Classification
Dimension Reduction

Clustering and PCA

Introduction

- Here we are interested in Prediction but on discovering structures.
  - Are there natural groups?
  - Are there latent structures or dimensions?
- It is subjective and there is no preset goal in this analysis.
- It is difficult to assess the results of these analysis.
- It is also referred as unsupervised learning problem.

Principal Components Analysis

Principal Components Analysis (PCA)

- It helps in reducing dimensions hence called dimension reduction tool
- It uncovers the underlying directions of high variability
- We can use principal components thus discovered for supervised learning.
- In nutshell, PCA produces derived variables which are more efficient.
- PCA also serves as a tool for data visualization.

What are principal components?

- We can visualize a p dimensional feature set through a scatterplot by taking two variables at a time.
- Basically, our n observations in p dimensional space but not all dimensions are equally valuable.
- So we compose each of these variables with varying weights to derive new features.
- The new features, called as factors, has higher concentration of value in some features and lower in others.
- We discard low value concentration features and use only high value ones.

Principal Components, Mathematically

- $F_i = \emptyset_{1i}X_1 + \emptyset_{2i}X_2 + \cdots + \emptyset_{pi}X_\mathrm{p}$
  - Where $\emptyset_i$ is referred to as loadings of the principal component i
- The loadings are constraint so that the sum of square equals 1.
- We determine the optimal weights$\emptyset_i$ to determine the factors by solving for one factor at a time.
- The factors thus determined are uncorrelated with each other or equivalently orthogonal to each other.
- These factors need to be interpreted and labeled for further use in modeling.

More on PCA

- PCA is sensitive to scaling.
- All variables are centered at zero and scale to 1 standard deviation.
- In some cases where we scale of variable is not very diverse, we can avoid scaling.
- Only absolute values of the loading components are considered for determining whether a variable describes the factor or not.
- The proportion of variance explained determines the quality of factors extracted.
- Typically, we follow 1/3 variables explaining 2/3 variance.

More on PCA

- Typically we use scree plot to determine number of factors to retain.
- It is determined by eigen value of 1 or more.
- It is often the case that signal in a data is concentrated in first few principal components.

CLustering

Clustering

- Clustering is finding subgroups in data.
- We partition into groups with similar observations.
- Clustering is similar to Factor analysis except that Factor analysis groups similar variables and clustering groups similar observations.
- One of the important application is segmentation in marketing.
- There are primarily two types of clustering
  - K Means – We prespecify K clusters in advance
  - Hierarchical - We determine number of clusters after clustering.

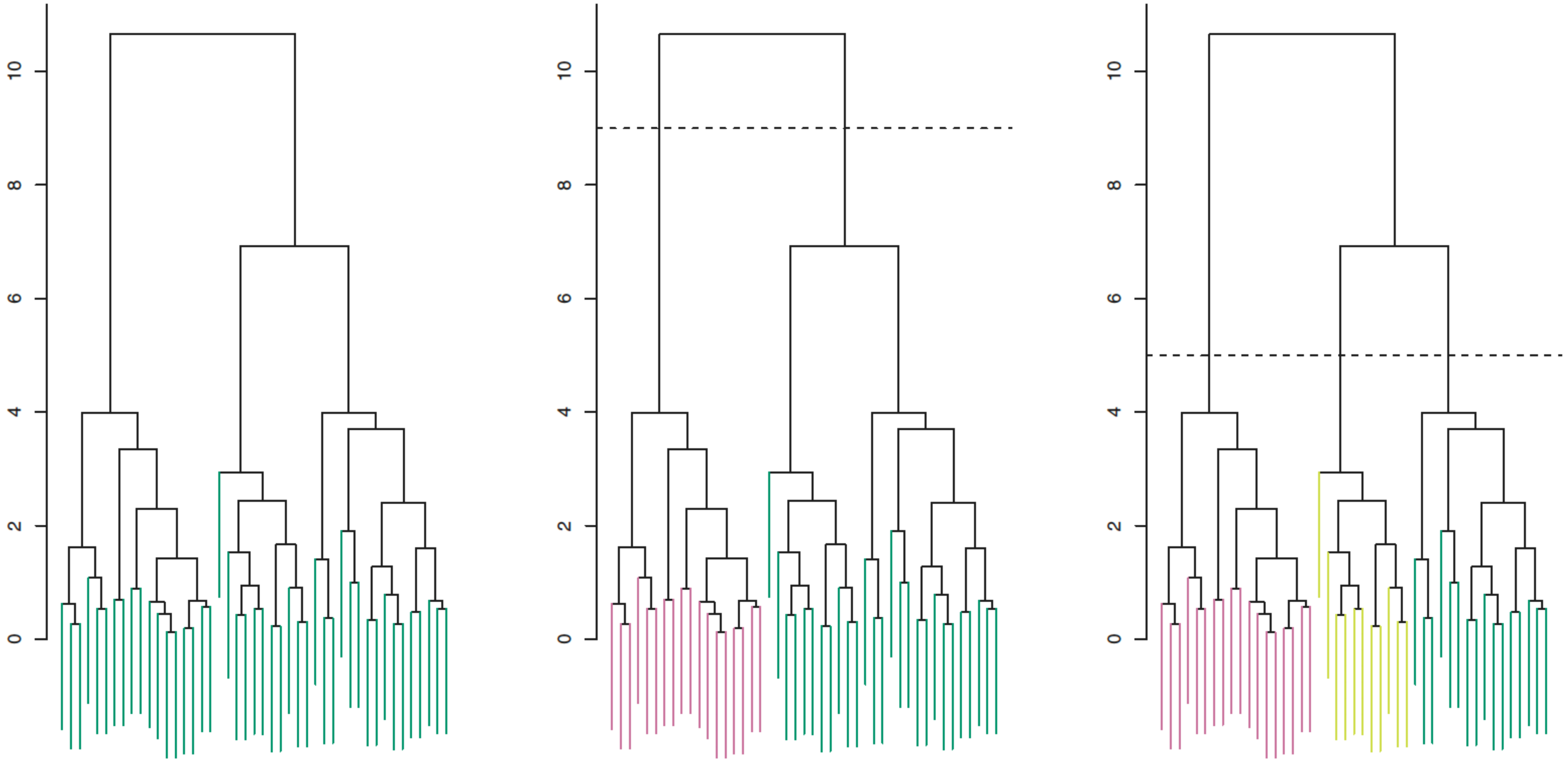
K Means Clustering

- Partitions data set into K distinct non-overlapping clusters.
- Typically we choose odd number of clusters.
- A good clustering is one where the within-cluster variation is as small as possible and across-cluster variation is as large as possible.
- It is measured by F Statistic given by $F = \frac{Across-cluster\ variance}{Within-cluster\ variance}$
- We run several trials with different number of clusters and choose the one with highest value of F.

Hierarchical Clustering

- It is a non-parametric clustering method.
- There is no need to specify K in advance.
- It is not sensitive to outliers or any data distribution.
- It is tree based and as an output produces Dendrogram.
- The tree has a single all encompassing group on top and each observation as leave of the tree.

Dendrogram



Decisions to be made in Clustering

- There are many decisions to be made subjectively. Such as
  - Kmeans
    - How many clusters to specify
    - Should we standardize or not
  - Hierarchical
    - What dissimilarity measure to use
    - What linkage to use
    - Where do we cut the dendrogram