\left(L_i=k|X_i\right)$, $k=1,2,\cdots,K$, and we can then assign the ith observation to $\arg\max_k\hat{p}_{ik}$.
- Note that $X_i\sim\sum_{k=1}^K p_kN\left(\mu_k,\sigma^2\mathbf{I}\right)$, so this is a Gaussian mixture model (GMM), also referred to as a soft clustering method, while K-means is hard.

Hierarchical Clustering

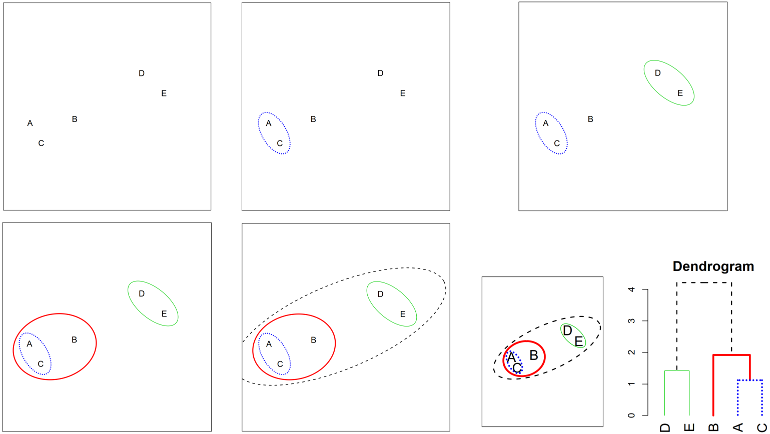
- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage (later we discuss strategies for choosing K).
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- We here describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

  - The top-down or divisive clustering reverts the bottom-up clustering, starting from the entire data set as a single cluster, and recursively dividing one of the existing clusters into two child clusters at each iteration until each cluster contains only one data point.

  - The ultimate clustering results of these two methods are usually different.

  - The top-down clustering is less popular than the bottom-up clustering in the literature, but useful when we want to partition the data into a relatively small number of clusters.

Hierarchical Clustering: the Idea



Hierarchical Clustering Algorithm

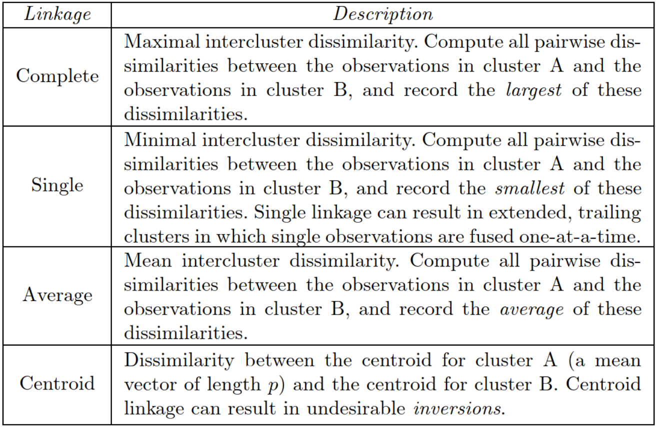
- The approach in words:
  1. Start with each point in its own cluster.
  2. Identify the closest two clusters and merge (or fuse) them.
  3. Repeat.
  4. Ends when all points are in a single cluster.
- But how to define whether two clusters (instead of two observations) are "closest" in Step 2?

  - In Table 12.3, complete and average linkages are more popular than single and centroid linkages which suffer from some drawbacks.

  - The dendrogram typically depends strongly on the type of linkage used. [see Figure 12.14]
- The height in the dendrogram indicates the dissimilarity between the two clusters in Step 2.
- For any two observations, the height where branches containing them are first fused is how dissimilar they are.

  - The similarity of two observations are not based on their proximity along the horizontal axis but along the vertical axis. [see Figure 12.12]

Linkages



Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

TABLE 12.3. *A summary of the four most commonly-used types of linkage in hierarchical clustering.*

- Inversion in centroid linkage means that two clusters are fused at a height below either of the individual clusters.

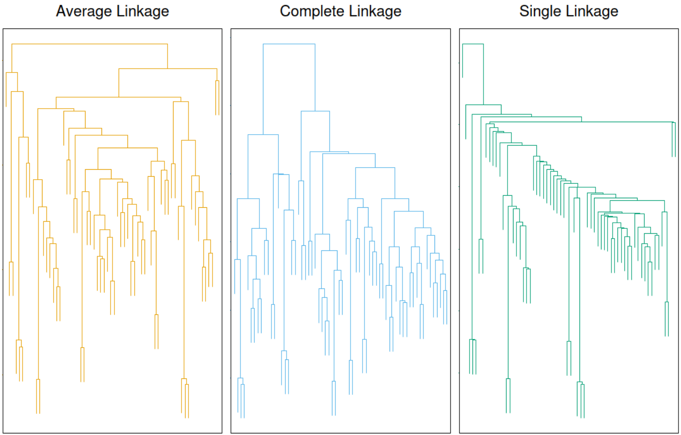


FIGURE 12.14. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

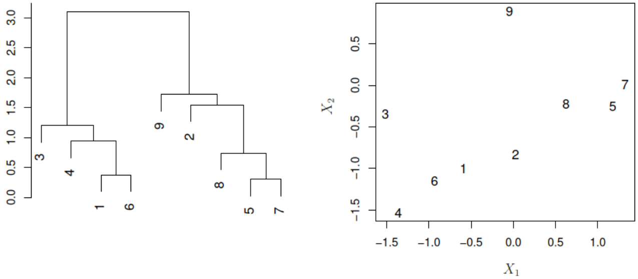


FIGURE 12.12. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space. Left: a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6*. However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7*, even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7*.*

An Example

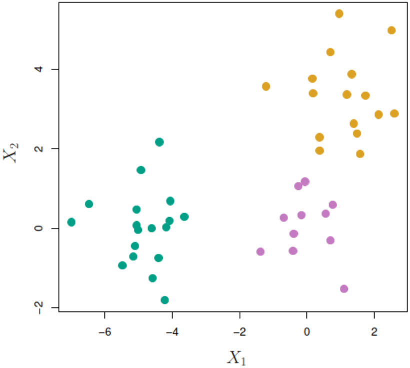


FIGURE 12.10. *Forty-five observations generated in two-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.*

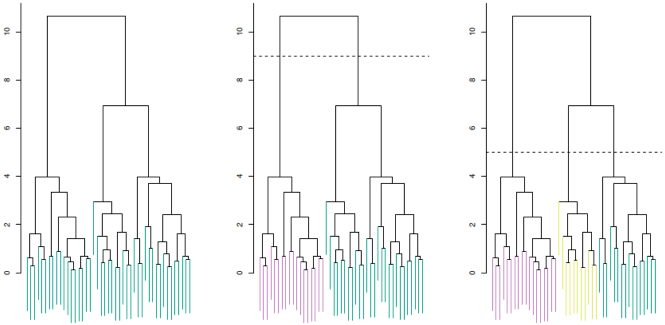


FIGURE 12.11. Left: *dendrogram obtained from hierarchically clustering the data from Figure 12.10 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.* Right: *the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.*

More on Figure 12.11

- From the middle and right panels of Figure 12.11, the height of cut serves the same role as the K in K-means clustering: it controls the number of clusters.
- One single dendrogram can be used to obtain any number of clusters; in practice, people often use eyeballs to select the number of clusters, based on the heights of the fusion and the number of clusters desired.
  - In this example, $K = 2$ or 3 clusters seem to make sense.
- Why the name "hierarchical"? Clusters obtained by cutting the dendrogram at a given height are necessarily nested within the clusters obtained by cutting the dendrogram at any greater height.
  - May not make sense, e.g., if $K = 2$, then the clustering $\{$men, women$\}$ is the best, but if $K = 3$, then the clustering $\{$American, Chinese, German$\}$ are the best.
  - Due to such situations, hierarchical clustering can sometimes yield worse results than K-means clustering for a given K.

Choice of Dissimilarity Measure

- To define linkage, we need first specify a dissimilarity measure.
- So far we have used Euclidean distance.
- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.[2]
  - This dissimilarity measure is equivalent to the Euclidean distance when the data are standardized to have mean 0 and standard deviation (SD) 1. [Exercise]
- This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation profiles for each pair of observations.
  - Correlation-based distance focuses on the shapes of observation profiles rather than their magnitudes. [see Figure 12.15]
- Two key factors determining the choice of dissimilarity measure: the type of data being clustered and the scientific question at hand. [see Figure 12.16, left panel]
  - It is often also preferable to scale all variables to have SD one when either some items are purchased more frequently than others or they are measured on different scales. [see Figure 12.16, middle panel]
  - Of course, if the spending is of main concern, then scaling is not necessary. [see Figure 12.16, right panel]

[2] This is NOT a distance; no triangle inequality.

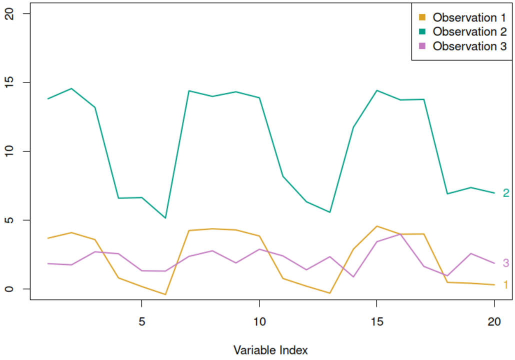


FIGURE 12.15. *Three observations with measurements on 20 variables are shown. Observations 1 and 3 have similar values for each variable and so there is a small Euclidean distance between them. But they are very weakly correlated, so they have a large correlation-based distance. On the other hand, observations 1 and 2 have quite different values for each variable, and so there is a large Euclidean distance between them. But they are highly correlated, so there is a small correlation-based distance between them.*

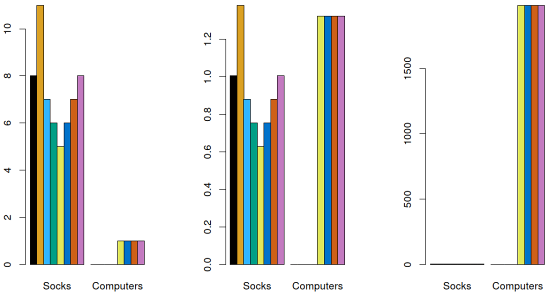


FIGURE 12.16. *An eclectic online retailer sells two items: socks and computers.*

Figure: Left: eight consumers and two items; Middle: standardization of the two items; Right: dollars instead of numbers.

- The target of the retailer is to identify subgroups of similar shoppers.
- x_i includes the number of times the shopper i purchased each item.
- If Euclidean distance is used, then overall frequency is the key clustering factor, while if correlation-based distance is used, then preference is the key factor.

Practical Issues in Clustering

- Scaling of the variables matters! Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have SD one.
- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
- How many clusters to choose? (in both K-means and hierarchical clustering). Difficult problem. No agreed-upon method.
- Which features should we use to drive the clustering?

Continued

- Are the clusters representing true subgroups in the data, or simply a result of clustering the noise? Use a test to assess that?
- A few outliers may be different from each other and from all other observations, and they may heavily distort the clustering if they are forced to be clustered to one of the K subgroups.
  - Use a soft version of K-means clustering, making probabilistic (rather than deterministic) assignments of points to cluster centers.
  - Clustering methods are not robust to perturbations to the data, e.g., removing a few observations at random may have a huge impact on the clustering.
- Suggestion: a tempered approach – try different parameters and cluster subsets of the data to check the sensitivity of clustering, and use these results as the starting point of more rigorous analyses.

Lab: Clustering

(Section 12.5.3)

- Clustering
  - *K*-Means Clustering
  - Hierarchical Clustering