STATISTICAL METHODS IN DATA MINING

DR. ALPER VAHAPLAR

Clustering in R

- ✓ Clustering Examples in R
- ✓ Execute the following commands one by one in R (or R Studio)

```
› ali = c(24, 80, 172)
› veli = c(24, 80, 172)
› ayse = c(24, 50, 172)
› fatma = c(22, 40, 160)
› d = rbind(ali, veli)
› d = rbind(d, ayse)
› d = rbind(d, Fatma)
› dist(d, method = "minkowski", p = 3, upper = TRUE,
  diag=TRUE)
```

Clustering in R

- ✓ Hierarchical Clustering Examples in R
- ✓ Execute the following commands one by one in R (or R Studio)

```
› uz = dist(d, method = "euclidean", upper = TRUE, diag=TRUE)
› #Hierarchical agglomerative clustering
› kume = hclust(uz, method="single")
› plot(kume)
› kume$height
› kume$order
› kume$merge
```

Clustering in R

- ✓ Hierarchical Clustering Examples in R
- ✓ Execute the following commands one by one in R (or R Studio)

```
› rm(list=ls()) #clear all variables
› data(iris) # Loading iris data...
› sapply(iris[,1:4], fivenum) # five number summary
› summary(iris[,1:4])
› x = iris[sample(nrow(iris), 30, replace = FALSE),] #sample 30
  rows
› d =dist(x[,1:4]) # calculate distance matrix
› kume = hclust(d, method = "complete") #hierachical cluster
› plot(kume)
› k = cutree(kume, k=3) # stop in 3 clusters
› table(x$Species, k) # compare with the actual classes
```

Clustering in R

- ✓ K-means Clustering Examples in R
- ✓ Execute the following commands one by one in R (or R Studio)

```
› rm(list=ls()) #clear all variables
› data(iris) # Loading iris data...
› x = iris[sample(nrow(iris), 30, replace = FALSE),] #sample
  30 rows
› # k-means clustering
› kume = kmeans(x[,1:4], centers=3) #k-means cluster with k=3
› kume$centers
› kume$cluster
› kume$size
```

Clustering in R

- ✓ K-means Clustering Examples in R
- ✓ Execute the following commands one by one in R (or R Studio)

```
› kume$withinss
› kume$tot.withinss #sum(kume$withinss)
› kume$betweenss
› kume$totss # kume$tot.withinss + kume$betweenss
› table(kume$cluster, x$Species) # compare with the actual
  classes
```

Clustering in R

- ✓ K-means Clustering Examples in R
- ✓ Which value of "**k**" is the best in k-means?
- ✓ `Smallest kume$tot.withinss is better`

```
› twss=sapply(1:10,function(k){kmeans(x[,1:4], k, nstart = 20,
  iter.max = 20)$tot.withinss})

› plot(1:10, twss, type= "b", xlab = "Number of clusters(k)",
  ylab = "Within cluster sum of squares")
```