STK2100- Machine Learning and Statistical Methods for Prediction and Classification

Spring 2026

Unsupervised learning

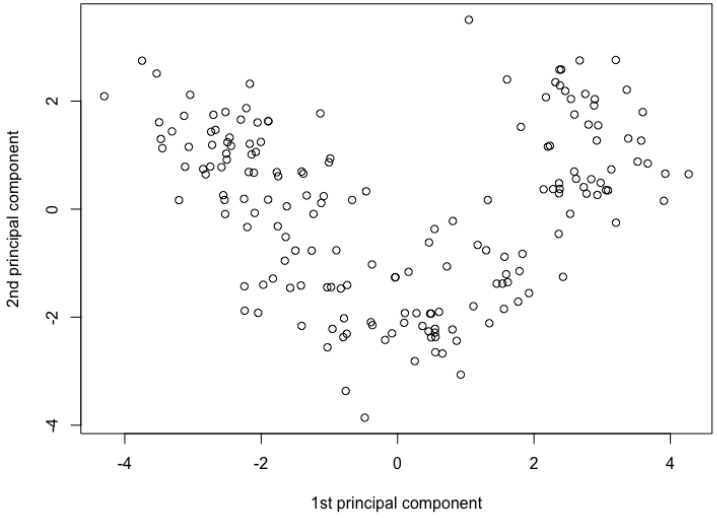
Covers Sections 14.1 and 14.3 (except 14.3.9)

Ingrid Hobæk Haff
Department of Mathematics
University of Oslo

Example: Wine data

Variables:

- Alcohol percentage (`Alcohol`)
- Malic acid (`Malic`)
- Ash (`Ash`)
- Alcalinity of ash (`Alcalinity`)
- Magnesium (`Magnesium`)
- Total phenols (`Phenols`)
- Flavanoids (`Flavanoids`)
- Nonflavanoid phenols (`Nonflavanoids`)
- Proanthocyanins (`Proanthocyanins`)
- Color intensity (`Color`)
- Hue (`Hue`)
- D280/OD315 of diluted wines (`Dilution`)
- Proline (`Proline`)



We want to identify types of wines with similar characteristics

Example: DNA microarray data

Variables:

- Sample, each corresponding to one cancer patient
- Gene expression from 6830 genes

We want to understand how the genes and samples are organised

Unsupervised learning

- So far, we have considered supervised learning, with data of the type $\{(\mathbf{x}_i, y_i), i = 1, ..., N\}$, where the response y_i is either numeric (regression), or categorical (classification)
- Now, we will focus on situations where there is no response y_i, only variables $\{\mathbf{x}_i, i = 1, ..., N\}$
- The aim is to cluster the data into a set of groups/categories
- Examples:
  - Segmentation of a company's customer base
  - Clustering genes
  - Recognising communities in social networks
  - Clustering movies for recommender systems

Unsupervised learning

- This is unsupervised learning, whichs refers to "learning without a teacher", and consists in inferring properties of $p(\mathbf{x})$
- One of the challenges, is that we cannot define a loss function in the same way
- This also makes it difficult to assess the validity and performance of unsupervised learning methods

Cluster analysis

- There are three types of clustering methods
  - combinatorial approches
  - model based approaches
  - mode seeking approaches
- Model based approaches define a mixture model, with latent cluster membership variables C_i, $i = 1, \ldots, N$, and define a model for $(\mathbf{X}_i, C_i)$, typically

$$\begin{aligned} \Pr(C_i = k) &= \Pr(\mathbf{X}_i \text{ is in cluster } k) = \alpha_k, \quad k = 1, \ldots, K \\ \mathbf{X}_i | C_i = k &\sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k = 1, \ldots, K \end{aligned}$$

- The parameters $(\alpha_1, \ldots, \alpha_K, \mu_1, \ldots, \mu_K, \Sigma)$ are then estimated based on the data $\{\mathbf{x}_i, i = 1, ..., N\}$
- Mode seeking consists in modelling $p(\mathbf{x})$ non-parametrically, and searching for the modes of the distribution
- Here, we will focus on algorithmic based approaches

Combinatorial approaches to cluster analysis

- Clusters are built by merging observations that are close
- This requires the definition of a distance/dissimilarity measure $d(\mathbf{x}_i, \mathbf{x}_{i'})$ of the distance/dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$
- These must fulfil

$$
\begin{aligned}
d(\mathbf{x}_i, \mathbf{x}_{i'}) &\geq 0 \quad \text{Positivity} \\
d(\mathbf{x}_i, \mathbf{x}_i) &= 0 \\
d(\mathbf{x}_i, \mathbf{x}_{i'}) &= d(\mathbf{x}_{i'}, \mathbf{x}_i) \quad \text{Symmetry} \\
d(\mathbf{x}_i, \mathbf{x}_{i'}) + d(\mathbf{x}_{i'}, \mathbf{x}_{i''}) &\geq d(\mathbf{x}_i, \mathbf{x}_{i''}) \quad \text{Triangle inequality}
\end{aligned}
$$

Distance measures

- After some normalisation, it is common to let

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} d_j(x_{i,j}, x_{i',j}) \quad \text{Sum over all components}$$

- The typical choice for quantitative variables is

$$d_j(x_{i,j}, x_{i',j}) = (x_{i,j} - x_{i',j})^2 \quad \text{Squared Euclidean}$$

- The typical choice for qualitative variables is

$$d_j(x_{i,j}, x_{i',j}) = 1 - I(x_{i,j} = x_{i',j})$$

Alternative distance measures

Name	$d(i, i')$
Euclidean distance simple: $w_j = 1$ weighted with variance $w_j = 1/s_j^2$ weighted with range: $w_j = 1/R_j^2$	$\left(\sum_{j=1}^{p} w_j \, (x_{ij} - x_{i'j})^2\right)^{1/2}$
Mahalanobis distance (where Σ is a positively defined matrix)	$\left\{(\tilde{x}_i - \tilde{x}_{i'})^\top \Sigma^{-1} (\tilde{x}_i - \tilde{x}_{i'})\right\}^{1/2}$
Minkowsky distance (for a parameter $\lambda \geq 1$)	$\left(\sum_{j=1}^{p} w_j \, (x_{ij} - y_{i'j})^\lambda\right)^{1/\lambda}$
Manhattan distance	$\sum_{j=1}^{p} w_j \, \lvert x_{ij} - x_{i'j} \rvert$
Canberra metric (one of several variants), where terms in which denominator is 0 are excluded	$\sum_{j=1}^{p} \dfrac{\lvert x_{ij} - x_{i'j} \rvert}{\lvert x_{ij} \rvert + \lvert x_{i'j} \rvert}$
L_∞ norm	$\max_j \lvert x_{ij} - x_{i'j} \rvert$

Non-hierarchical clustering methods

- Combinatorial clustering approaches can again be divided into hierarchical and non-hierarchical methods
- In non-hierarchical, which we will consider first, one first has to specify the number K of clusters
- Then, one must make a first allocation of each observation to one of the clusters, resulting in $C(i) \in \{1, \ldots, K\}$, $i = 1 \ldots, N$
- Further, let $d_{ii'} \equiv d(x_i, x_{i'})$, and define the loss

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{\{i:C(i)=k\}} \sum_{\{i':C(i')=k\}} d_{ii'}$$

- The cluster assignments are then adjusted in order to minimise the loss $W(C)$

Non-hierarchical clustering methods

- If T is the total point scatter, we have

$$
\begin{aligned}
T &= \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} d_{ii'} \\
&= \frac{1}{2}\sum_{k=1}^{K}\sum_{i:C(i)=k}\left(\sum_{\{i':C(i')=k\}} d_{ii'} + \sum_{\{i':C(i')\neq k\}} d_{ii'}\right) \\
&= W(C) + B(C)
\end{aligned}
$$

- $W(C)$ is called the within cluster point scatter
- $B(C)$ is called the between cluster point scatter
- Note that since T does not depend on the clusters, minimising $W(C)$ is equivalent to maximising $B(C)$
- Going through all possible divisions of the points into the K clusters is a huge combinatorial optimisation problem
- In practice, one must use a greedy algorithm, such as K-means or K-medoids

K-means clustering

- In K-means clustering, the squared Euclidean distance measure $d_{ii'} = ||\mathbf{x}_i - \mathbf{x}_{i'}||^2$ is used
- The resulting loss function is then given by

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{\{i:C(i)=k\}} \sum_{\{i':C(i')=k\}} ||\mathbf{x}_i - \mathbf{x}_{i'}||^2 = \sum_{k=1}^{K} N_k \sum_{\{i:C(i)=k\}} ||\mathbf{x}_i - \bar{\mathbf{x}}_k||^2,$$

  with $N_k = \sum_{i=1}^{N} I(C_i = k)$ and $\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{\{i:C_i=k\}} \mathbf{x}_i$
- The objective is then to find the clustering C^*, which is the solution to

$$C^* = \min_C \sum_{k=1}^{K} N_k \sum_{\{i:C(i)=k\}} ||\mathbf{x}_i - \bar{\mathbf{x}}_k||^2$$

K-means clustering

- Since for any subset S of the observations

$$\bar{\mathbf{x}}_S = \arg\min_{\mathbf{m}} \sum_{i \in S} ||\mathbf{x}_i - \mathbf{m}||^2,$$

C^* can be obtained by solving

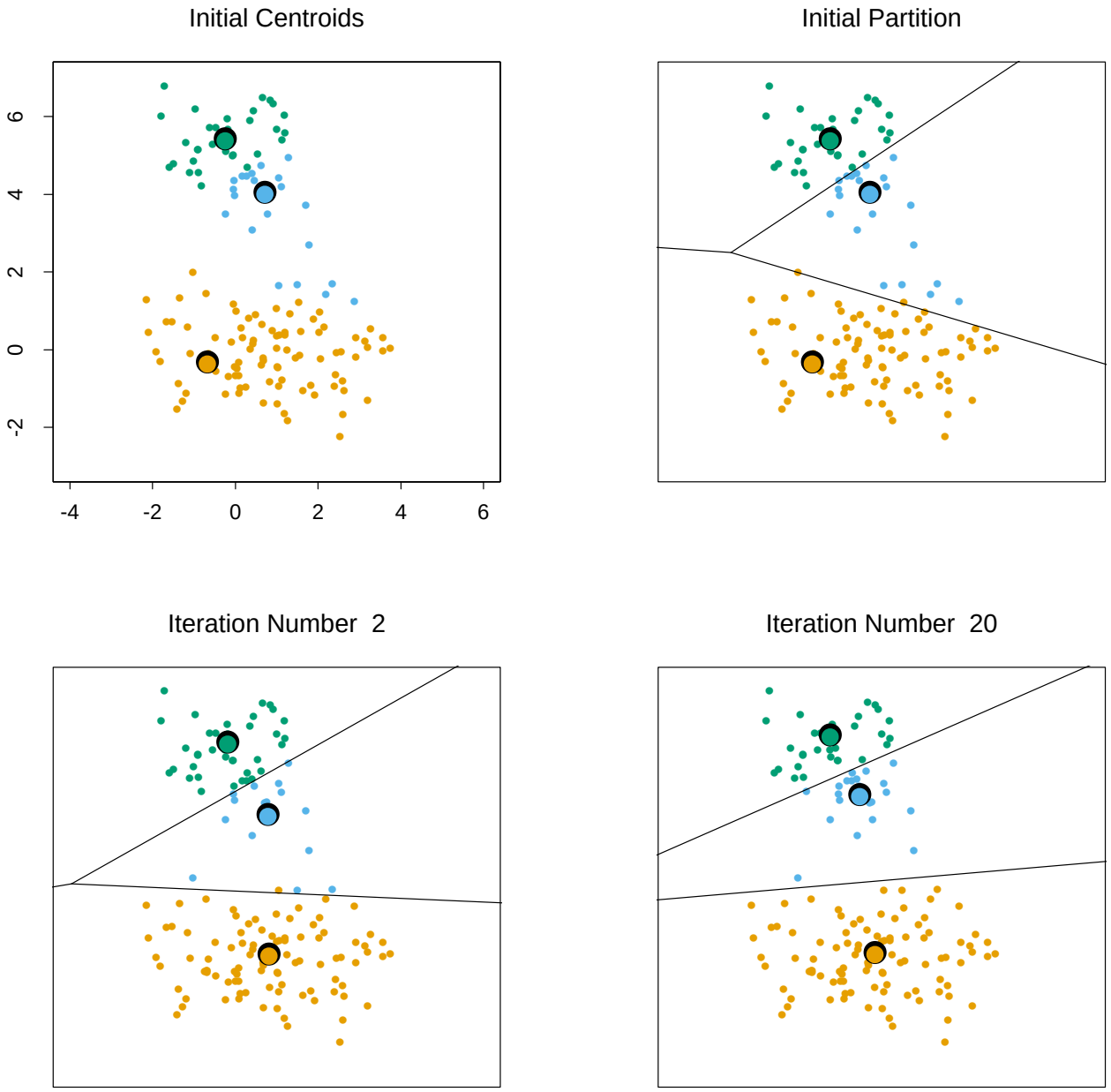
$$C^* = \min_{C, \mathbf{m}_1, \ldots, \mathbf{m}_K} \sum_{k=1}^{K} N_k \sum_{\{i : C(i) = k\}} ||\mathbf{x}_i - \mathbf{m}_k||^2 \qquad (1)$$

- This is done by iterating the steps
  1. identifying the centroids $\mathbf{m}_1, \ldots, \mathbf{m}_K$ that minimise (1)
  2. allocating observations to closest centroid to minimise (1) w.r.t C, i.e.

$$C(i) = \arg\min_k ||\mathbf{x}_i - \mathbf{m}_k||^2$$

- The two steps are iterated until the cluster assignments no longer change

Illustration of K-means clustering



Model-based clustering with mixture of Gaussians

- Let us consider model based clustering based on a mixture of K Gaussians, i.e. assuming the model

$$\begin{aligned} \Pr(C_i = k) &= \Pr(\mathbf{X}_i \text{ is in cluster } k) = \alpha_k, \quad k = 1, \ldots, K \\ \mathbf{X}_i | C_i = k &\sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k = 1, \ldots, K \end{aligned}$$

- Letting $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma})$, the likelihood function of this model is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= f(\mathbf{x}_1, ..., \mathbf{x}_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \Pr(C_i = k) p(\mathbf{x}_i | C_i = k) \\ &= \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \alpha_k \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)} \right) \end{aligned}$$

- The parameters $\boldsymbol{\theta}$ are estimated by maximising the above likelihood, using an algorithm, that iterates between making cluster assignments $\widehat{C}_i$ for fixed parameters and maximising the likelihood for fixed cluster assignments

Model-based clustering with mixture of Gaussians

- Now, let $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ in the mixture model
- Then

$$\begin{aligned} p(\mathbf{x}_i | C_i = k) &= \frac{1}{(2\pi)^{p/2}\sigma^p} e^{-\frac{1}{2\sigma^2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T(\mathbf{x}_i - \boldsymbol{\mu}_k)} \\ &= (2\pi)^{-p/2}\sigma^{-p} e^{-\frac{1}{2\sigma^2}||\mathbf{x}_i - \boldsymbol{\mu}_k||^2} \end{aligned}$$

- For given cluster assignments $\widehat{C}_i$, the parameters $\boldsymbol{\theta}$ are estimated by

$$\hat{\alpha}_k = \frac{N_k}{N}, \quad \hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{\{i:\widehat{C}_i = k\}} \mathbf{x}_i, \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{k=1}^{K} N_k \sum_{\{i:\widehat{C}_i = k\}} ||\mathbf{x}_i - \bar{\mathbf{x}}_k||^2$$

Model-based clustering with mixture of Gaussians

- For given parameter values $\boldsymbol{\theta}$, the cluster assignments $\widehat{C}_i$s are given by

$$
\begin{aligned}
\widehat{C}_i &= \arg\max_k \left(\alpha_k p(\mathbf{x}_i | C_i = k)\right) = \arg\max_k \left(\log\left(\alpha_k p(\mathbf{x}_i | C_i = k)\right)\right) \\
&= \arg\max_k \left(\log \alpha_k - \frac{p}{2}\log(2\pi) - p\log\sigma - \frac{1}{2\sigma^2}||\mathbf{x}_i - \boldsymbol{\mu}_k||^2\right) \\
&= \arg\max_k \left(\log \alpha_k - \frac{1}{2\sigma^2}||\mathbf{x}_i - \boldsymbol{\mu}_k||^2\right)
\end{aligned}
$$

- These two steps are iterated until the results stabilise
- This can be seen as a soft K-means clustering, where the cluster assignments are probabilistic, rather than deterministic
- When $\sigma^2 \to 0$, the probabilities of assigning a data point to each of the clusters approach 0 and 1, and in the limit the two methods coincide

K-medoids clustering

- The K-means algorithm is only appropriate when the dissimilarity measure is the squared Euclidean distance
- However, this requires that all the variables are quantitative
- Further, it is not robust to outliers
- An alternative is to replace the squared Euclidean distance with a more general dissimilarity measure
- The centroid $\mathbf{m}_k$ of cluster k is then chosen as the observation x_{i^*} that minimises the total distance to all the other points in the cluster, i.e.

$$i^* = \underset{\{i:C(i)=k\}}{\arg\min} \sum_{\{i':C(i')=k\}} d(\mathbf{x}_i, \mathbf{x}_{i'})$$

resulting the the K-medoids algorithm

Practical issues

- The methods considered so far require the number K of clusters to be specified, as well as initialisation of the cluster assignments C_i, $i = 1, \ldots, N$, or the cluster centres $\mathbf{m}_1, ..., \mathbf{m}_K$
- The centres can for instance be initialised with random draws, or based on forward stepwise assignment
- Specifying K is more difficult, and should depend on the aim of the clustering, as different K can give completely different clusters
- One option is to use the gap curve

Practical issues

- First, one performs K-means or K-medoids clustering for $K = 1, \ldots, K_{max}$, where K_{max} is the maximum number of clusters that one wants, and one lets $W_1, \ldots, W_{K_{max}}$ be the corresponding within cluster dissimilarities
- Then, one plots the $\log W_K$s as a function of K
- Subsequently, one computes and plots the corresponding curve for data uniformly distributed over the range of the data.
- Finally, one chooses the value of K for which the gap between the two curves is the largest

Example: Wine data

See R code

Hierarchical clustering methods

- A weakness of the combinatorial clustering methods considered so far, is that the clusters are not nested as K decreases
- A type of clustering methods that result in nested clusters is hierarchical clustering methods
- These do not require the specification of K, nor initialisation of the clusters, simply a dissimilarity measure $d(\mathbf{x}_i, \mathbf{x}_{i'})$
- The hierarchical methods may again be grouped into agglomerative and divisive methods

Agglomerative methods - formalization

- One needs a measure of dissimilarity between groups G and G', fo instance

$$
\begin{aligned}
d_S(G, G') &= \min_{i \in G, i' \in G'} d(i, i') \quad \text{Single link} \\
d_C(G, G') &= \max_{i \in G, i' \in G'} d(i, i') \quad \text{Complete link} \\
d_S(G, G') &= \frac{1}{N_G N_{G'}} \sum_{i \in G, i' \in G'} d(i, i') \quad \text{Average link}
\end{aligned}
$$

- One starts with one cluster for each observation, i.e. $K = N$, and the group dissimilarities are then simply the dissmilarities $d(i, i') = d(\mathbf{x}_i, \mathbf{x}_{i'})$ between two observations
- Then, one iteratively merges the two groups that are most similar to each other, i.e. the groups G and G' with the smallest $d(G, G')$, until $K = 1$
- The result is a hierarchical/tree structure, in which the clusters are nested, so that the division into $K + 1$ clusters "contains" the division into K clusters

Divisive clustering methods

- Divisive methods start with $K = 1$, i.e. a single cluster G, containing all the observations, and iteratively splits clusters in two until $K = N$
- There are several such algorithms, and one of them is as follows
- A new cluster H is started with the observation in G that has the largest average dissimilarity from all the other observations
- Then, one continues adding the observation in G, whose average distance from those in H, minus that for the remaining observations in G is largest, to H
- This continues until the corresponding difference in averages becomes negative, and the division into $K = 2$ clusters is done

Divisive clustering methods

- For each $K = 3, \ldots, N$, this splitting procedure is then applied to the cluster with the largest diameter

$$D_G = \max_{i,i' \in G} d(i, i')$$

or the largest average dissimilarity

$$\bar{d}_G = \frac{1}{N_G^2} \sum_{i,i' \in G} d(i, i')$$

- The hierarchical structure can be plotted in a dendrogram, which is a binary tree where the heights of the branches are equal to the dissimilarity between the corresponding clusters

Example: Wine data

See R code

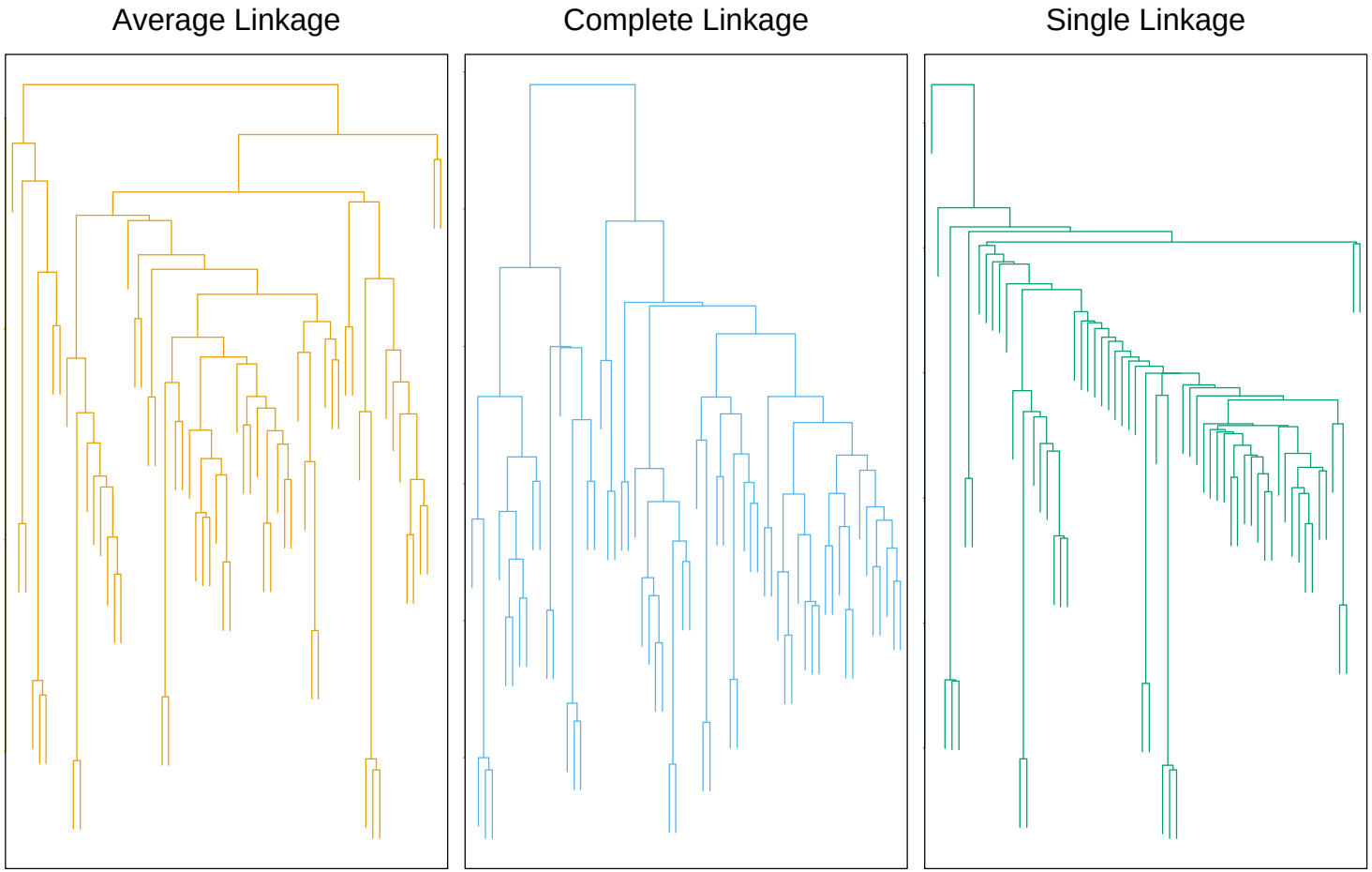



Illustration of dendrograms for the DNA microarray data