Unsupervised learning – introduction

October 7, 2019

General statement of the problem

- One has a set of N observations $(x_1, x_2, ..., x_N)$ of a random p-vector X having joint density $f(X)$.
- The goal is to directly infer some properties of this probability density **without help of a 'supervisor' or a 'teacher'** who would provide correct answers or assessment of the degree-of-error for each observation.
- The dimension of X can be much higher than in supervised learning, and the properties of interest are often complicated and not easily formalized: some structural relations between variables, the patterns of behaviors, etc.
- Often the 'discovered' properties constitutes a starting point for further investigation, possibly, through supervised methods.

Example – genes and microarray data

- Suppose that the observations $(x_1, x_2, ..., x_N)$ represents gene activities of a certain group of population in which certain various pathological features was observed, say, cancer.
- The data on the pathologies are not given but a distant goal is to find some relation between them and the genes activities.
- The goal is to identify some gene patterns and group individuals with respect to these patterns – this would be a non-supervised learning problem.
- Then by succeeding in the above and thus having the population classified by these patterns, one can further search if these patterns are responsible for some pathologies.
- For example, if the certain groups are more inclined to get certain cancer, this could be achieved by designing a supervised learning problem, classification problem.

Learning without a teacher

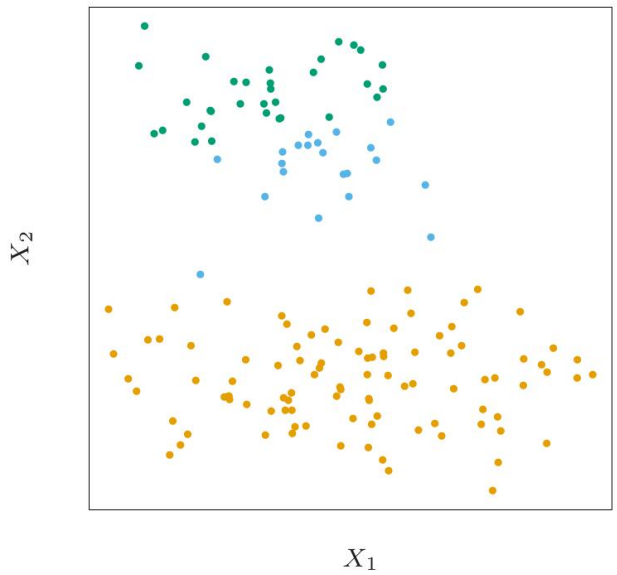
- With supervised learning, due to availability of values of Y in training and testing, there is a clear measure of success, or lack thereof, that can be used to judge adequacy in particular situations and to compare the effectiveness of different methods over various situations. Methods can be validated, for example, through cross-validation.
- In the context of unsupervised learning, there is no such direct measure of success.
  It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms.
- Heuristic arguments for judgments as to the quality of the results.
- Effectiveness often is a matter of opinion and cannot be verified directly.

Basic idea of clustering

- The idea behind cluster analysis (data segmentation) is simple:

Identify **groupings** or **clusters** of individuals that are not readily apparent to the researcher.

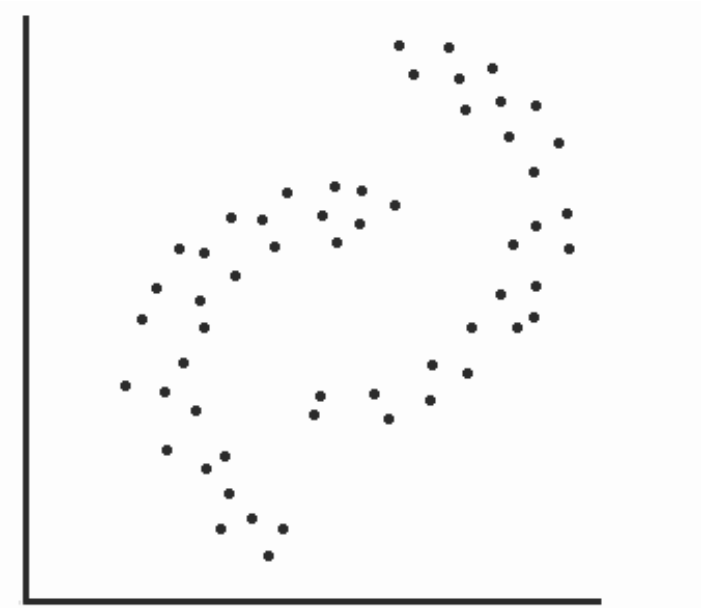
- Important aspect of it is using multiple variables, which are more difficult to analyze by visual inspection – similarities can be “hidden” in high dimensions.
- The figure below gives a simplistic example of three clusters (two clusters and one data segmentation) defined by two variables.



- Central to cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered.
- A clustering method attempts to group the objects based on the definition of similarity.

What kind of clusters?

- The problem with cluster analysis is that in all but the simplest of cases uniquely defined clusters may not exist.
- Cluster analysis may classify the same observations into completely different groupings depending on the choice of a method.
- Cluster analysis tends to be good at finding spherical clusters and has great difficulty with curved clusters.



Similarity and distance

Clustering means grouping observations into subgroups in such a way that observations within subgroups are “similar”.
For example

- group languages into families using characteristics of the languages
- divide animals and plants into different species and families using a variety of characteristics

Clustering algorithms typically consist of the followings steps:

1. Determine “distances” or similarities between all pairs of objects. These distances or similarities define a symmetric matrix:

   dissimilarity matrix.

2. Run an algorithm that takes this matrix as the input.

Measuring similarities

- Two objects (i and j) having multivariate values $\mathbf{x}_i$ and $\mathbf{x}_j$ are assigned a measure of dissimilarity d_{ij} with the following properties:

$$d_{ij} \geq 0$$

$$d_{ii} = 0$$

$$d_{ij} = d_{ji}$$

- Some measures of dissimilarity are also distances (satisfying the triangular inequality).

Metric variables

Common distance measures:

- ‘Cityblock’ $$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

- Euclidian distance $$d_{ij} = \sqrt{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^2}$$

- or more generally

Minkowski distance $$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r}$$

Other measures

- Clustering can be based on the variable 'correlation' between two observations

$$\rho_{ik} = \frac{\sum_{j=1}^{p}(x_{ij} - \bar{x}_{i\cdot})(x_{ik} - \bar{x}_{k\cdot})}{\sqrt{\sum_{j=1}^{p}(x_{ij} - \bar{x}_{i\cdot})^2 \sum_{j=1}^{p}(x_{kj} - \bar{x}_{k\cdot})^2}}$$

Note that the correlation is averaged over variables in an observation x not over observations – high correlation (close to one) means that variables between two observation depend nearly linearly one on another.

- Ordinal variables: code them to $(i - 1/2)/M$, $i = 1, \ldots, M$, where M is the number of ordinal variables.
- Categorical variables: Take zero-one distance, i.e.
  - if a variable has the same value for two observations the distance is 'zero', otherwise is 'one'
  - count number of 'ones' as the distance: a lot of non-zeros the observations are distant
  - other integers can be used to emphasize different kinds of dissimilarities

Hierarchical cluster methods

- This kind of clustering starts with the calculation of the ‘distances’ of each individual to all other individuals in the dataset.
- Groups are formed by the process of agglomeration or division.
- **Agglomeration**
  - Start with the most refined grouping, i.e. each individual constitute a separate group – singeltons.
  - Then through certain agglomeration algorithms we arrive to a smaller number of larger groups made of many ‘similar’ members.
  - Eventually we end up with the single most crude group of all individuals.
- **Division**
  - Not so popular as agglomeration, it starts with one the most crude grouping made of all individuals
  - By process division of larger groups into smaller ones we arrive through certain algorithms to larger number of smaller groups made of only the most similar members.
  - Eventually we end up with singletons

Agglomeration algorithm – general scheme

We want to cluster n objects.

1. Initiate the process with n clusters; one for each individual or object.
2. Two groups A and B that based on their distance or dissimilarity d_{AB} are closest to each other among all cluster pairs at a given stage of the algorithm are merged with one another.
3. Calculate dissimilarities between the new group and all other clusters.
4. Repeat Steps 2 and 3 until finally all individuals are in one single group.

The sequence of grouping operations can be illustrated as a tree diagram aka **dendrogram** that is then used to identify clusters.

Division procedure

This is ‘agglomeration in reverse’:

1. All *n* objects start in a single group (number of groups=1).
2. This is then split into two groups using one of a number of rules for choosing the best split of one group into two groups.
3. Each of the two groups are in turn split, and so on until all individuals are in groups of their own.

The sequence of grouping operations can be inspect visually or by some numerical analysis of the tree diagram **dendrogram** – identification of the groups is made in the same manner as in agglomeration technique.

Why is it harder to divide, than to agglomerate?

Defining distances between clusters

- Suppose at a certain step of algorithm the two groups A and B were agglomerate to one group (AB).
- For any other cluster C the distances between A and C: d_{AC} and B and C: d_{BC} were given
- To define the algorithm one has to define how the distance from (AB) to any other cluster C: $d_{(AB)C}$ will be measured, i.e. the relation between $d_{(AB)C}$ and the pair (d_{AC}, d_{BC}) has to be given.
- Occasionally, d_{AB} is also used to define $d_{(AB)C}$.

Dissimilarities between clusters

- **Single linkage: Nearest neighbor** clustering computes the similarity between two groups as the similarity of the closest pair of observations between the two groups.

$$d_{(AB)C} = \min(d_{AC}, d_{BC})$$

- **Complete linkage: Farthest neighbor** clustering uses the farthest pair of observations between two groups to determine the similarity of the two groups.

$$d_{(AB)C} = \max(d_{AC}, d_{BC})$$

Another linkage method

Average linkage clustering uses the average similarity of observations between two groups as the measure between the two groups.

$$d_{(AB)C} = \frac{|A|d_{AC} + |B|d_{BC}}{|A| + |B|}$$

where $|A|$ denotes the number of elements in A.

McQuittys method

$$d_{(AB)C} = (d_{AC} + d_{BC})/2$$

Some other averaging methods

- **Gowers method**

$$d_{(AB)C} = 0.5d_{AC} + 0.5d_{BC} - 0.25d_{AB}$$

The term $-0.25d_{AB}$ 'encourages' merging distant clusters through clusters that 'lie' between them.

- **Centroid method**

$$d_{(AB)C} = \frac{|A|d_{AC} + |B|d_{BC}}{|A| + |B|} + \frac{|A| \cdot |B|}{(|A| + |B|)^2} d_{AB}$$

The term $\frac{|A| \cdot |B|}{(|A|+|B|)^2} d_{AB}$ penalize for merging too large too distant clusters because if both A and B are large than $|A| \cdot |B|$ are on the same magnitude as $(|A| + |B|)^2$ but if one is evidently smaller than $(|A| + |B|)^2$ is much bigger than $|A| \cdot |B|$.

Method based on Sum of Squares

- **Ward's method** is distinct from all the other methods because it uses an analysis of variance approach to evaluate the distances between clusters.
- In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step.
- Let ESS_A be the sum of squares for a cluster A

$$\sum_{\mathbf{x} \in A} (\mathbf{x} - \bar{\mathbf{x}}_A)'(\mathbf{x} - \bar{\mathbf{x}}_A)$$

- Combine two clusters A and B together, then the new sum of squares satisfies

$$ESS_{(AB)} = ESS_A + ESS_B + |A|(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_{AB})^2 + |B|(\bar{\mathbf{x}}_B - \bar{\mathbf{x}}_{AB})^2$$

- Put these cluster together that minimize the increase of the sum of squares

$$ESS_{(AB)} - ESS_A - ESS_B = |A|(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_{AB})^2 + |B|(\bar{\mathbf{x}}_B - \bar{\mathbf{x}}_{AB})^2$$

Partition methods

- The partition methods do not require to go through all levels of grouping (i.e. from the singletons to the single group made of all members).
- Partition methods break the observation into distinct nonoverlapping groups.
- There are many different partition methods, we briefly discuss two of them.

k-means clustering – “MacQueens metod”

- The *k*-means clustering is particularly appropriate when the number of clusters or the approximate number of clusters is known apriori.
- Unlike hierarchical cluster analysis, the *k*-means clustering can not produce all possible clusters of *n* observations.
- The *k*-means cluster analysis programs begin by creating the *k* clusters according to some arbitrary procedure.
- The program calculates the means or centroids of each of the clusters.
- If one of the observations is closer to the centroid of another cluster then the observation is made a member of that cluster.
- This process is repeated until none of the observations is reassigned to a different cluster.

Algorithm

Algorithm 14.1 *K-means Clustering.*

1. For a given cluster assignment C, the total cluster variance (14.33) is minimized with respect to $\{m_1, \ldots, m_K\}$ yielding the means of the currently assigned clusters (14.32).

2. Given a current set of means $\{m_1, \ldots, m_K\}$, (14.33) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2. \qquad (14.34)$$

3. Steps 1 and 2 are iterated until the assignments do not change.

$$\bar{x}_S = \underset{m}{\operatorname{argmin}} \sum_{i \in S} \|x_i - m\|^2 \qquad (14.32)$$

$$\sum_{k=1}^{K} N_k \sum_{C(i)=k} \|x_i - m_k\|^2 \qquad (14.33)$$

Comments

- In the measurement of the closeness to group-center, the Euclidean distance is often used.
- The process of partioning is sensitive to the starting point and can result in different clusters for different starting grouping.
- The choice of k is not obvious.

k-medians clustering

- K-medians clustering is a variation on the k-means method.
- The same process is followed except that medians are used instead of means.
- K-medians would be appropriate when you need a more stable measure of the group centers.

The choice of clusters problem

- In both the approaches to clustering: hierarchical and partition methods, there is a problem of determining the final form and the number of the clusters.
- In the hierarchical clustering, we have available the tree that illustrates how the clusters form but we do not know at which level we should cut the tree to form the final clusters.
- In the partition method, we start with the initial number of cluster but we need to identify this number to be optimal.
- A criterion that could help to decide the optimality of the cluster selection is needed.

Dissimilarity measures

- One common approach is to use a some kind of dissimilarity measure that quantify dissimilarity for any clustering.
- In a one cluster everything is the same and there is no dissimilarity.
- If there are more than just one cluster, then the dissimilarity of them should be measured how far they are from each other.
- The measure should decrease if we go from a higher number of clusters to a smaller one.
- Any reduction of the number of clusters that results in a huge drop of the dissimilarity measure should be considered undesirable (we put together very dissimilar objects).

The Gap statistics

- One can compare the reduction of the dissimilarity reductions against a completely randomly distributed observations (no clusters)
- Then to choose the clustering for which the drop of the dissimilarity encompassed by the reduction of the number of clusters is the largest as compared to the completely random distribution of the x-variables. (If the drop large it means that we are joining groups that are distant.)
- This applies both to the hierarchical clustering and the choice of the number of the clusters in the partition method.
- The method has been proven quite effective and utilized in a formal testing problem through the Gap statistic that calibrates drop of the dissimilarity against totally random distribution of the points.
- Instead, for the hierarchical model one can utilize the graphs that result from the clustering and use them to decide for the clear cut between the clusters.
- The height of branches (from the top) often represent how dissimilar the clusters are by adding distances between clusters at a given step of the algorithm.

Choosing k in the partition method

- Perform the procedure for different k.
- One possibility is to choose k such that the between cluster variability relative the within cluster variability is maximized.
- With

$$\mathbf{B} = \sum_i (\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{i\cdot} - \bar{\mathbf{x}}_{\cdot\cdot})'$$

and

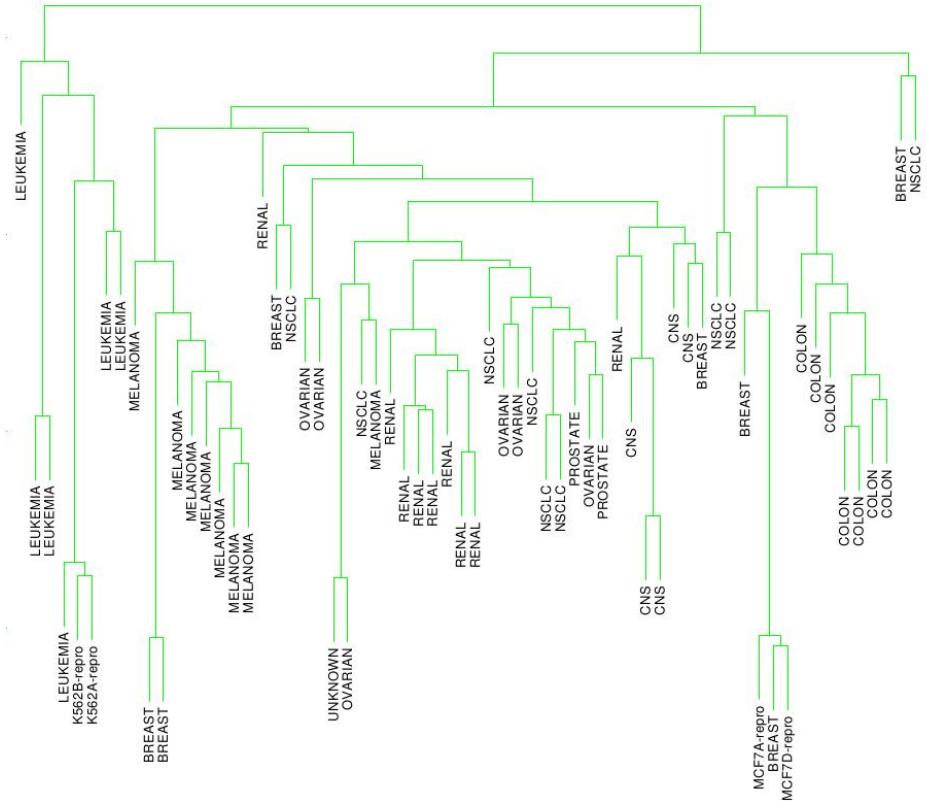
$$\mathbf{W} = \sum_i \sum_j (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i\cdot})'$$

we want to maximize $|\mathbf{B}|/|\mathbf{B} + \mathbf{W}|$ or $\text{tr}(\mathbf{B}\mathbf{W}^{-1})$ (possibly with some penalty on large number of clusters).

Human tumor data

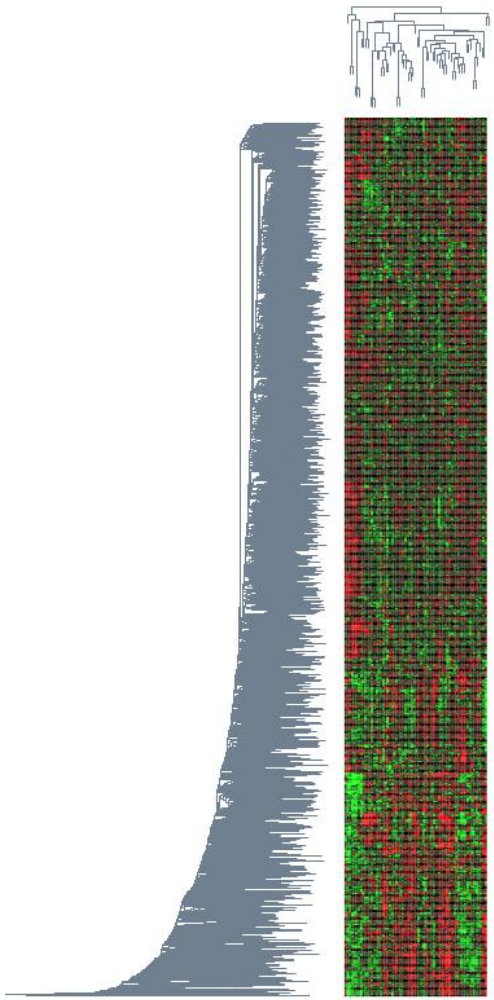
- The data are a 6830 $\times$ 64 matrix of real numbers, each representing an expression measurement for a gene (row) and sample (column).
- Here we cluster the samples, each of which is a vector of length 6830, corresponding to expression values for the 6830 genes.
- Each sample has a label such as breast (for breast cancer), melanoma, and so on; we don't use these labels in the clustering, but will examine posthoc which labels fall into which clusters.

Clustering trees – dendrogram



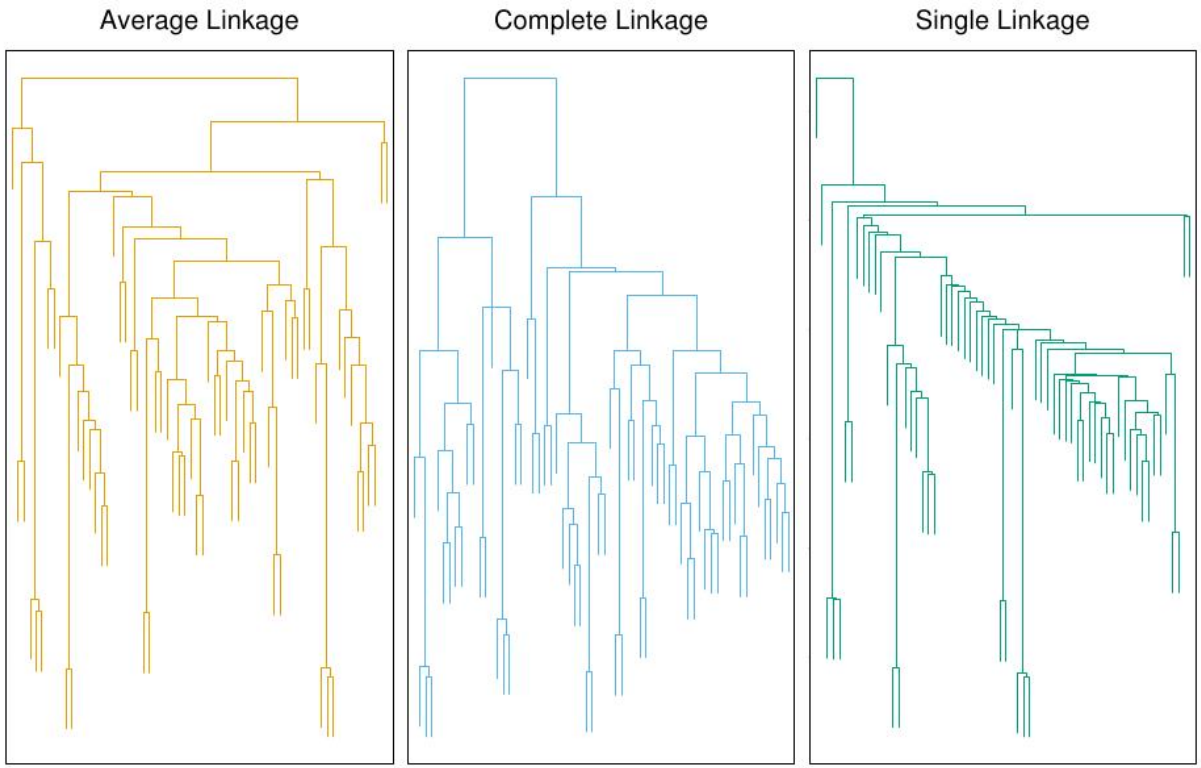
- Dendrogram resulting from average linkage agglomerative clustering of the samples (columns) of the microarray data.
- Hierarchical clustering is successful at clustering simple cancers together.
- By cutting off the dendrogram at various heights, different numbers of clusters emerge, and the sets of clusters are nested within one another.

Ordering in hierarchical clustering



- Genes (rows) and samples (columns) of the expression matrix are arranged in orderings derived from hierarchical clustering
- To produce the row ordering: at each merge, the subtree with the tighter cluster is placed to the left (toward the bottom in the rotated dendrogram in the figure.)
- Individual genes are the tightest clusters possible, and merges involving two individual genes place them in order by their observation number.
- The same rule was used for the columns.
- By grouping genes we obtain genes with more similar roles (activities) across all samples at the bottom.

Other clustering methods



- The left panel shows the dendrogram resulting from average linkage agglomerative clustering of the samples (columns) of the microarray data.
- The middle and right panels show the result using complete and single linkage. Average and complete linkage gave similar results, while single linkage produced unbalanced groups with long thin clusters.

K-means clustering

- We applied *K*-means clustering with *K* running from 1 to 10
- The total within-sum of squares for each clustering is shown in the figure.
- No kink in the sum of squares curve to locate the optimal number of clusters
- for illustration we chose K = 3 giving the three clusters shown in the table

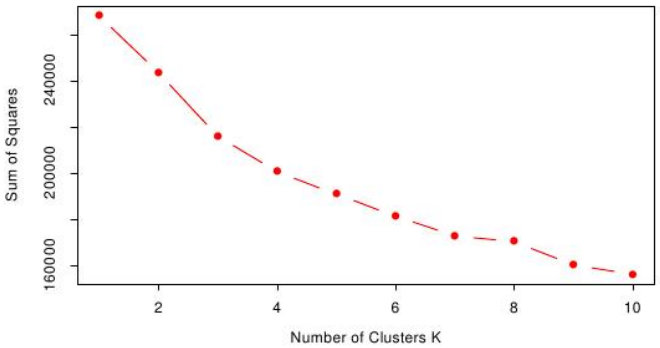


TABLE 14.2. *Human tumor data: number of cancer cases of each type, in each of the three clusters from K-means clustering.*

Cluster	Breast	CNS	Colon	K562	Leukemia	MCF7
1	3	5	0	0	0	0
2	2	0	0	2	6	2
3	2	0	7	0	0	0
Cluster	Melanoma	NSCLC	Ovarian	Prostate	Renal	Unknown
1	1	7	6	2	9	1
2	7	2	0	0	0	0
3	0	0	0	0	0	0

- Method is grouping together samples of the same cancer.
- *K*-means clustering has shortcomings
- it does not give a linear ordering of objects within a cluster;
- as the number of clusters *K* is changed, the cluster memberships can change in arbitrary ways: the clusters need not be nested within the three clusters above.
- Hierarchical clustering is preferable