CSCI446/946 Big Data Analytics

Week 4 Statistical Methods and Cluster Analysis

School of Computing and Information Technology
University of Wollongong Australia

Brief Recap

Last week:

- Introduction to R
- Exploratory Data Analysis
  - Visualization
- Descriptive statistics

Today's lecture

- Statistical Methods for Evaluation
  - Hypothesis Testing, ANOVA
- Clustering
  - Overview
  - K-means clustering
    - Overview of the Method
    - Determining the Number of Clusters
    - Diagnostics
    - Reasons to Choose and Cautions
  - DBScan clustering
  - Additional Algorithms

Statistical Methods for Evaluation

- Statistics is crucial because it may exist throughout the entire Data Analytics Lifecycle
  - Initial data exploration and data preparation
  - Model planning and model building
    - Best input variables, predictability
  - Evaluation of the final models
    - Accuracy, better than guess or another one?
  - Assessment of the new models when deployed
    - Sound prediction? Have desired effect?

Statistical Methods for Evaluation

- Hypothesis Testing
  - Form an assertion and test it with data
  - Common assumption (there is no statistically significant difference)
    - Null hypothesis (H_0) vs Alternative hypothesis (H_A)
- Example: identify the effect of drug A compared to drug B on patients
  - What are the H_0 and H_A ?
- A hypothesis is formed before validation
  - It can define expectations.

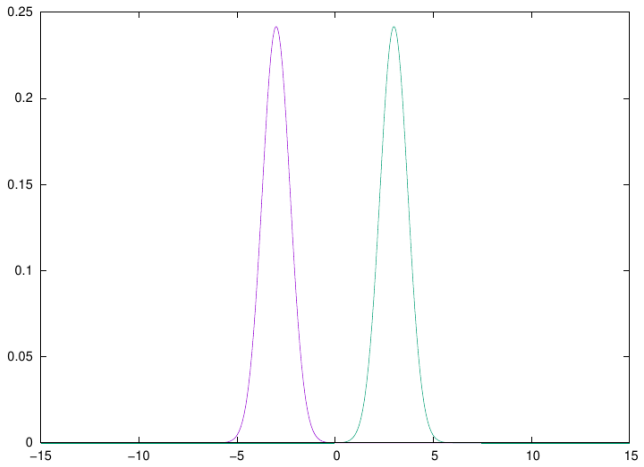
Statistical Methods for Evaluation

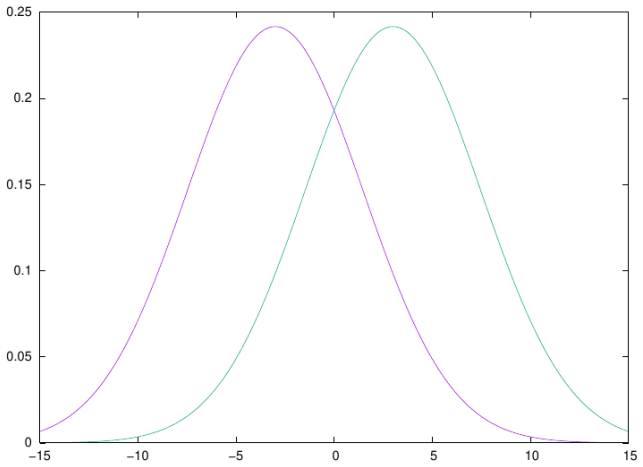
- Hypothesis Testing
  - Clearly state Null and Alternative hypotheses
  - **Either** reject the null hypothesis in favour of the alternative **or** not reject the null hypothesis

Application	Null Hypothesis	Alternative Hypothesis
Accuracy Forecast	Model X *does not predict* better than the existing model.	Model X *predicts* better than the existing model.
Recommendation Engine	Algorithm Y *does not produce* better recommendations than the current algorithm being used.	Algorithm Y *produces* better recommendations than the current algorithm being used.
Regression Modeling	This variable *does not affect* the outcome because its coefficient is *zero*.	This variable *affects* outcome because its coefficient is not *zero*.

Statistical Methods for Evaluation

- Difference of Means (A common hypothesis test)
  - Assume we have two populations, one with mean=-3 and the other with mean=3
    - By comparing the means can we say that the difference between the two populations is significant?
    - Answer depends on variance.





Statistical Methods for Evaluation

- Student's *t*-test
  - Assumes that distributions of the two populations have equal but unknown variance.
  - Assumes that each population is normally distributed.

Signal

Noise

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

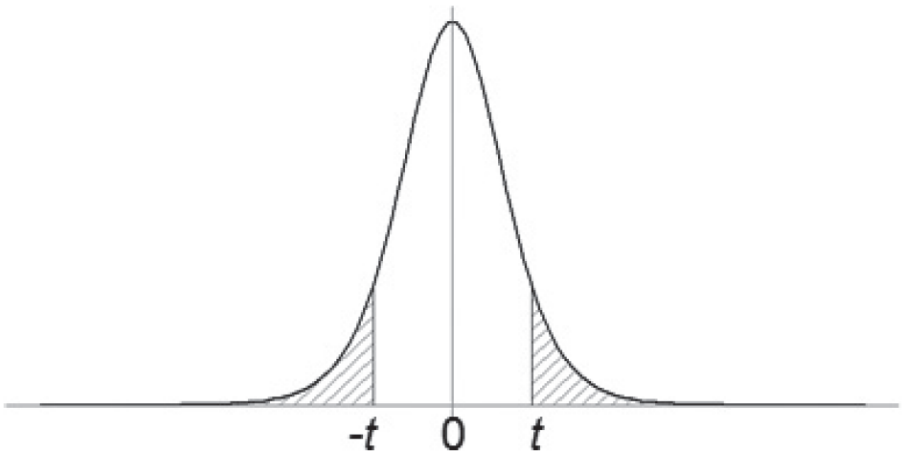
T (the *t-statistic*) follows a *t-distribution* with (n_1+n_2-2) degree of freedom

Statistical Methods for Evaluation

- Student's *t*-test
  - The further T is from zero the more significant the difference between the populations. If T is large then one would reject the null hypothesis

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

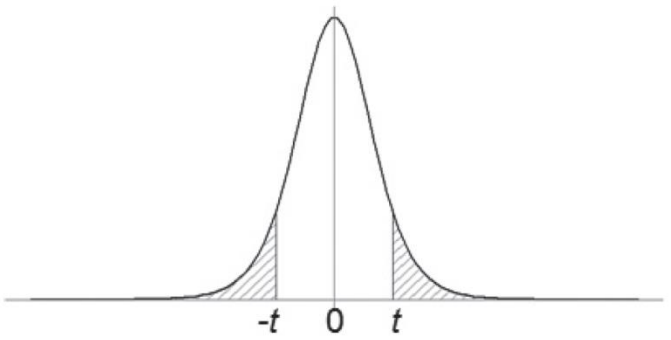


Statistical Methods for Evaluation

- Student's *t*-test
  - Significance level of the test (α): the probability of rejecting the null hypothesis, when the null hypothesis is actually TRUE
    - It is common to use $\alpha = 0.05$
  - Find T* such that $P(|T| \geq T^*) = \alpha$
  - Reject H_0 if $|T| \geq T^*$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$



Statistical Methods for Evaluation

- Student's *t*-test (an example)

```
# generate random observations from the two populations
x <- rnorm(10, mean=100, sd=5)     # normal distribution centered at 100
y <- rnorm(20, mean=105, sd=5)     # normal distribution centered at 105

t.test(x, y, var.equal=TRUE)       # run the Student's t-test
Two Sample t-test

data:  x and y
t = -1.7828, df = 28, p-value = 0.08547
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -6.1611557  0.4271893
sample estimates:
  mean of x mean of y
102.2136  105.0806
```

Statistical Methods for Evaluation

- Welch's *t*-test
  - Shall be used when the equal population variance assumption is NOT justified
  - It uses the sample variance for each population instead of the pooled sample variance
  - Still assumes two populations are normal.

$$T_{welch} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Statistical Methods for Evaluation

- Welch's *t*-test

```
t.test(x, y, var.equal=FALSE)       # run the Welch's t-test

Welch Two Sample t-test

data:  x and y
t = -1.6596, df = 15.118, p-value = 0.1176
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -6.546629  0.812663
sample estimates:
  mean of x mean of y
102.2136  105.0806
```

Statistical Methods for Evaluation

- Recommended viewing: An excellent video tutorial on the t-test can be found here:
  - https://www.youtube.com/watch?v=pTmLQvMM-1M

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test
  - What if the two populations are not normal?
- Parametric test (i.e. student's t-test) vs Nonparametric test (i.e. Wilcoxon rank-sum test)
  - Parametric test
    - Makes assumptions about the population distributions from which the samples are drawn
  - Nonparametric test
    - Shall be used if the populations cannot be assumed (or transformed) to be normal

Statistical Methods for Evaluation

- Wilcoxon Rank-Sum Test

```
wilcox.test(x, y, conf.int = TRUE)

Wilcoxon rank sum test

data:  x and y
W = 55, p-value = 0.04903
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  -6.2596774 -0.1240618
sample estimates:
  difference in location
-3.417658
```

p-value: the probability of the rank-sums of this magnitude being observed assuming that the population distributions are identical

Statistical Methods for Evaluation

- Type I and Type II Errors
  - Type I error: the rejection of the null hypothesis when the null hypothesis is TRUE
  - The probability of type I error is denoted by α
  - Type II error: the acceptance of the null hypothesis when the null hypothesis is FALSE
  - The probability of type II error is denoted by β
- Power (statistical power)
  - The probability of correctly rejecting the null hypothesis (1- β)

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - What if there are more than two populations?
  - Multiple *t*-test may not perform well then.
- A generalization of the hypothesis testing
  - ANOVA tests if any of the population means differ from the other population means
  - Each population is assumed to be normal and have the same variance

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)

$$H_o: \mu_1 = \mu_2 = \ldots = \mu_n$$

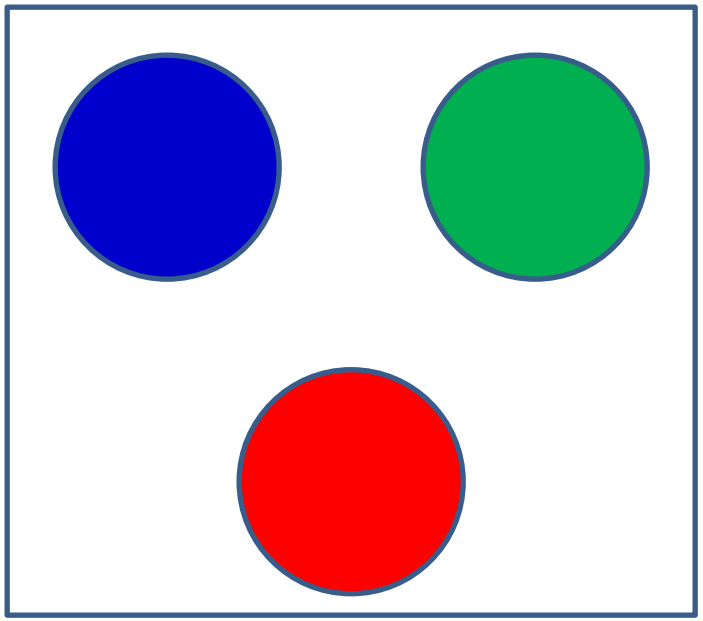
$$H_A: \mu_i \neq \mu_j \text{ for at least one pair of } i, j$$

- Compute *F*-test statistic
  - Between-groups mean sum of squares
  - Within-groups mean sum of squares

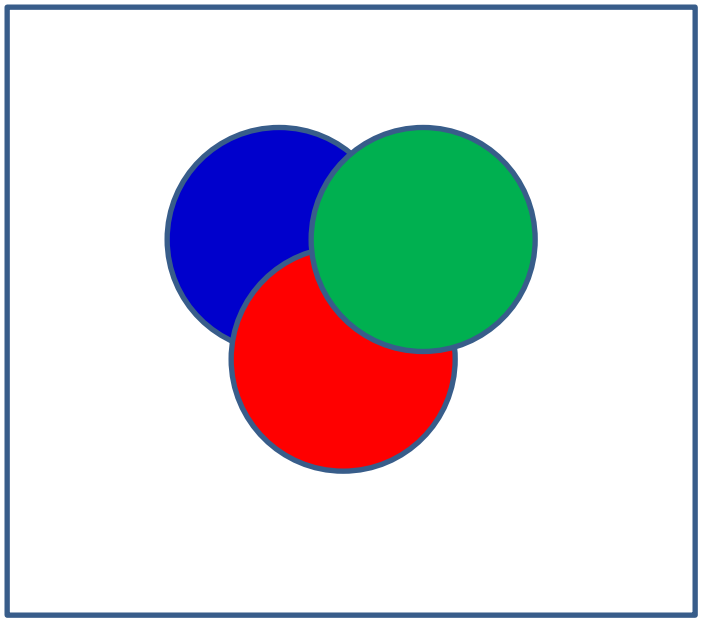
$$S_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x}_0)^2 \qquad S_W^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)



$$F = \frac{S_B^2}{S_W^2}$$



$$S_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \cdot (\bar{x}_i - \bar{x}_0)^2$$

$$S_W^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - Measures how different the means are relative to the variability within each group
  - The larger the *F*-test statistic, the greater the likelihood that the difference of means are due to something other than chance alone
  - The *F*-test statistic follows an *F*-distribution

$$F = \frac{S_B^2}{S_W^2}$$

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)

```
# fit ANOVA test
model <- aov(purchase_amt ~ offers, data=offertest)

summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
offers        2 225222  112611   130.6 <2e-16 ***
Residuals   497 428470     862
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Shall we accept or reject the null hypothesis?

Statistical Methods for Evaluation

- ANOVA (Analysis of Variance)
  - Additional tests for each pair of groups
  - Tukey's Honest Significant Difference (HSD)

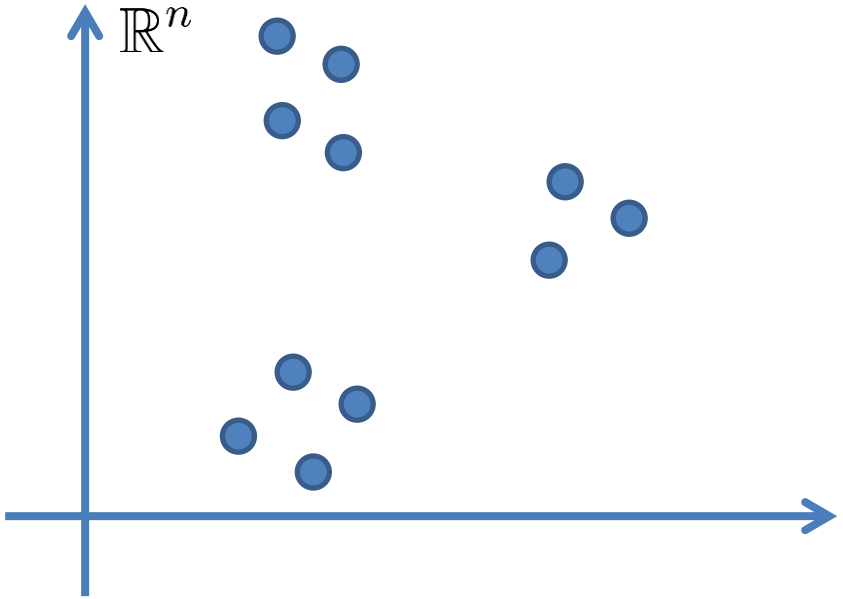
```
TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = purchase_amt ~ offers, data = offertest)

$offers
                     diff        lwr      upr     p adj
offer1-nopromo 40.961437 33.4638483 48.45903 0.0000000
offer2-nopromo 48.120286 40.5189446 55.72163 0.0000000
offer2-offer1   7.158849 -0.4315769 14.74928 0.0692895
```

Overview of Clustering

- Supervised vs. Unsupervised Techniques
  - Labelled data vs. Unlabelled data
- Unsupervised Techniques
  - Refers to the problem of finding hidden structure within unlabelled data
  - Clustering, density estimation, dimensionality reduction, etc.
- Clustering is an unsupervised technique

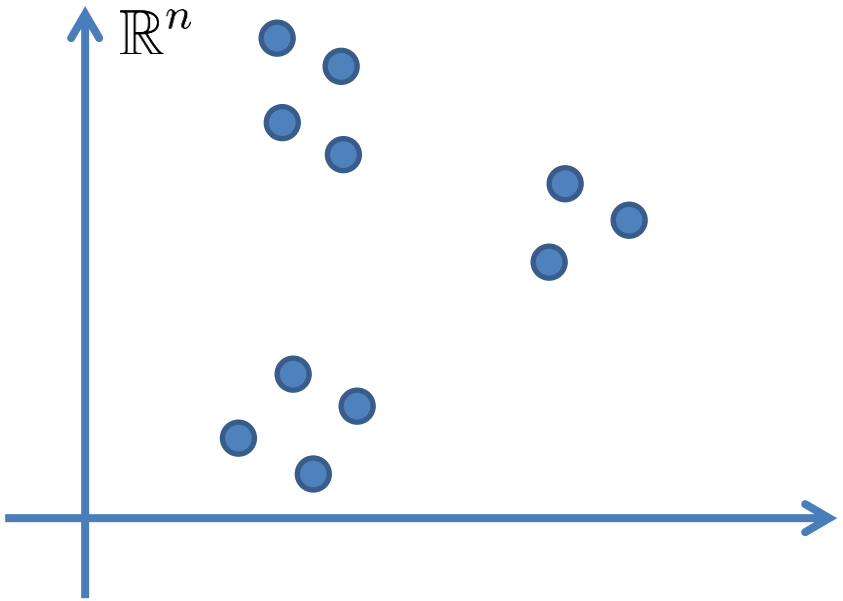
K-means Clustering

- Given a collection of m objects each with n measurable attributes
  - Mathematically, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m \in \mathbb{R}^n$
  - Each object is a point in an n-dimensional space



K-means Clustering

- For a chosen value of k, identify k clusters of objects based on the objects' proximity to the centre of the k groups



K-means Clustering

- Use Cases
  - Often used as a lead-in to classification
  - Once clusters are identified, labels can be applied to each cluster to do classification
- Applications
  - Image Processing
  - Medical (Clustering patients)
  - Customer grouping (find similar customers)

K-means Clustering

- Application to image processing
  - i.e. cluster colour space

Original image

$K = 2$

$K = 3$

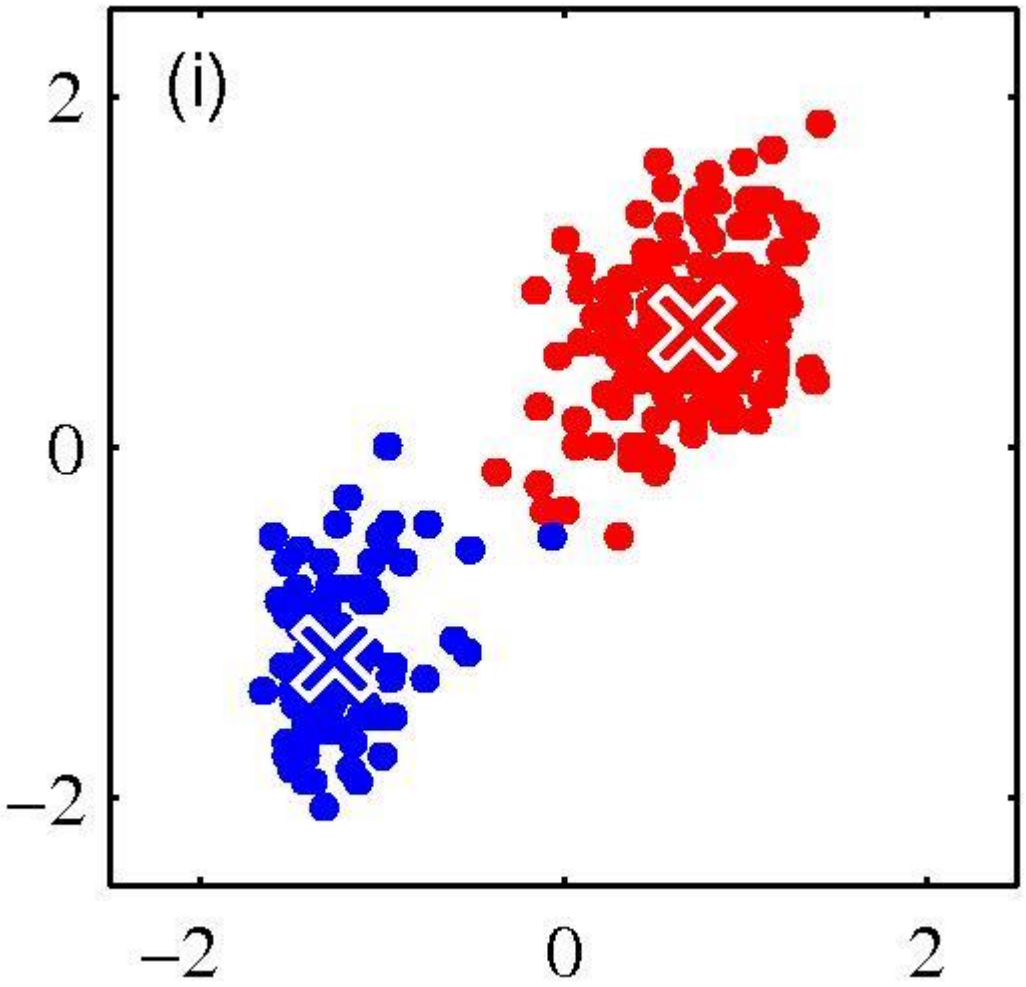
$K = 10$

Overview of K-means Clustering

- Four steps
  1. Choose a value of k, create k centroids, then initialize them by guessing their value.
  2. Compute the distance from each data point to each centroid. Assign each point to the closest centroid.
  3. Update the centroid of each cluster to become the center of gravity of the cluster.

  Repeat Steps 2 and 3 until convergence

Overview of K-means Clustering



Overview of K-means Clustering

- Compute the Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Compute the center of gravity

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^{m} \mathbf{x}_i}{m}$$

Determine the Number of Clusters

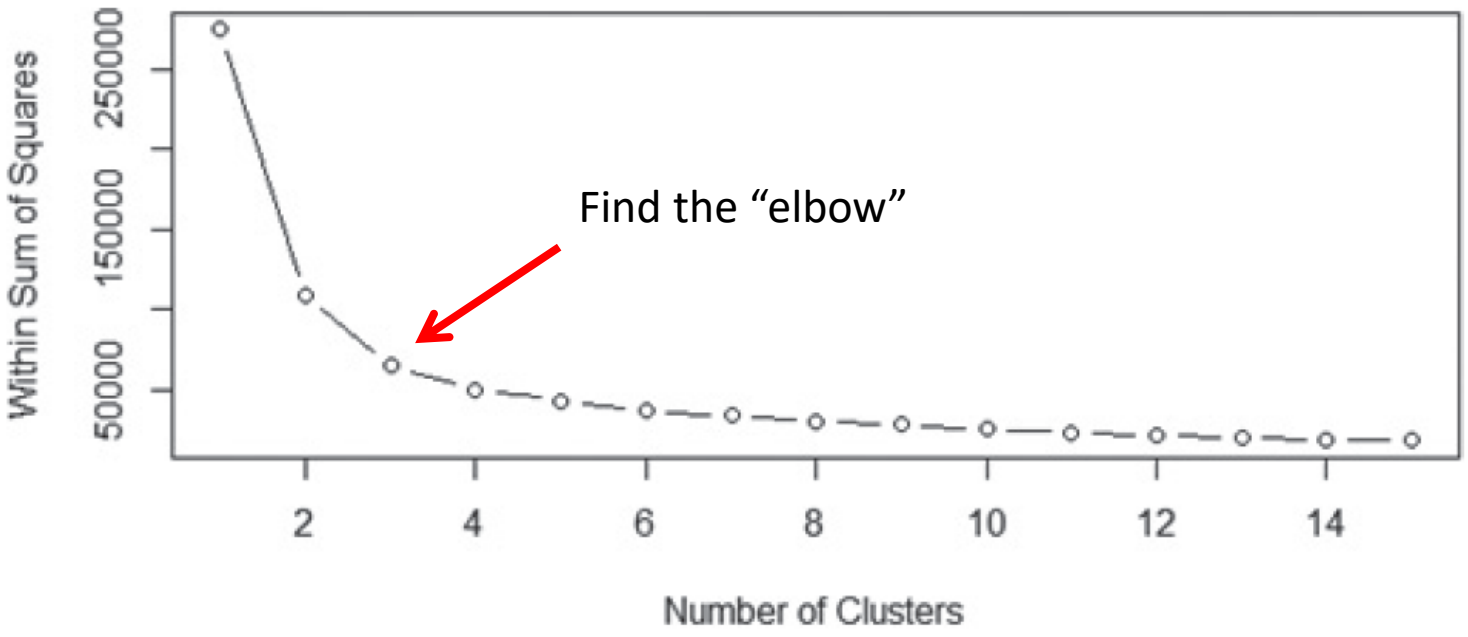
- What value of k shall be selected?
  - A reasonable guess, some predefined requirement
  - k-1, k, or k+1?
- Within Sum of Squares (WSS)
  - A heuristic
  - Sum of the squares of the distances between each data point and the closest centroid

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|_2^2; \quad r_{ij} \in \{0, 1\}$$

Determine the Number of Clusters

- Within Sum of Squares (WSS)

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|_2^2; \quad r_{ij} \in \{0, 1\}$$



Using R to Perform K-mean Clustering

- Task is to
  - Group 620 high school seniors based on their grades in “English”, “Math”, and “Science”

```
library(plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(graphics)
library(grid)
library(gridExtra)

#import the student grades
grade_input = as.data.frame(read.csv("c:/data/grades_km_input.csv"))
```

Using R to Perform K-mean Clustering

- Task is to
  - Group 620 high school seniors based on their grades in “English”, “Math”, and “Science”

```
kmdata_orig = as.matrix(grade_input[,c("Student","English", "Math","Science")])
kmdata <- kmdata_orig[,2:4]
```

```
kmdata[1:10,]
```

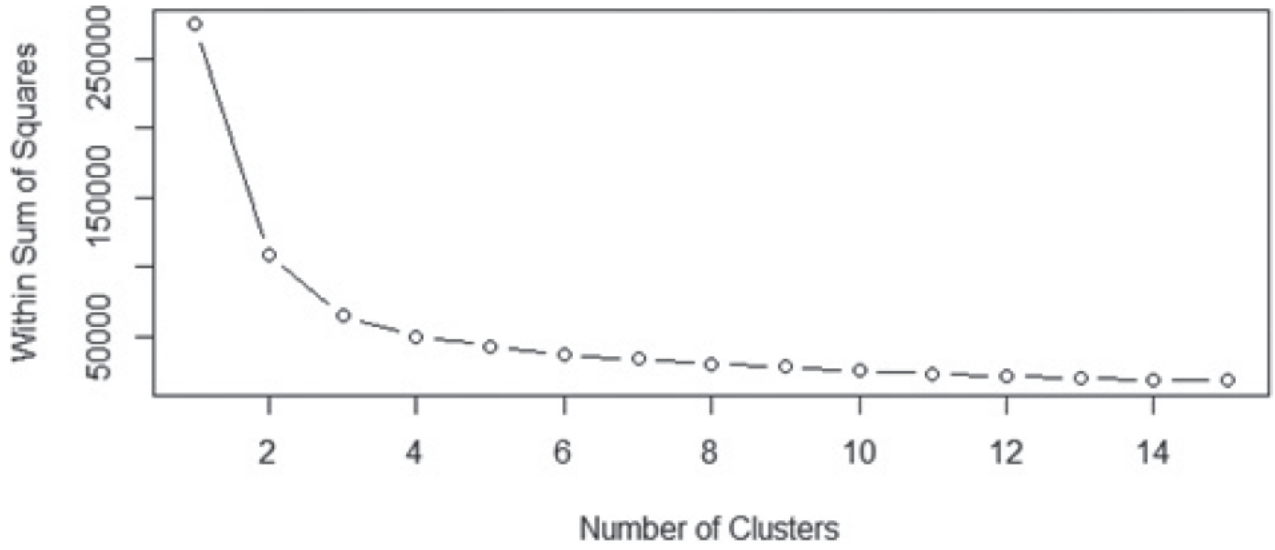
```
      English Math Science
 [1,]      99   96      97
 [2,]      99   96      97
 [3,]      98   97      97
 [4,]      95  100      95
 [5,]      95   96      96
 [6,]      96   97      96
 [7,]     100   96      97
 [8,]      95   98      98
 [9,]      98   96      96
[10,]      99   99      95
```

Using R to Perform K-mean Clustering

- Compute and plot WSS to choose k value

```
wss <- numeric(15)
for (k in 1:15) wss[k] <- sum(kmeans(kmdata, centers=k, nstart=25)$withinss)
```

```
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within Sum of
Squares")
```



Using R to Perform K-means Clustering

- Perform K-means Clustering

```
km = kmeans(kmdata,3, nstart=25)
km

K-means clustering with 3 clusters of sizes 158, 218, 244

Cluster means:
   English     Math  Science
1 97.21519 93.37342 94.86076
2 73.22018 64.62844 65.84862
3 85.84426 79.68033 81.50820

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
 [41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
 [81] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 1 1
[121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      3 3 3 3 3 3 3 3 3 3
```

Using R to Perform K-means Clustering

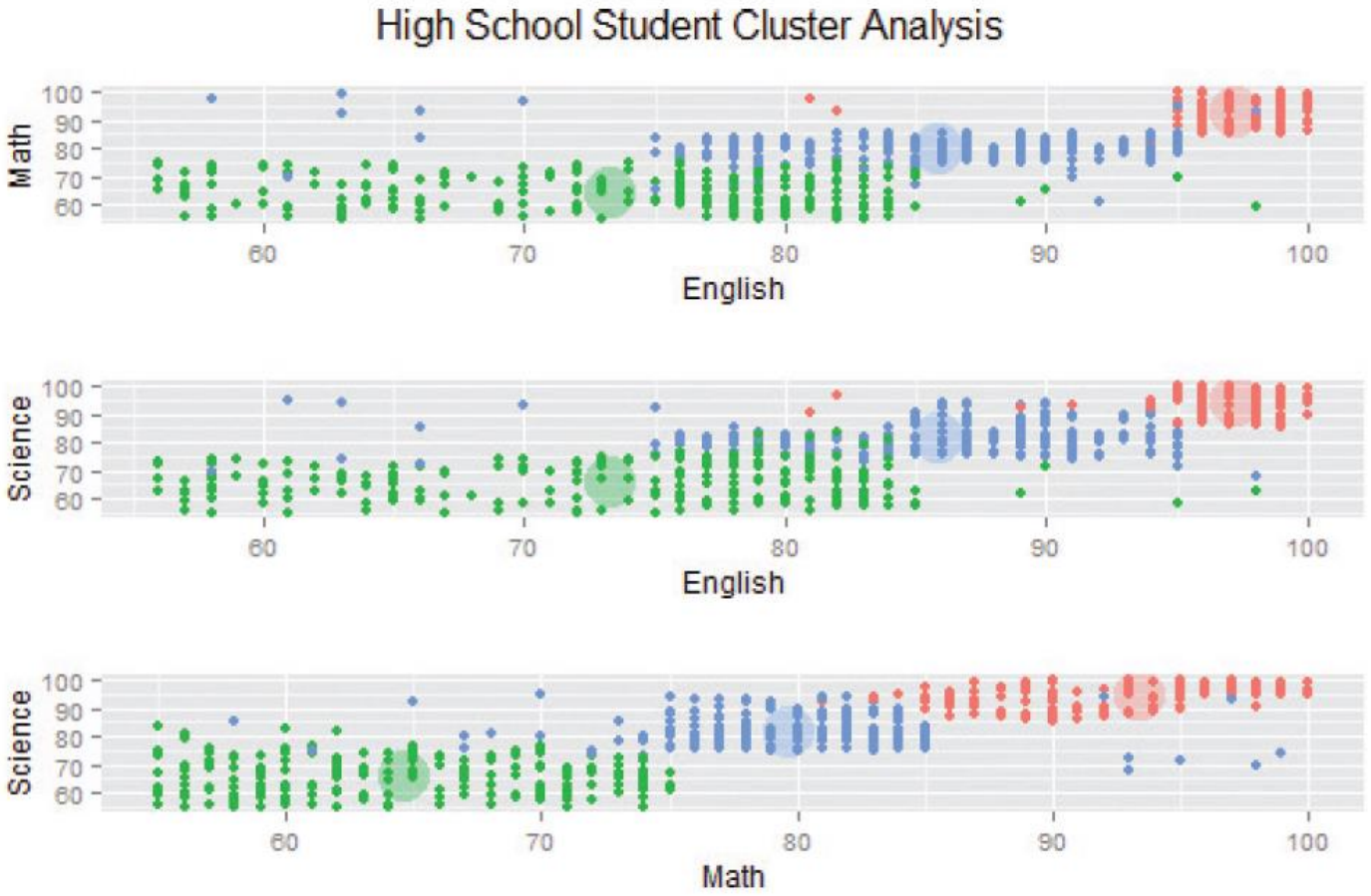
- Perform K-means Clustering

```
[521] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
      2 2 2 2 2 2 2 2 2 2
[561] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
      2 2 2 2 2 2 2 2 2 2
[601] 3 3 2 2 3 3 3 3 1 1 3 3 3 2 2 3 2 3 3 3

Within cluster sum of squares by cluster:
[1]  6692.589 34806.339 22984.131
 (between_SS / total_SS =  76.5 %)

Available components:

[1] "cluster"   "centers"  "totss"    "withinss"   "tot.withinss"
[6] "betweenss" "size"     "iter"     "ifault"
```

```
c( wss[3] , sum(km$withinss) )

[1] 64483.06 64483.06
```

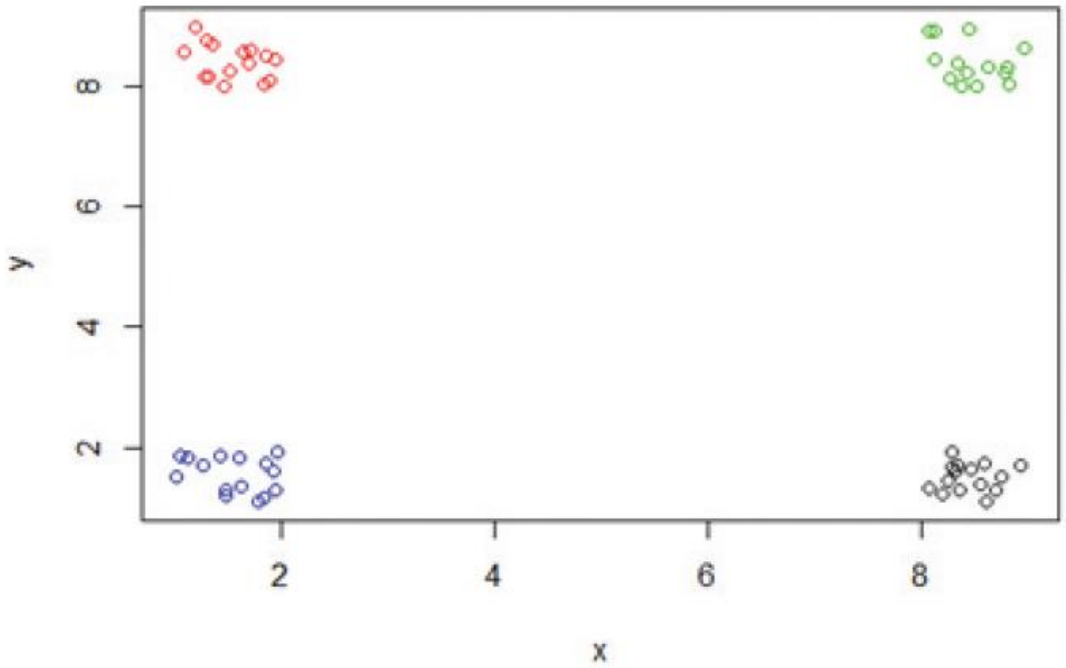
Using R to Perform K-means Clustering

- Visualize the identified clusters and centroids



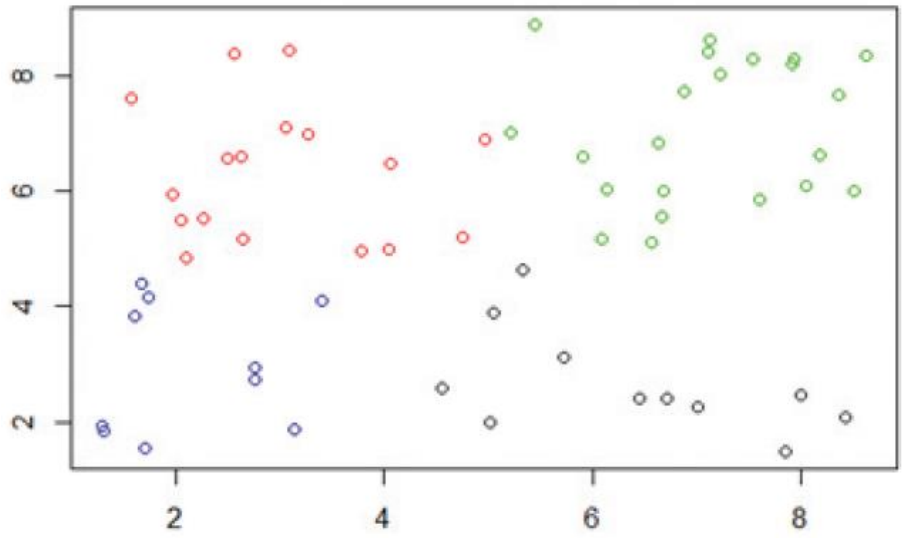
Diagnostics

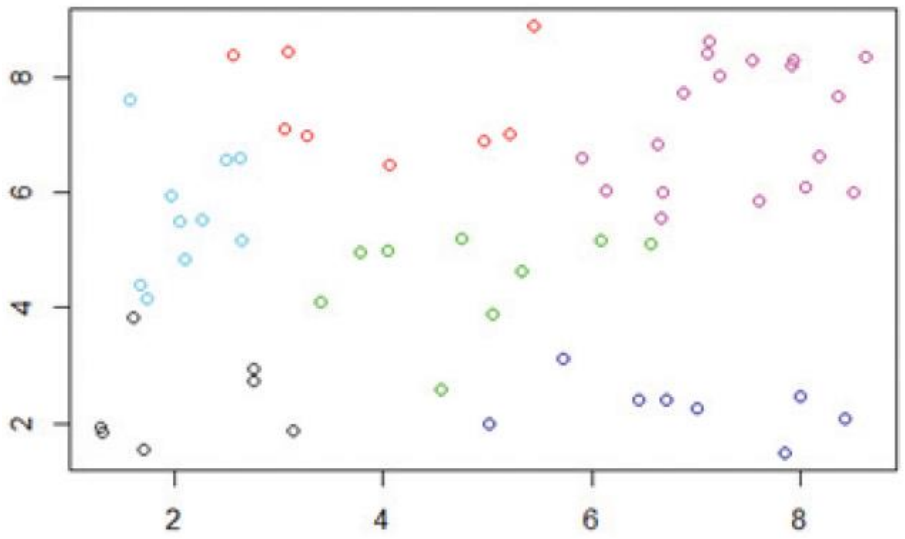
- The following questions shall be asked
  - Are the clusters well separated from each other?
  - Do any of the clusters have only a few points?
  - Do any of the centroids appear to be too close to each other?



Diagnostics

- A principle
  - If using more clusters does not better distinguish the groups, it is almost certainly better to go with fewer clusters





Reasons to Choose and Cautions

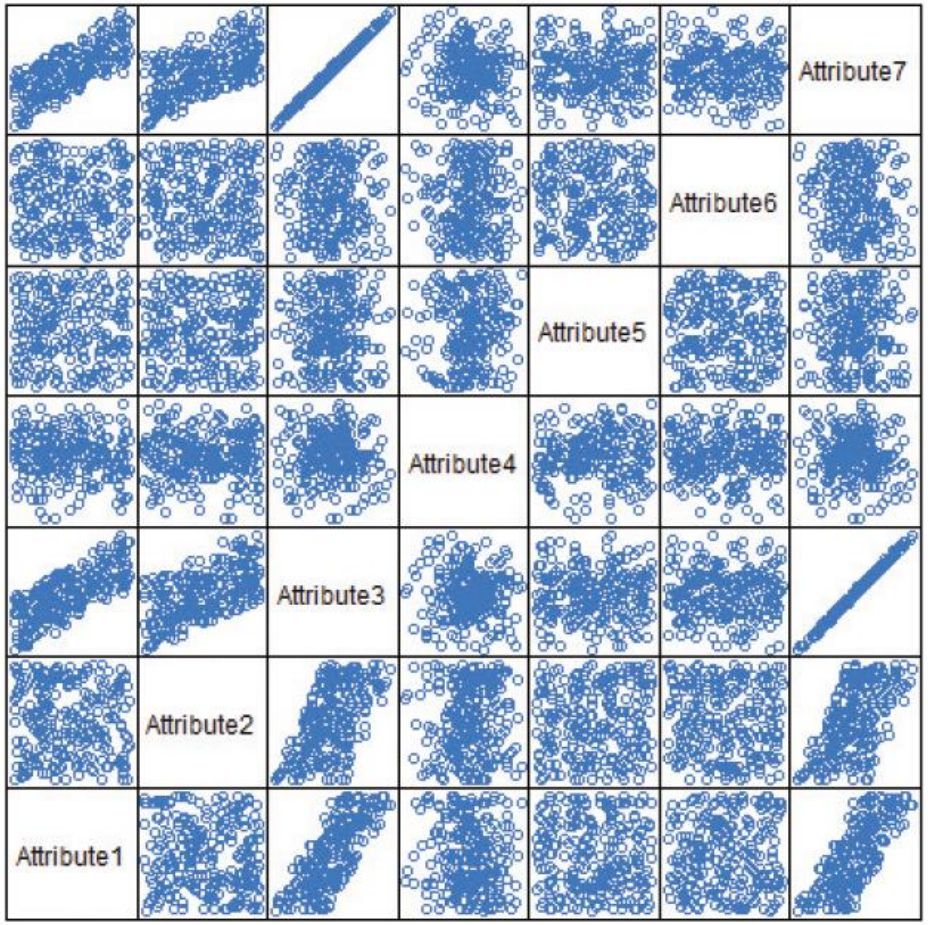
- Several decisions that must be made
  - What object attributions shall be included in clustering analysis?
  - What unit of measure shall be used for each attribute?
  - Do the attributes need to be rescaled?
    - One attribute could have a disproportionate effect

Reasons to Choose and Cautions

- Object attributes
  - Whether it will be known for a new object?
  - Best to reduce the number of attributes to the extent of possible
    - Avoid using too many variables (Why?)
    - Avoid using several similar variables (Why?)
- Identify any highly correlated attributes
- Feature selection: Information gain, PCA, etc.

Reasons to Choose and Cautions

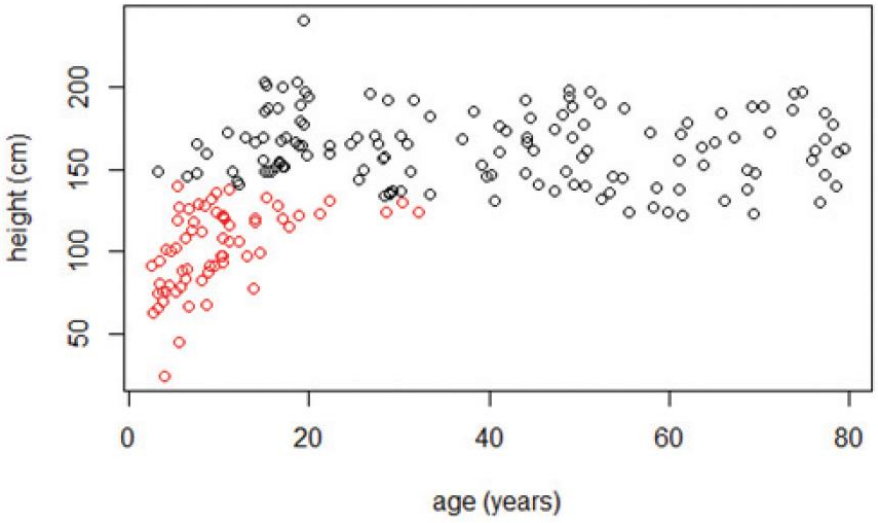
- Identify any highly correlated attributes

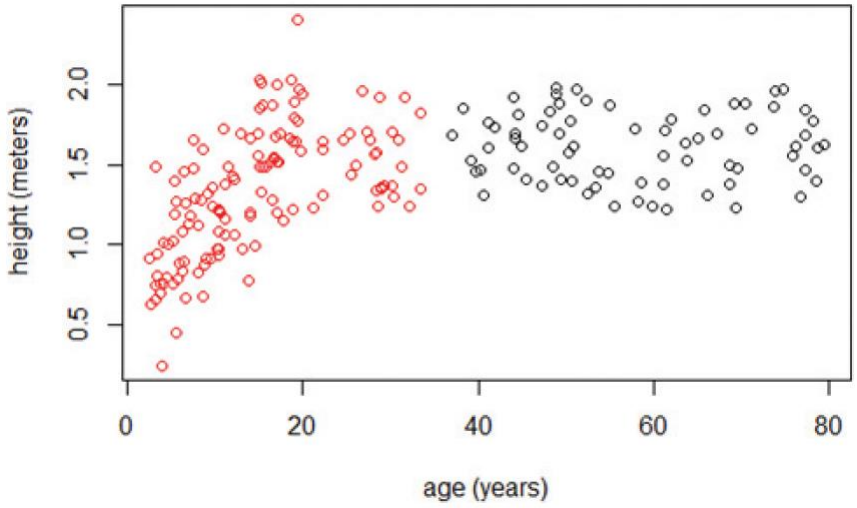


What is your observation?

Reasons to Choose and Cautions

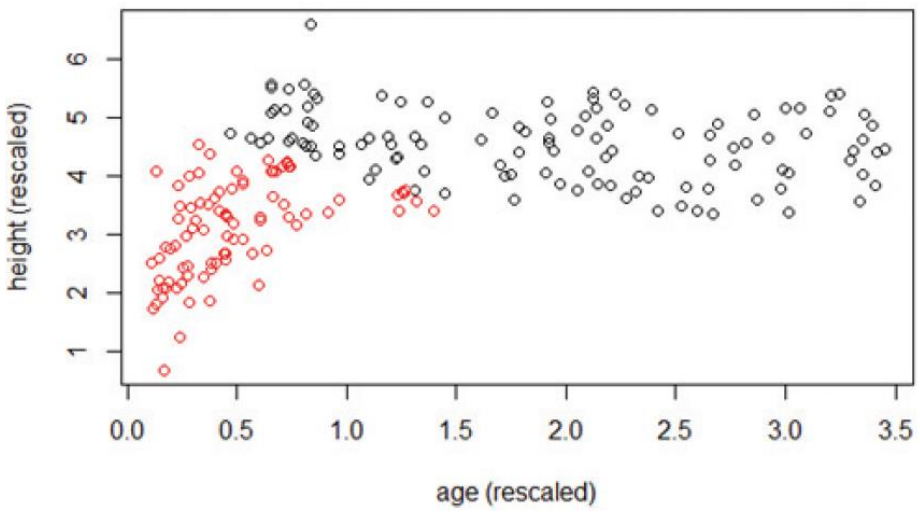
- Units of measure could affect clustering result

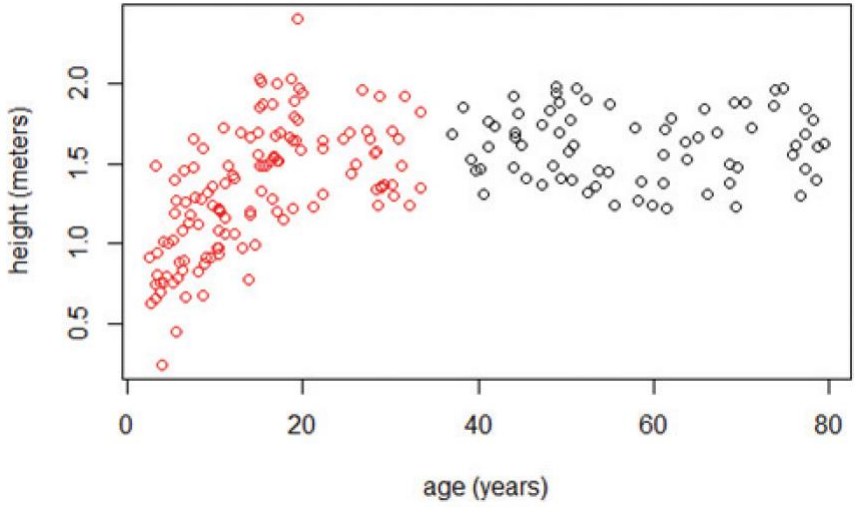




Reasons to Choose and Cautions

- Rescaling attributes affect clustering result
  - Divide each attribute by its standard deviation





Additional Considerations

- K-means clustering is sensitive to the starting positions of the initial centroids
  - Usually, we run the k-means clustering several times for a particular k value to choose the clustering result with the lowest WSS value
  - Implemented by the `nstart` option in `kmeans()`
- Other distances
  - Manhattan distance & the median of cluster

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{n} |x_i - y_i|$$

Additional Algorithms

- K-means clustering is easily applied to numeric data where the concept of distance can naturally be applied
- K-modes handles categorical data
  - Use the number of differences in the respective components of the attributes
    - What is the distance between (a,b,e,d) and (d,d,d,d)?
  - Implemented by the `kmode()` function
  - Caution: Sometimes it is better to convert categorial (or symbolic) data to numerical i.e. {hot, warm, cold} to {1,0,-1}.
    - Understand why!
    - Understand how to encode categorial values.

Additional Considerations

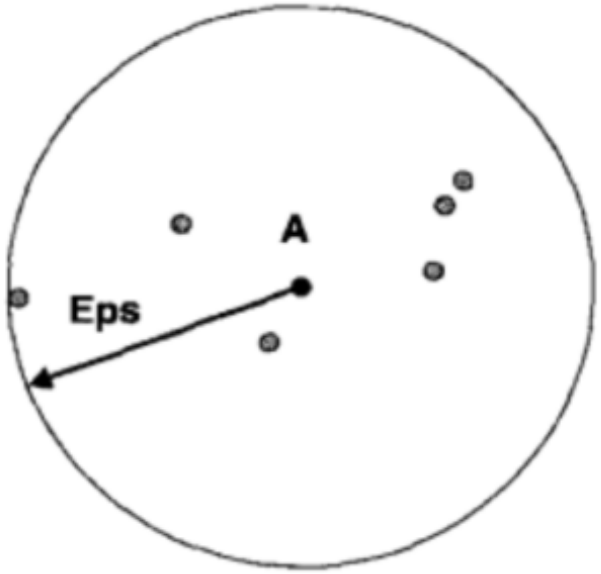
- Despite its popularity, K-means has problems:
  - When data contains noise and/or outliers
  - When clusters have non-globular shapes
  - When clusters vary in densities
  - When clusters differ significantly in size
  - Can reveal “empty” clusters
- Know your data (i.e via visualization) to verify whether K-means is suitable

Density Based Clustering

- Density-based clustering locates regions of high density that are separated from one another by regions of low density.
- In other words, clusters are dense regions in the data space, separated by regions of lower object density
- Major features of density-based clustering:
  - Discover clusters of arbitrary shape
  - Handle noise
  - Need density parameters as termination condition
- Example: DBSCAN

DBScan

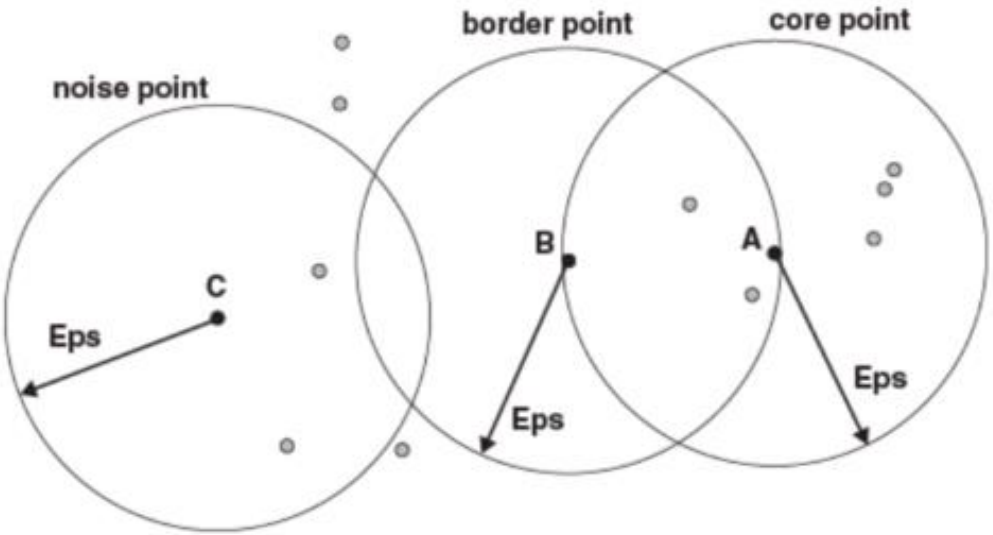
- Density is estimated for a particular point in the data set by counting the number of points within a specified radius, *Eps*, of that point. This includes the point itself.



- Example: the number of points within a radius of *Eps* of point A is 7, including A itself.
  - The density of A is 7.

DBSCAN

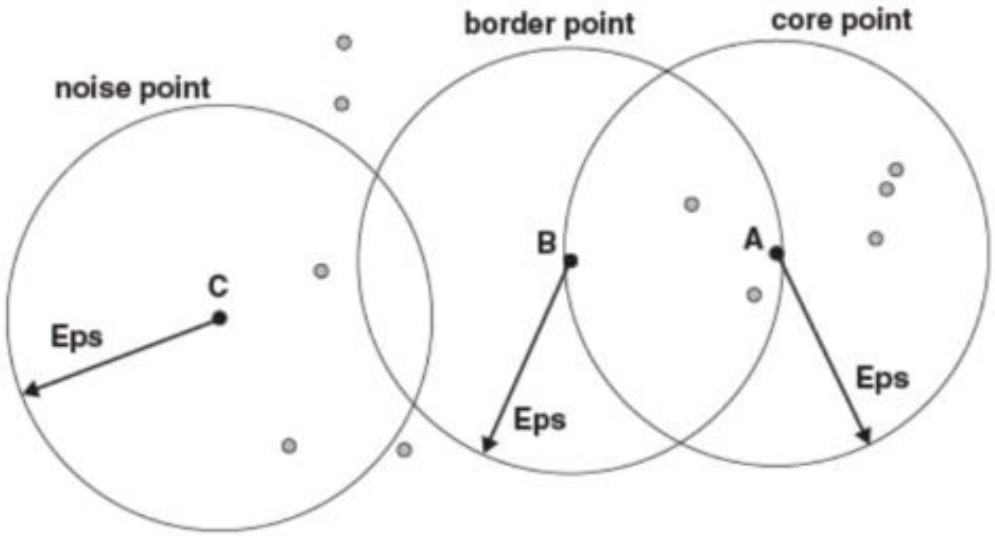
- Given a density threshold ($MinPts$) and a radius (Eps), the points in a dataset are classified into three types: core point, border point, and noise point.
  - Core points: Point whose density >= $MinPts$
    - Core points are in the interior of a density-based cluster.



Example: If MinPts = 6 then A is a core point because its density = 7 (7>6)

DBSCAN

- Three types: core point, border point, and noise point.
  - A border point is not a core point but falls within the neighborhood of a core point.

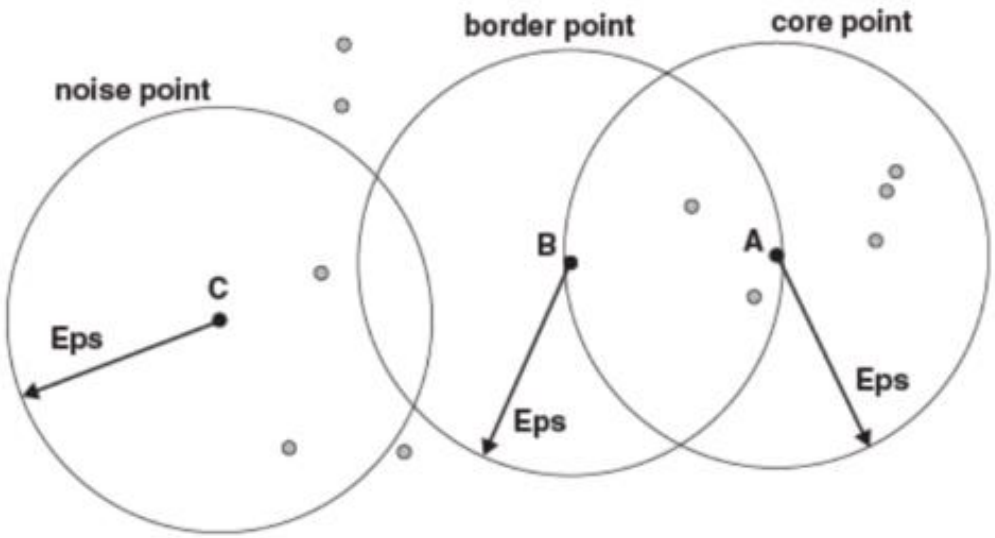


Example:

- The density of B is 4 and less than MinPts=6, so B is not a core point.
- But B falls within the neighbor of A (a core point).
- So, B is a border point.

DBSCAN

- Three types: core point, border point, and noise point.
  - A noise point is any point that is neither a core point nor a border point.



Example:

- The density of C is 3 which is less than MinPts=6, so C is not a core point.
- C doesn't fall within the neighborhood of any core point, so it is not a border point.
- So, C is a noise point.

DBSCAN

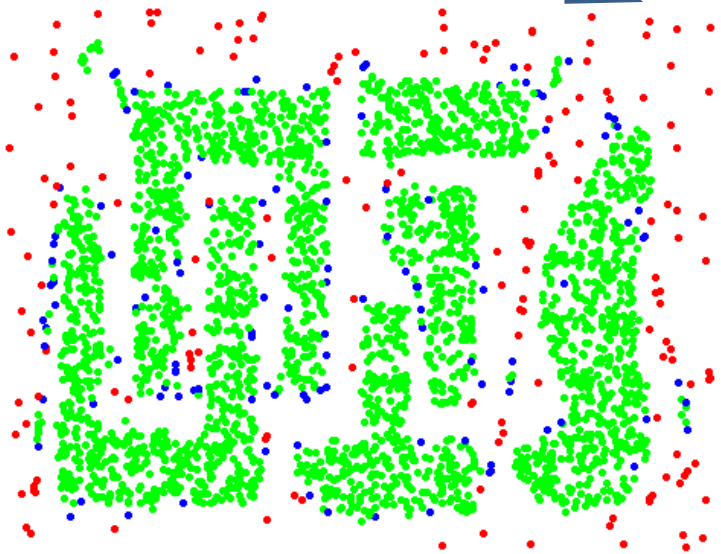
Steps of DBSCAN clustering

- Step 1: Label each point as either core, border, or noise point.
- Step 2: Mark each group of Eps connected core points as a separate cluster
- Step 3: Assign each border point to one of the clusters of its associate core points.

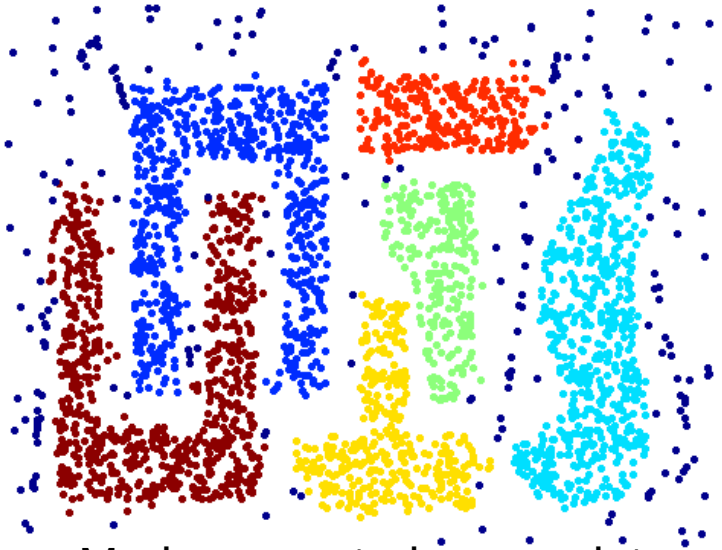
DBSCAN example

Original Points

Eps = 10, MinPts = 4



Mark core, border and noise points



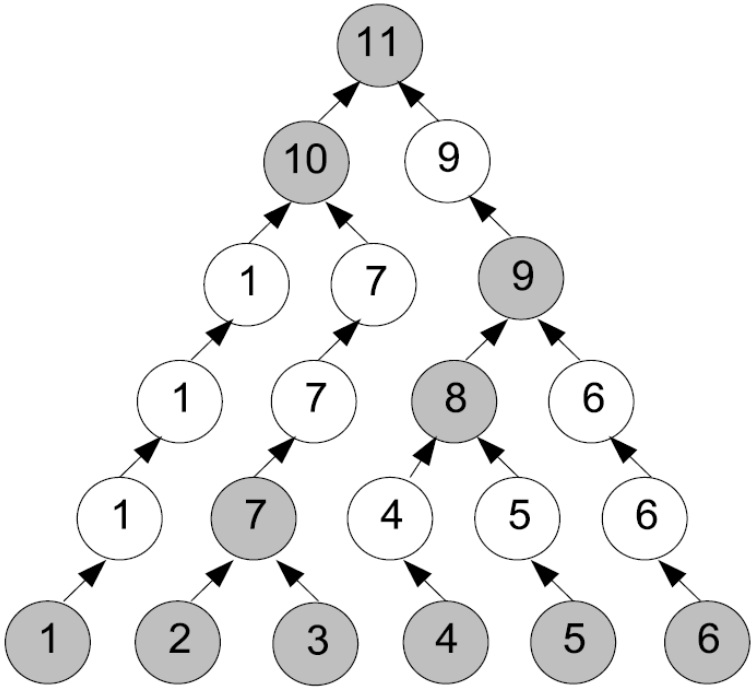
Mark connected core points

DBScan properties

- DBSCAN:
  - Resistant to noise and outliers
  - Can handle clusters of different shapes and sizes
  - Computational complexity is similar to K-means
- When DBSCAN does not work well
  - Varying densities
    - Can be overcome by using sampling
  - Sparse and high-dimensional data
    - Can be overcome by using topology preserving dimension reduction techniques.
- Using DBScan in R
  - `res <- dbscan(data, eps, MinPts = 5)`
  - http://www.sthda.com/english/wiki/wiki.php?id_contents=7940

Additional Algorithms

- Hierarchical Clustering (`hclust()`)
  - Hierarchical agglomerative clustering
  - Hierarchical divisive clustering



1. Each object is initially treated as a cluster
2. The clusters are then combined with the most similar cluster in each step
3. This process is repeated until one cluster (containing all objects) exists

Computationally very expensive $O(n^2)$ to $O(n^3)$ and thus rarely used in Big Data analytics.

Q&A

Images Courtesy of Google Image