Data Mining II

Ahmed Eleish

Data Science – ITWS/CSCI/ERTH-4350/6350

November 14th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

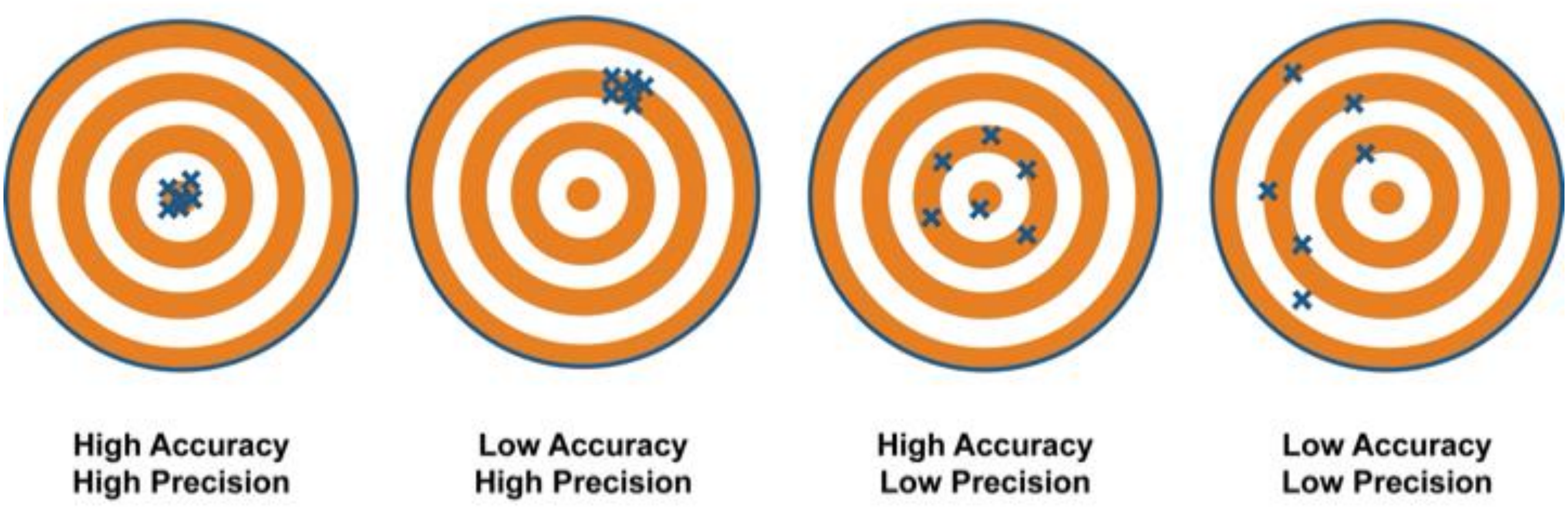
Contents

- Review of data mining concepts
- Linear regression
- k-nearest neighbors classification
- k-means clustering

Types of Data

Type of data	Level of measurement	Examples
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Results of some tests, e.g. positive/negative
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Accurate vs. Precise



http://climatica.org.uk/climate-science-information/uncertainty

Data Mining – What it is

- Extracting knowledge from large amounts of data
- Motivation
  - Our ability to collect data has expanded rapidly
  - It is impossible to analyze all of the data manually
  - Data contains valuable information that can aid in decision making
- Uses techniques from:
  - Pattern Recognition
  - Machine Learning
  - Statistics
- Data mining methods must be efficient and scalable (8~10 years ago, data mining could not be done on your Laptop).

Data Mining – Types of Mining

Supervised Learning

- Regression
  - Predict a continuous variable
- Classification
  - Predict a categorical variable (class label)
  - Labeled samples (ground truth) required

Unsupervised Learning

- Clustering
  - Detect structure in dataset
  - Divide samples into groups based on their similarity

Linear Regression

Regression

Linear Regression: In regression, fitting covariate and response data to a line is referred to as linear regression.
Covariate: A variable that is possibly predictive of the outcome under study control variable, ***explanatory variable, independent variable, predictor***
Response: dependent variable
Intercept: The expected value of the response variable when the value of the predictor variable is 0.
Slope: the average increase in Y associated with a one-unit increase in X

Reference/Resources:
The Elements of Statistical Learning. Hastie • Tibshirani • Friedman, 2nd Edition.
Introduction to Probability and Statistics, 4th Edition by Beaver.
Introduction to Statistical Learning with R, 7th Edition (ISLR).

Simple Linear Regression

- Let's take a look at the Least Squares Method for a single covariate (single regression).

- Utilizing the statistical notion of estimating parameters from data points, we find the estimates (coefficients) using the least squares method.

- We will look at evaluating linear models.

Least Squares Method

Equation of line: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

Let *n* be a positive integer. For a given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R} \times \mathbb{R}$,

- we obtain the intercept β_0 and slope β_1 using the least squares method.
- Residual Sum of Squares (RSS), the *i*th residual $e_i = y_i - \hat{y}_i$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Or

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

More precisely, we minimize RSS

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2$$

Sum of squared distances between (x_i, y_i) and $(x_i, \widehat{\beta_0}+\widehat{\beta_1}x_i)$ over i = 1,...,n

Assessing the Coefficient Estimates

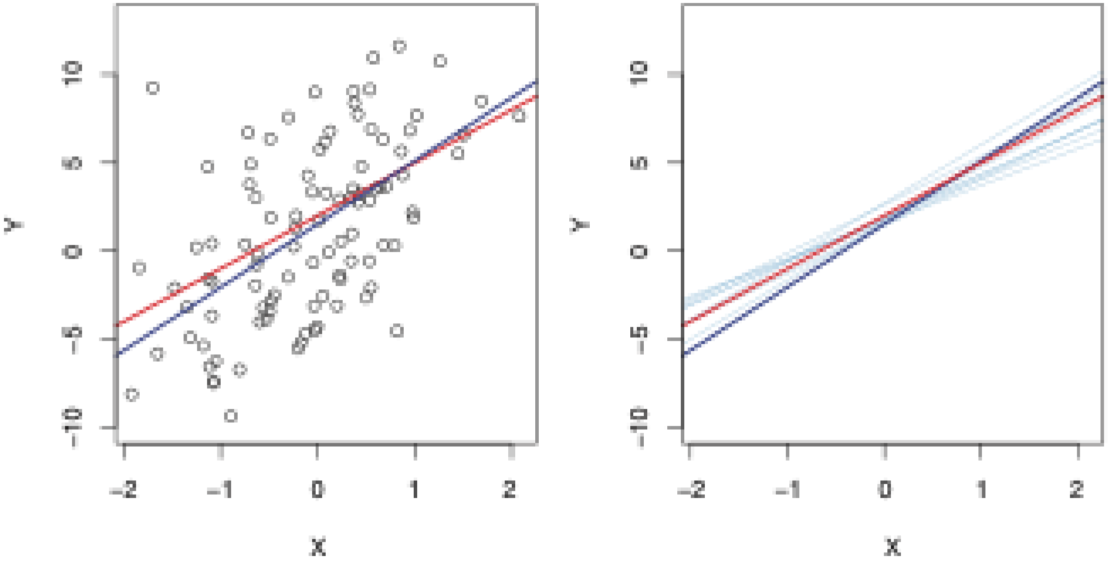
True relationship between X and Y: $Y = \beta_0 + \beta_1 X + \epsilon.$

- Where ϵ is a mean-zero random error

Red: true relationship

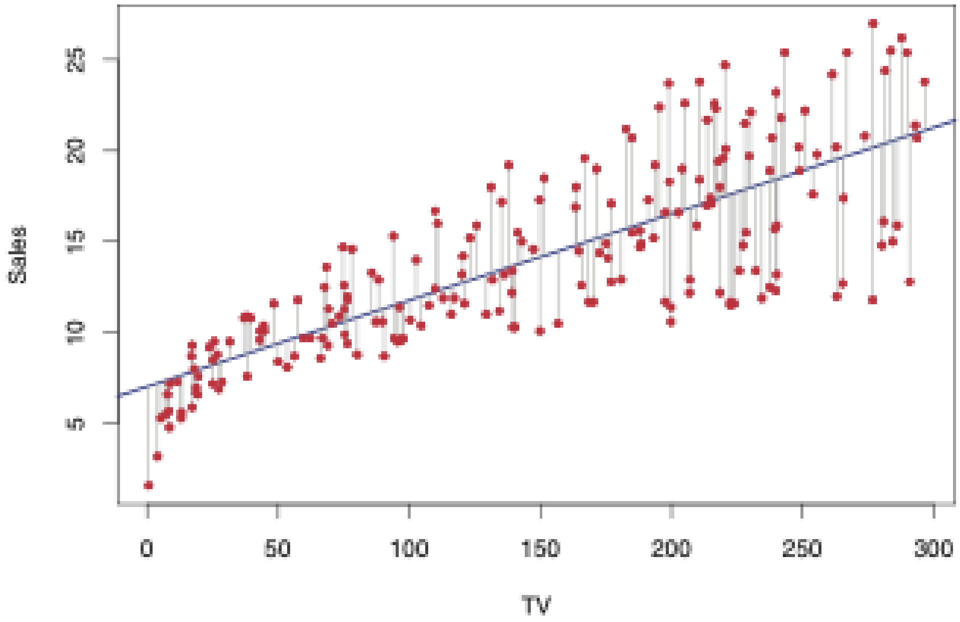
Dark Blue: least squares regression line

Light Blue: least squares regression lines for multiple random subsets



Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $



Evaluating Linear Models

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

Reject the null hypothesis!

Residual Standard Error

- Mean sales ≈ 14,000 units

RSE = 3.26 = 3,260 units
good/bad?

R^2

- measures the proportion of the variability in *Y* that can be explained using *X*
- has a value between 0,1

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

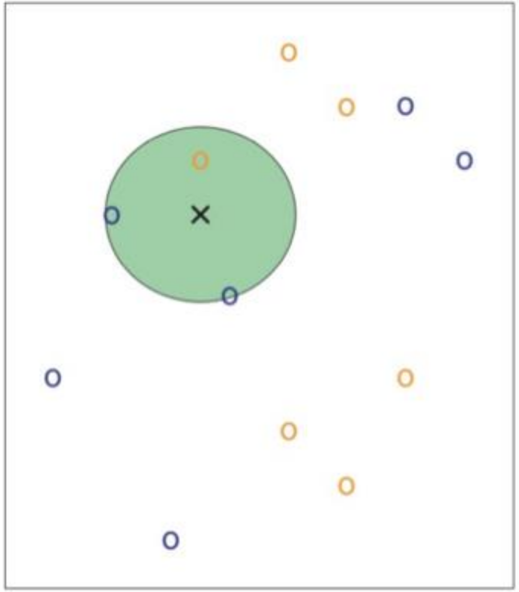
$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

k-Nearest Neighbors Classification

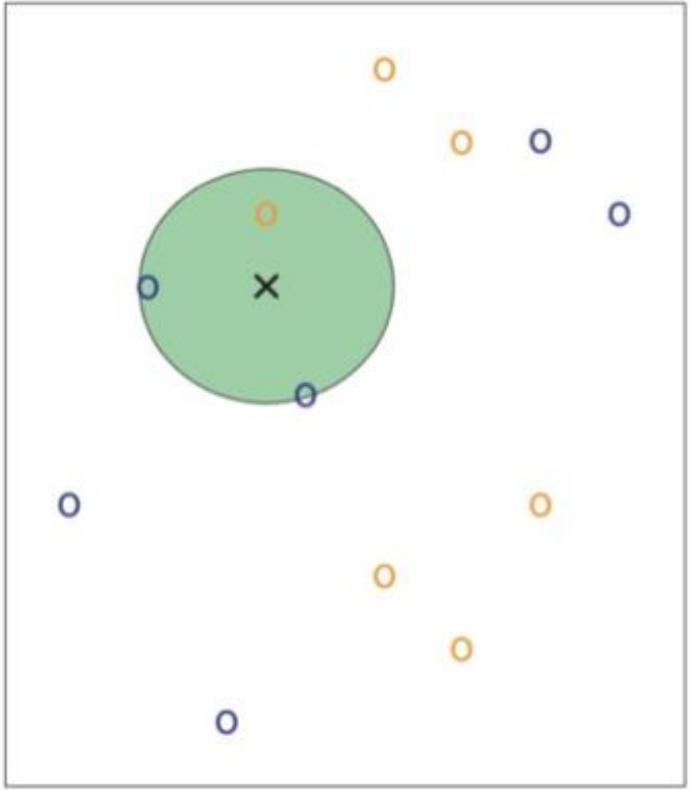
kNN Classifier

- In the figure a dataset is shown consisting 6 blue and 6 orange observations.
- Our goal is to make a prediction for the point labeled by the black cross.
- Suppose we choose K=3, then KNN will first identify the three observations that are closest to the black cross as shown in the figure.
- This neighborhood is shown as a circle. It consist of 2 blue points and 1 orange point, resulting in estimated probabilities of 2/3 for the blue class and 1/3 for the orange class.
- Hence, kNN will predict that the black cross belongs to the blue class.



1.Image/photo Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 2
Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 2 – KNN Classifier

6 blue points and 6 orange points



1.Image/photo Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 2
Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 2 – KNN Classifier

Classification Problem: iris flower

- Classifying Iris Species
- Let's assume a botanist is interested in distinguishing the species of some iris flowers that she has found. She has collected some measurements associated with each iris: length and width of the petals and length and width of sepals.
- She also has the measurements of some irises that have been previously identified by an expert botanist as belonging to the species
  - Setosa
  - Versicolor
  - Virginica
- **Problem: predict iris flower species from physical measurements**

Classification Accuracy

- ***Accuracy = (Number of correct predictions) / (Overall number of predictions)***

		Predicted Value	
		Positive	**Negative**
Real Value	**Positive**	TP	FP
	Negative	FN	TN

Evaluation Metrics

- ***Precision = (True Positive) / (True Positive + False Positive)***
- ***Recall = (True Positive) / (True Positive + False Negative)***
- ***F1 = 2 [(Recall * Precision) / (Recall + Precision)]***
    - ***F1 = (True Positive) / [True Positive + 1/2*(False Positive + False Negative)]***

k-Means Clustering

k-Means

- k-Means clustering is an unsupervised learning algorithm that, as the name hints, finds a fixed number (*k*) of clusters in a set of data.

- A *cluster* is a group of data points that are grouped together due to similarities in their features. When using a K-Means algorithm, a cluster is defined by a *centroid*, which is a point (either imaginary or real) at the center of a cluster.

- Every point in a data set is part of the cluster whose centroid is most closely located. To put it simply, K-Means finds *k* number of centroids, and then assigns all data points to the closest cluster, with the aim of keeping the centroids small

Resource: https://blog.easysol.net/machine-learning-algorithms-3/
https://blog.easysol.net/author/acorrea/

K-Means Algorithm

```
randomly chose k examples as initial centroids
while true:
    create k clusters by assigning each
        example to closest centroid
    compute k new centroids by averaging
        examples in each cluster
    if centroids don't change:
        break
```

Resource: MIT 6.0002 lecture 12 (MIT Open Courseware)
https://ocw.mit.edu/index.htm

Algorithm 10.1 *K-Means Clustering*

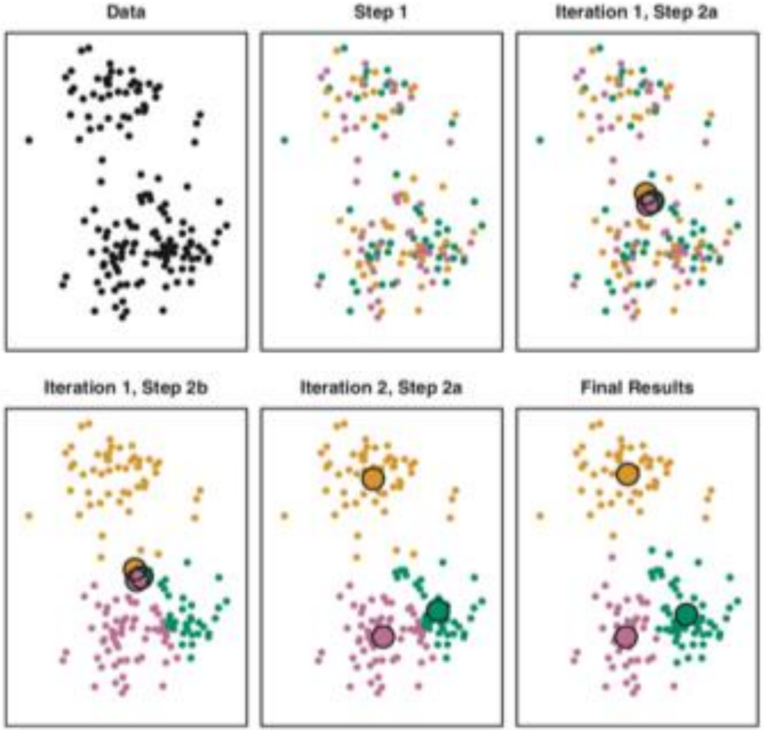
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

K-Means Algorithm

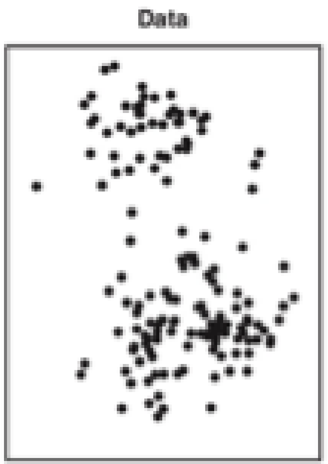
Algorithm 10.1 *K-Means Clustering*

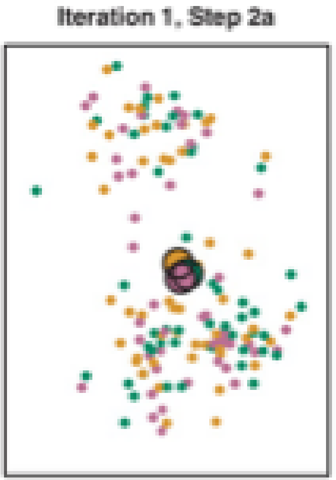
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

K-Means Algorithm





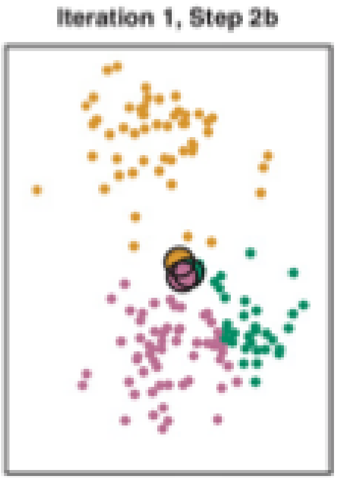
Observations (data) is shown

Step 1 of the algorithm: each observation is randomly assigned to a cluster

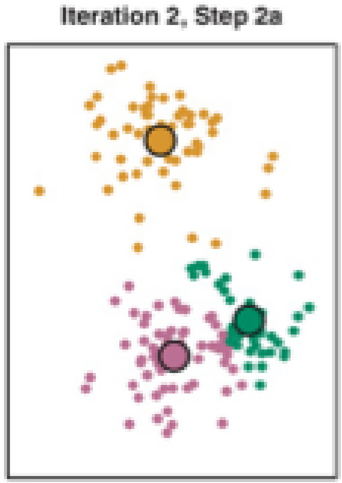
Iteration1 Step 2(a): The cluster centroids are computed; these are shown in large colored disks. Initially centroids are almost completely overlapping because the initial cluster assignment were chosen at random

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

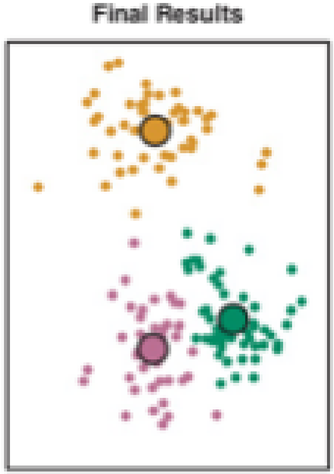
K-Means Algorithm



Iteration 1 Step 2(b) : each observation is assigned to the nearest centroid



Iteration 2, Step 2(a): the step 2(a) is once again performed, leading to new cluster centroids.



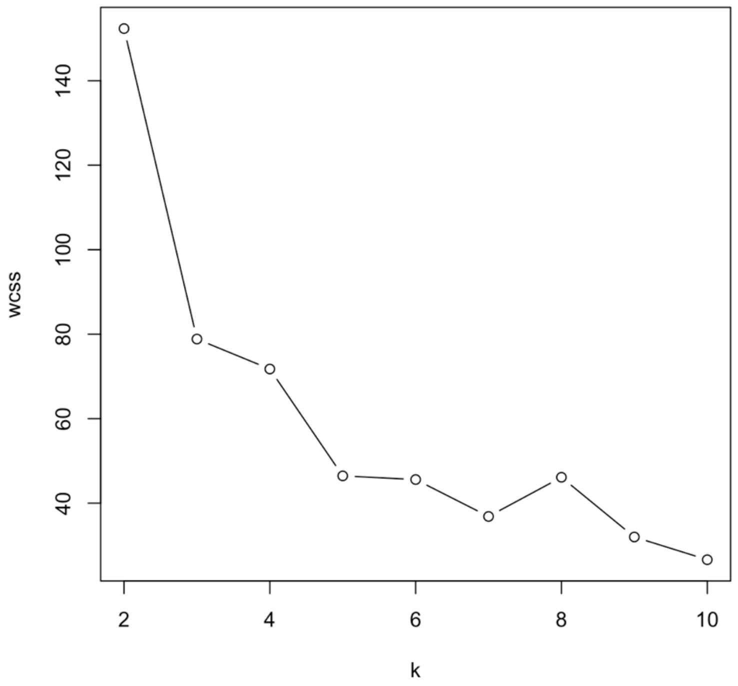
Final Results: the results obtained after ten iterations. You can see the distinct clusters with their centroids.

Image/Photo Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans
Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 10 – KMeans

- *k*-Means clustering Animation
- http://shabal.in/visuals/kmeans/6.html

Within-Cluster Sum of Squares (Elbow Method)

$$WCSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mathbf{c}_i||^2$$



Thanks!

Work with your teams!