Trendwise Analytics

Introduction to Machine learning

S.Mohan Kumar

Content

- What is Machine Learning
- Type of machine learning
  - Supervised
  - Unsupervised
  - Examples and applications
- Linear regression
  - Single variable - univariate
  - Understanding regression model
  - Classification
  - Logistic Regression
  - Decision Trees and Random Forest
  - Ensemble learning

Machine Learning

Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).

ML consists of:

-Programs
-Algorithms and statistical models
-Data and data mining

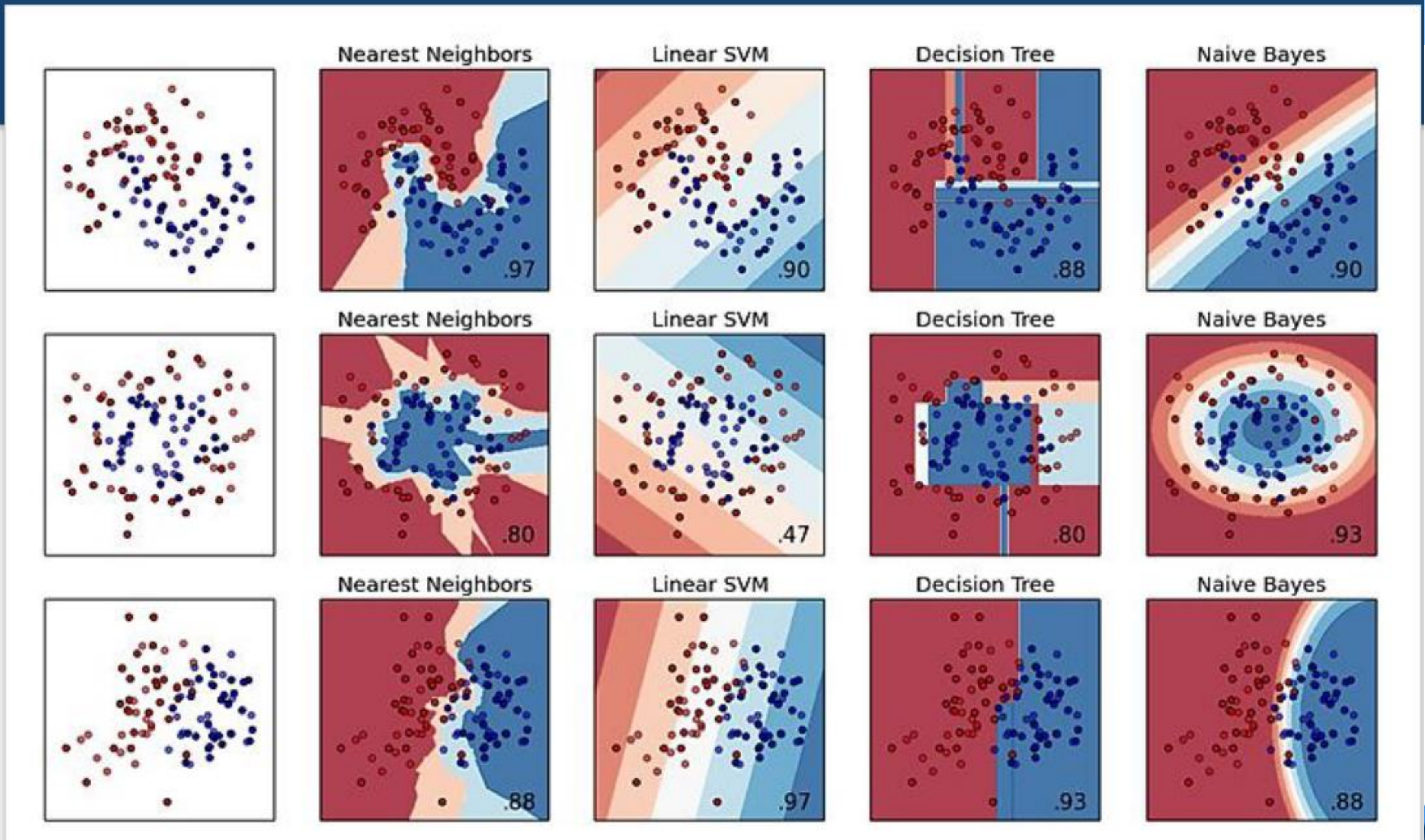
Machine learning

Supervised Learning: Learning with a labeled training set
Example: email spam detector with training set of already labeled emails

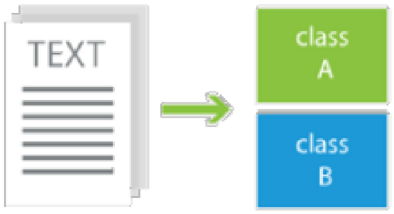
Unsupervised Learning: Discovering patterns in unlabeled data
Example: cluster similar documents based on the text content

Reinforcement Learning: learning based on feedback or reward
Example: learn to play chess by winning or losing

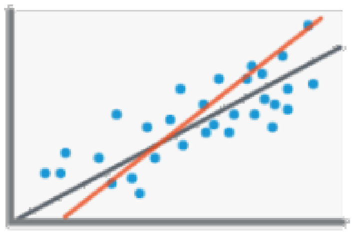
Algorithms


Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.97
.90
.88
.90
Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.80
.47
.80
.93
Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.88
.97
.93
.88

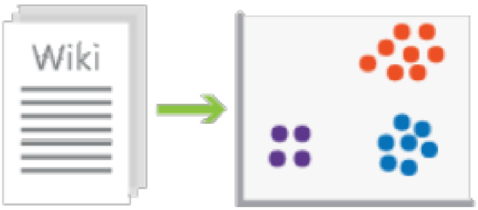
Data Mining Processes



Classification
(supervised – predictive)



Regression
(supervised – predictive)



Clustering
(unsupervised – descriptive)

Anomaly Detection
(unsupervised– descriptive)

Supervised learning

- Regression
-
- Classification
-

Machine learning

Supervised Learning: Learning with a labeled training set

Example: email spam detector with training set of already labeled emails

Supervised learning

- Regression – for continuous values
-
- Classification – for discrete values (classes)
-

Regression

- Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).

- In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

- One need to have the knowledge of dependent and independent variables

- Dependent variable: denoted **Y**
- Independent variables: denoted **$X_1, X_2, \ldots, X_k$**

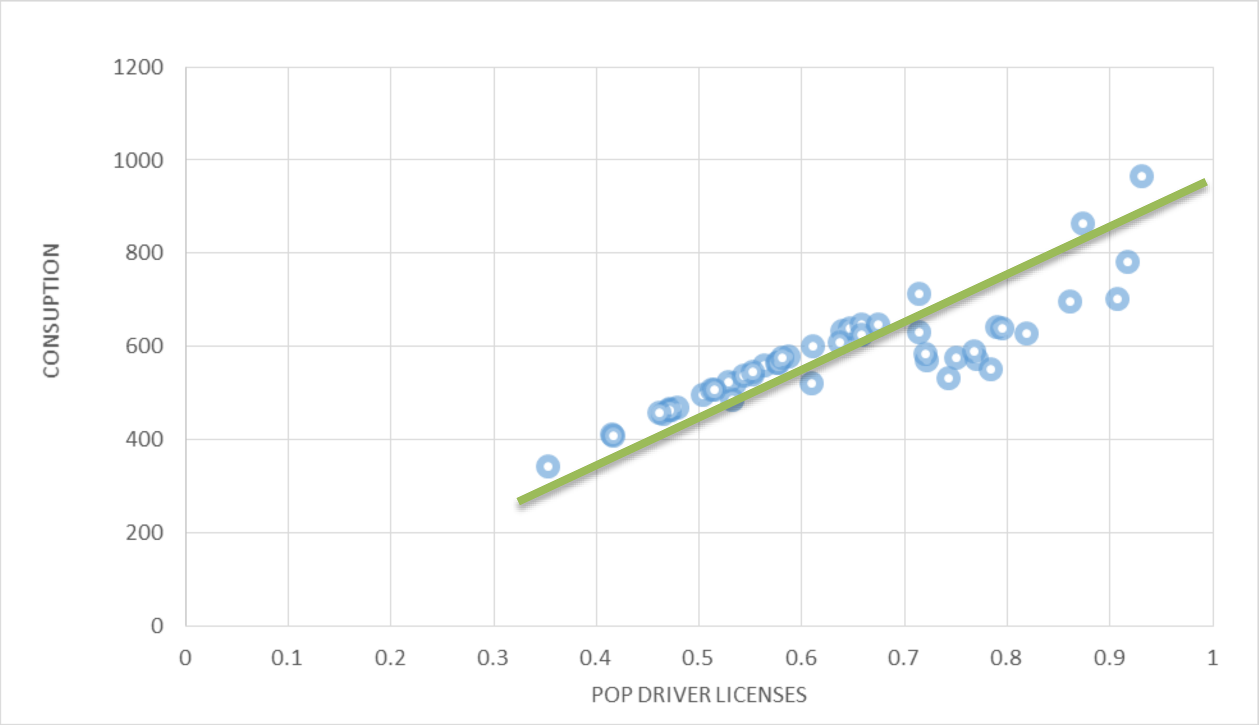
Simple Linear Regression Analysis

- If you know something about X, this knowledge helps you predict something about Y.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Above model is referred to as **simple linear regression**. We would be interested in estimating β_0 and β_1 from the data we collect.
- Variables:
  - X = Independent Variable (we provide this)
  - Y = Dependent Variable (we observe this)
- Parameters:
  - β_0 = Y-Intercept
  - β_1 = Slope
  - ε ~ Normal Random Variable [Noise]

Regression line fitting



- What is the best fit for our data?
  - The one which goes through the core of the data
  - The one which minimizes the error

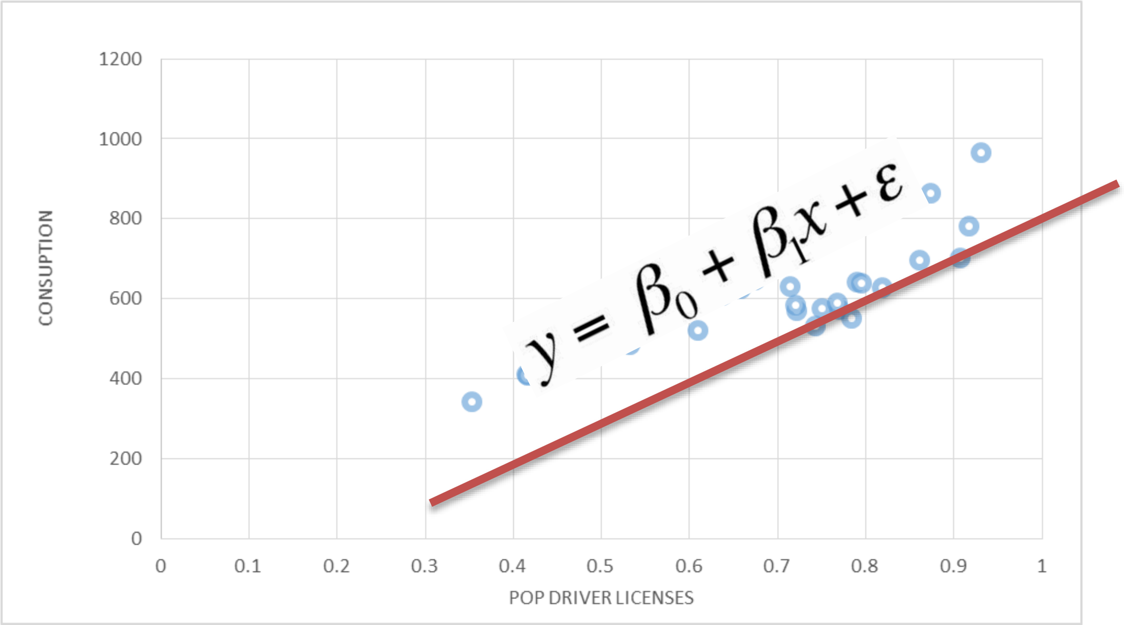
Least squares Estimation

- X: x1, x2, x3, x4, x5, x6, x7,……..
- Y:y1, y2, y3, y4, y5, y6, y7…….
- Imagine a line through all the points
- Deviation from each point (residual or error)
- Square of the deviation
- Minimizing sum of squares of deviation

$$\begin{aligned}\sum \mathbf{e}^2 &= \sum (\mathbf{y} - \hat{\mathbf{y}})^2 \\ &= \sum (\mathbf{y} - (\beta_0 + \beta_1 \mathbf{x}))^2\end{aligned}$$

β_0 and β_1 are obtained by minimize the sum of the squared residuals

Regression line



- The line goes through the core of the data since the parameters are obtained after minimizing the error function
- The above line the best fit for the data.
- We can go ahead and use it for prediction, substitute X and you will get Y

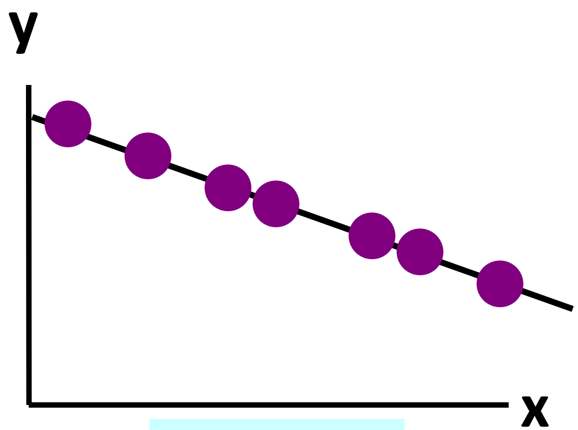
Coefficient of determination

- The coefficient of determination is also called R-squared and is denoted as R^2

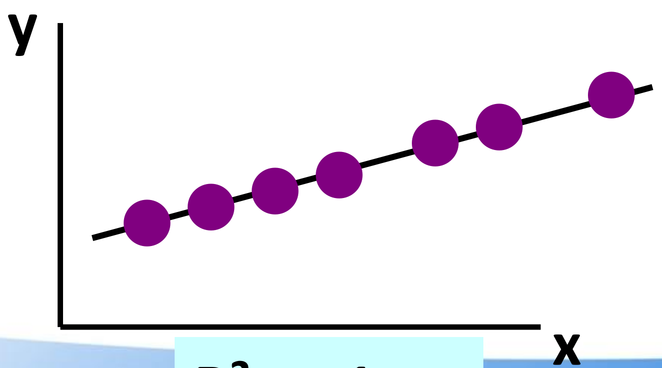
$$R^2 = \frac{SSR}{SST} \quad \text{where} \quad 0 \le R^2 \le 1$$

In the single independent variable case, the coefficient of determination is equal to square of simple correlation coefficient

Type of relationship



$R^2 = 1$



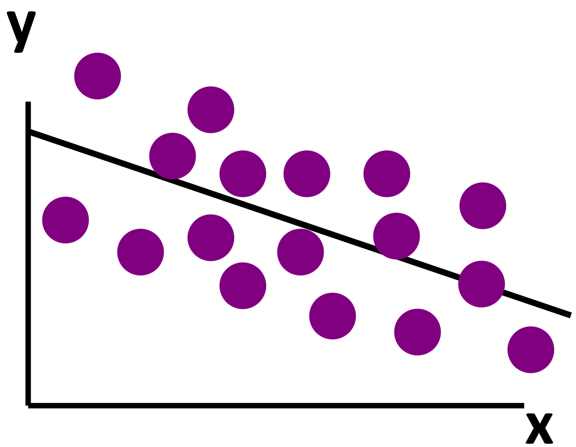
$R^2 = +1$

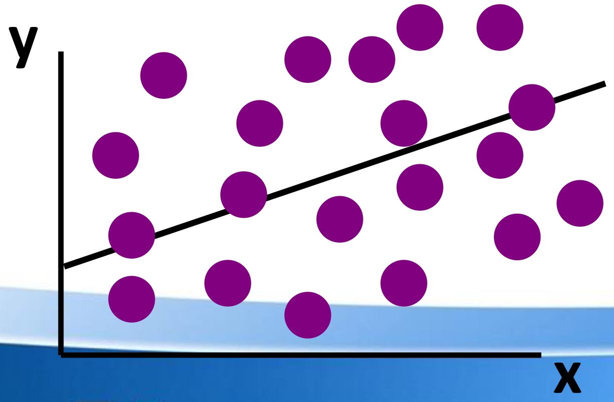
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

Type of relationship



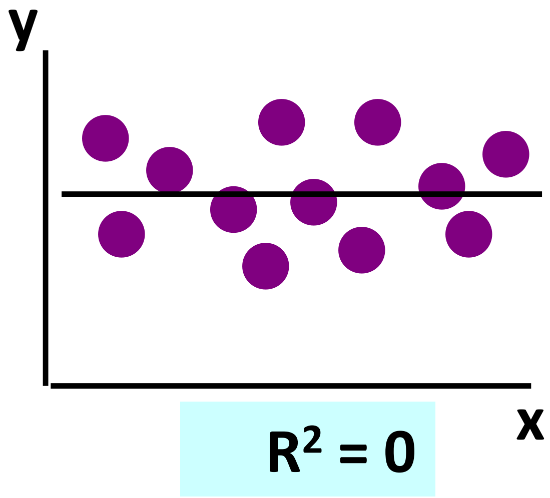


$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

Type of relationship



$R^2 = 0$

No linear relationship between x and y:

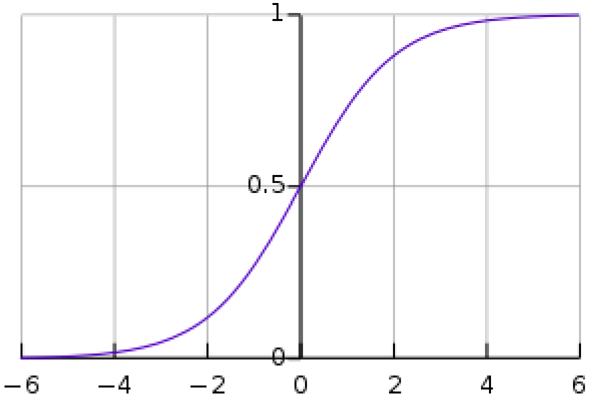
The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Classification examples

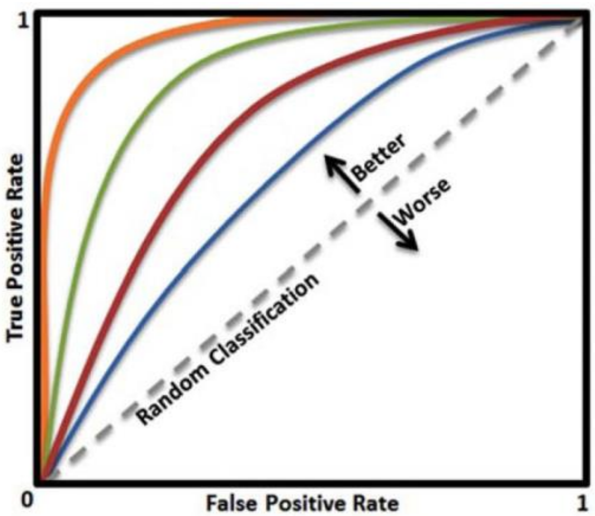
- Binary classification examples:
- Email – spam or not spam
- Gaming - Win vs Loss
- Sales - Buying vs Not buying
- Loans – Default vs Non Default
- Fraud identification –Fraud vs Non Fraud
-
- Multiclass Examples
-
- Image recognition – cat, dog, elephant
- Number recognition – 1,2,3....
- Voice recognition – speaker1, speaker2

Logistic regression

-
- Logistic regression is a technique used for
- classification.
- The results are discrete.
- The name comes from log
- Logistic function:
-
-



Logistic regression - accuracy



- Confusion matrix
- ROC curve
- Accuracy = sum of diagonal values
- ------------------------------
- Total observations

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Decision Trees

Unsupervised learning

Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm—which is one way of distinguishing unsupervised learning from supervised learning and reinforcement learning.

Wikipedia

Clustering

It is a type of unsupervised learning that:

• Forms clusters of similar objects automatically

• Segments the data so that each training example is assigned to a segment

Clustering – use cases

These include:

• Grouping the content of a website or product in a retail business

• Segmenting customers or users in different groups on the basis of their metadata and behavioral characteristics

• Segmenting communities in ecology

• Finding clusters of similar genes

• Creating image segments to be used in image analysis applications

Clustering – models

Some examples of clustering models are:

- K-means clustering
- Hierarchical Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Clustering

Clustering – k -Means model

K-means tries to:

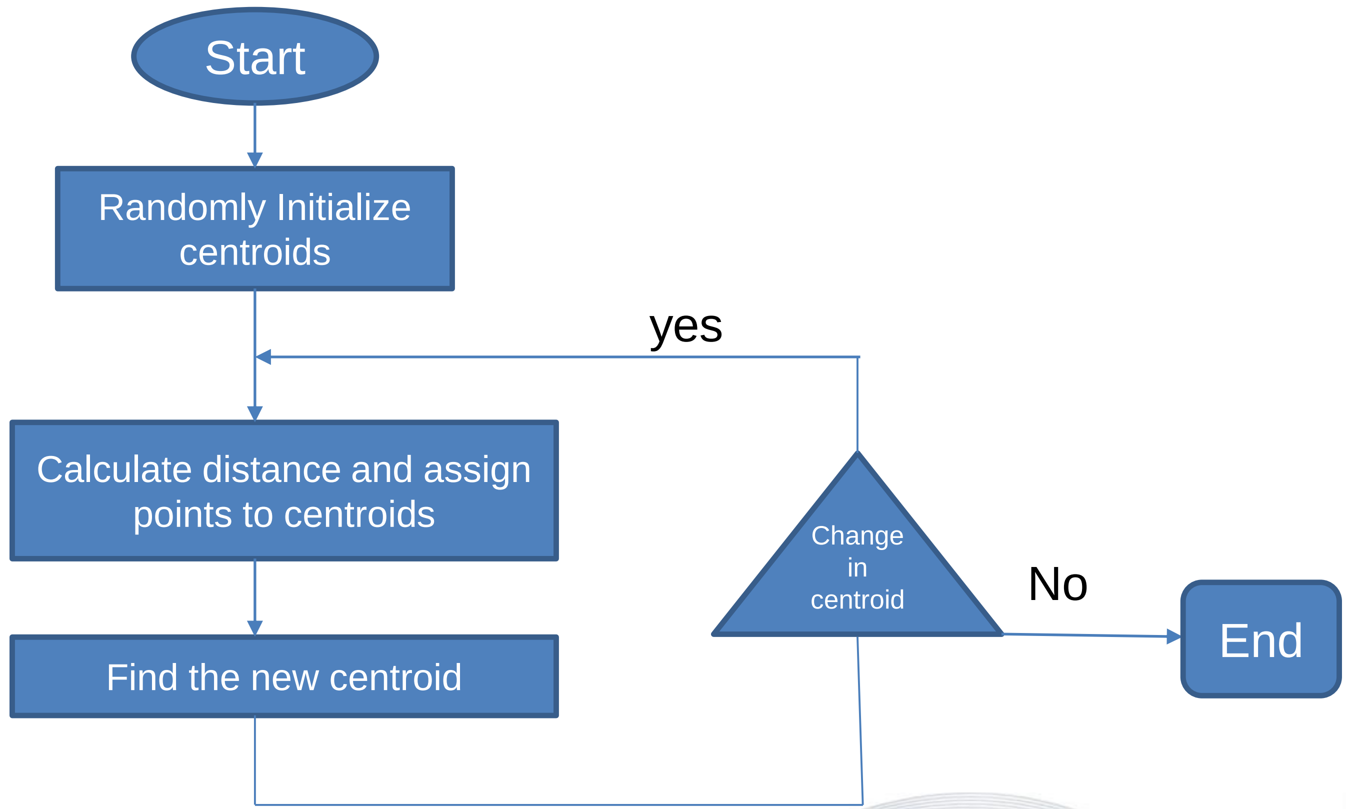
• Partition a set of data points into K distinct clusters
• Find clusters to minimize the sum of squared errors (WCSS) in every cluster

Clustering : k -Means algorithm

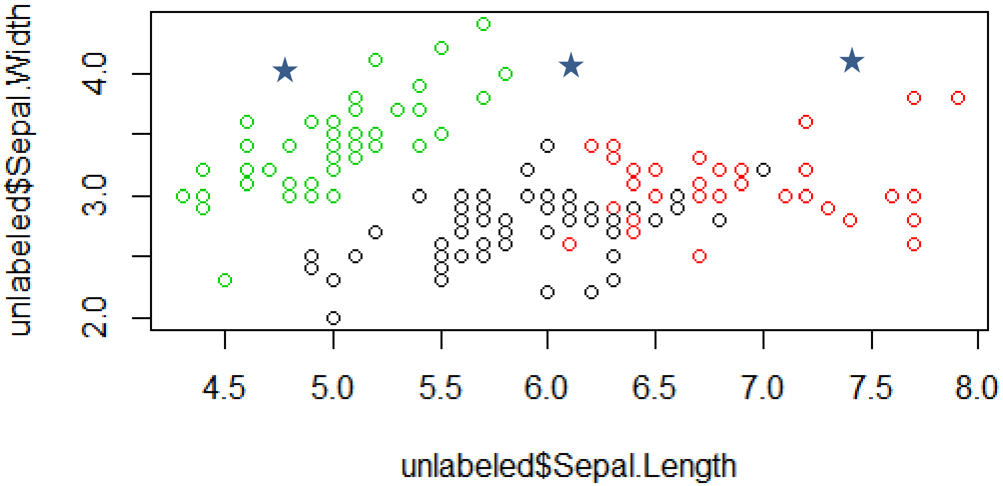
It includes the following steps:

- The k centroids are assigned to a point randomly
- Every point in the dataset is assigned to a cluster
- All centroids are updated by taking the mean of all the points in that cluster

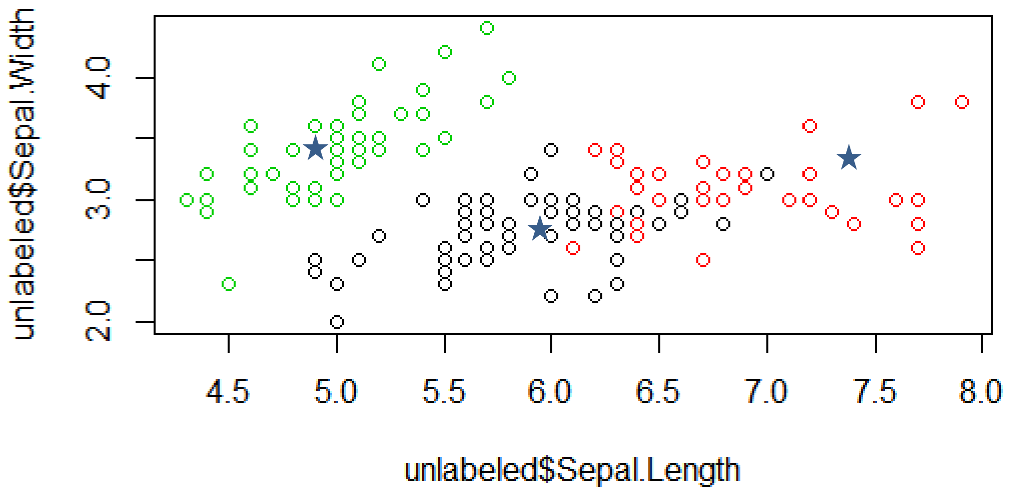
Clustering : k -Means algorithm



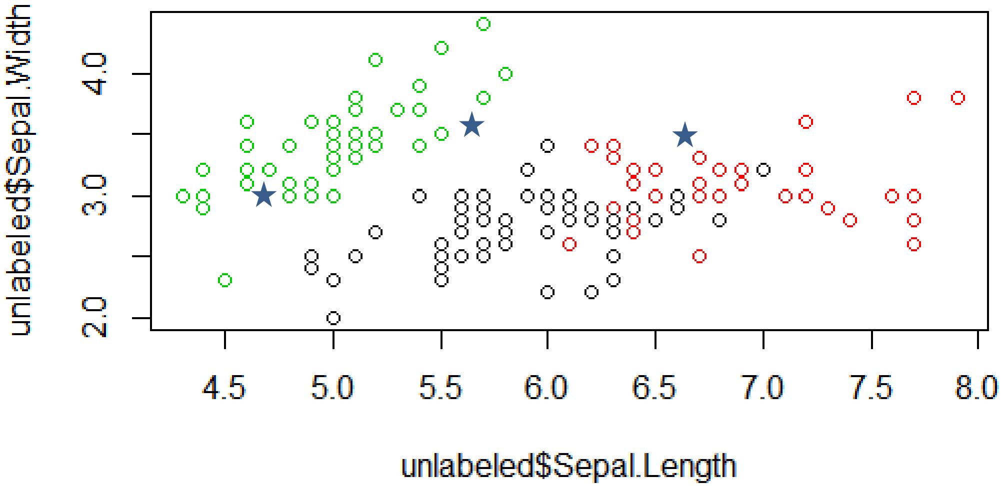
Clustering : k -Means algorithm



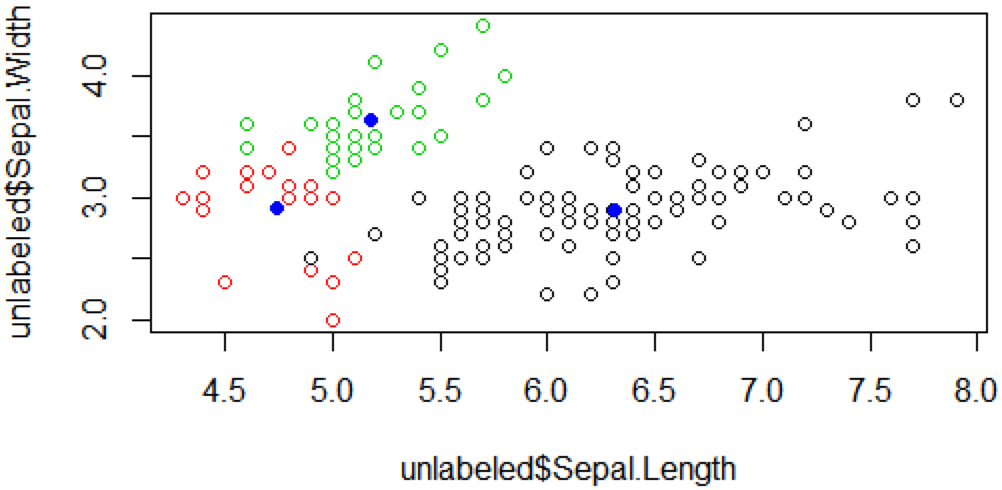
Clustering : k -Means algorithm



Clustering : k -Means algorithm

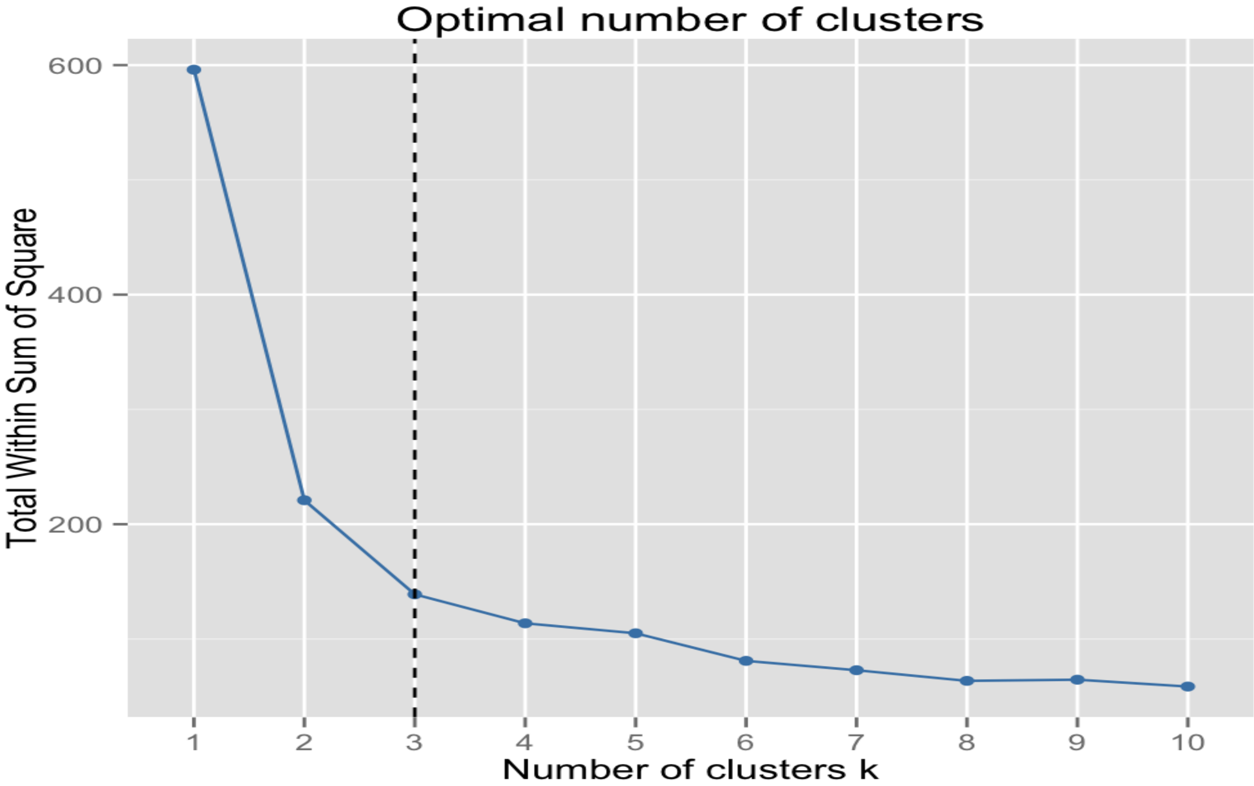


Clustering : k -Means algorithm



https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif

Clustering : choosing the value of k



Clustering : initialization