STAT 3690 Lecture 35

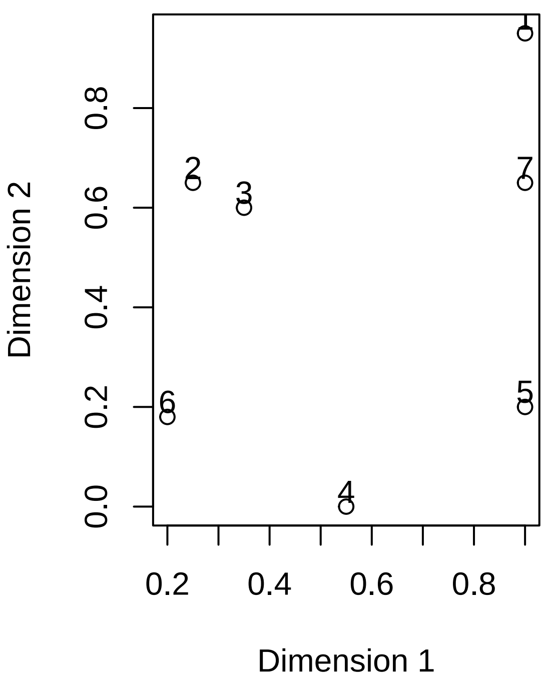
zhiyanggeezhou.github.io

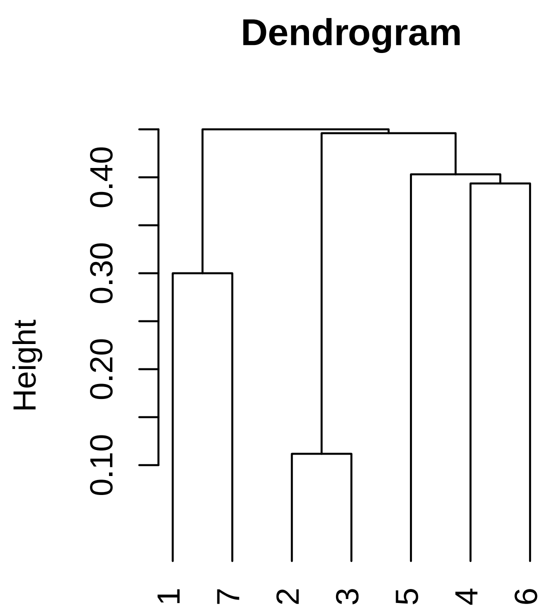
Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Apr 25, 2022

Hierarchical clustering

- A simple example
  - Step 1: {1}, {2}, {3}, {4}, {5}, {6}, {7};
  - Step 2: {1}, {2, 3}, {4}, {5}, {6}, {7};
  - Step 3: {1, 7}, {2, 3}, {4}, {5}, {6};
  - Step 4: {1, 7}, {2, 3}, {4, 5}, {6};
  - Step 5: {1, 7}, {2, 3, 6}, {4, 5};
  - Step 6: {1, 7}, {2, 3, 4, 5, 6};
  - Step 7: {1, 2, 3, 4, 5, 6, 7}.





- Dendrogram: a tree displaying a hierarchical sequence of clustering assignments
  - Node representing a group

        * Leaf node representing a singleton (i.e., a group containing a single data point)
        * Root node representing the group containing all the data points
        * Internal node: has two children nodes, representing the the groups that were merged to form it
    - Height: draw each internal node at a height proportional to the dissimilarity between its two children nodes (if fix the leaf nodes at height zero)
- Distances
    - Dissimilarity d_{ij}: (Euclidean) distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$
    - Linkage: distance between groups G and H
        * Options
            · Single linkage
$$d_{\text{single}}(G,H) = \min_{i\in G, j\in H} d_{ij}$$
            · Complete linkage
$$d_{\text{complete}}(G,H) = \max_{i\in G, j\in H} d_{ij}$$
            · Average linkage
$$d_{\text{average}}(G,H) = \frac{1}{n_G n_H} \sum_{i\in G, j\in H} d_{ij}$$
            · Centroid linkage
$$d_{\text{centroid}}(G,H) = \|\bar{\boldsymbol{x}}_G - \bar{\boldsymbol{x}}_H\|_2$$
            · Minimax linkage
$$d_{\text{minimax}}(G,H) = \min_{i\in G\cup H} \max_{j\in G\cup H} d_{ij}$$
        * Situation-dependent

- Example (hierarchical clustering for `iris`)

```
options(digits = 4)
d = dist(iris[,1:4])

tree.sing = hclust(d,method="single")
tree.comp = hclust(d,method="complete")
tree.avg = hclust(d,method="average")
tree.cent = hclust(d,method="centroid")

par(mfrow=c(2,2))
plot(tree.sing,hang=-1e-10, main='Single',xlab = '', sub = '')
plot(tree.comp,hang=-1e-10, main='Complete',xlab = '', sub = '')
plot(tree.avg,hang=-1e-10, main='Average',xlab = '', sub = '')
plot(tree.cent,hang=-1e-10, main='Centroid',xlab = '', sub = '')

# determine K for clustering with single linkage
x = iris[,1:4]
Ks = 2:20
Ws = numeric(length(Ks))
Bs = numeric(length(Ks))
CHs = numeric(length(Ks))

for(l in 1:length(Ks)){
  labs = cutree(tree.sing, k = Ks[l])
  nks = numeric(Ks[l])
  centers = matrix(0, nrow = Ks[l], ncol = ncol(x))
```

```
  for (k in 1:Ks[l]){
    nks[k] = nrow(x[labs == k,])
    centers[k,] = colMeans(x[labs == k,])
    Ws[l] = Ws[l]+sum(sweep(x[labs == k,], 2, centers[k,])^2)
  }
  Bs[l] = sum(nks * rowSums(sweep(centers, 2, colMeans(x))^2))
  CHs[l] = (Bs[l]/(Ks[l]-1))/(Ws[l]/(nrow(x)-Ks[l]))
}
plot(Ks, CHs,
  type="b", pch = 19,
  xlab="Number of clusters K",
  ylab="CH index")
```

Modern alternatives

- Density-based spatial clustering of applications with noise (DBSCAN, M. Ester, H. Kriegel, J. Sander, X. Xu (1996), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*.)
- t-distributed stochastic neighbor embedding (t-SNE, S. Roweis & G. Hinton (2002). *Conference on Neural Information Processing Systems (NIPS)*.)
- Uniform manifold approximation and projection for dimension reduction (UMAP, L. McInnes & J. Healy (2018), arXiv:1802.03426)