Unsupervised Learning

K-Means

What is Unsupervised Learning?

- Understand the underlying data of structure.
- How to summarize it without going into much details
- Eg 1. Group customers by purchasing patterns
- Eg 2. Compress the data (Dimension Reduction)

Supervised vs Unsupervised

Supervised Learning

- Finds pattern for a prediction task
- Eg. classify dog/cat from photo
- Have features and labels

Unsupervised Learning

- Just finds pattern for data
- Without prediction in mind
- Eg. Customer segmentation /
- No Labels only data(features)

Supervised vs Unsupervised

Supervised Learning
Unsupervised Learning
Positive ID:
George Clooney
Negative ID
Assessing
Assessing..
Positive ID:
George Clooney
DAMIANI
tiff.
Kontagent

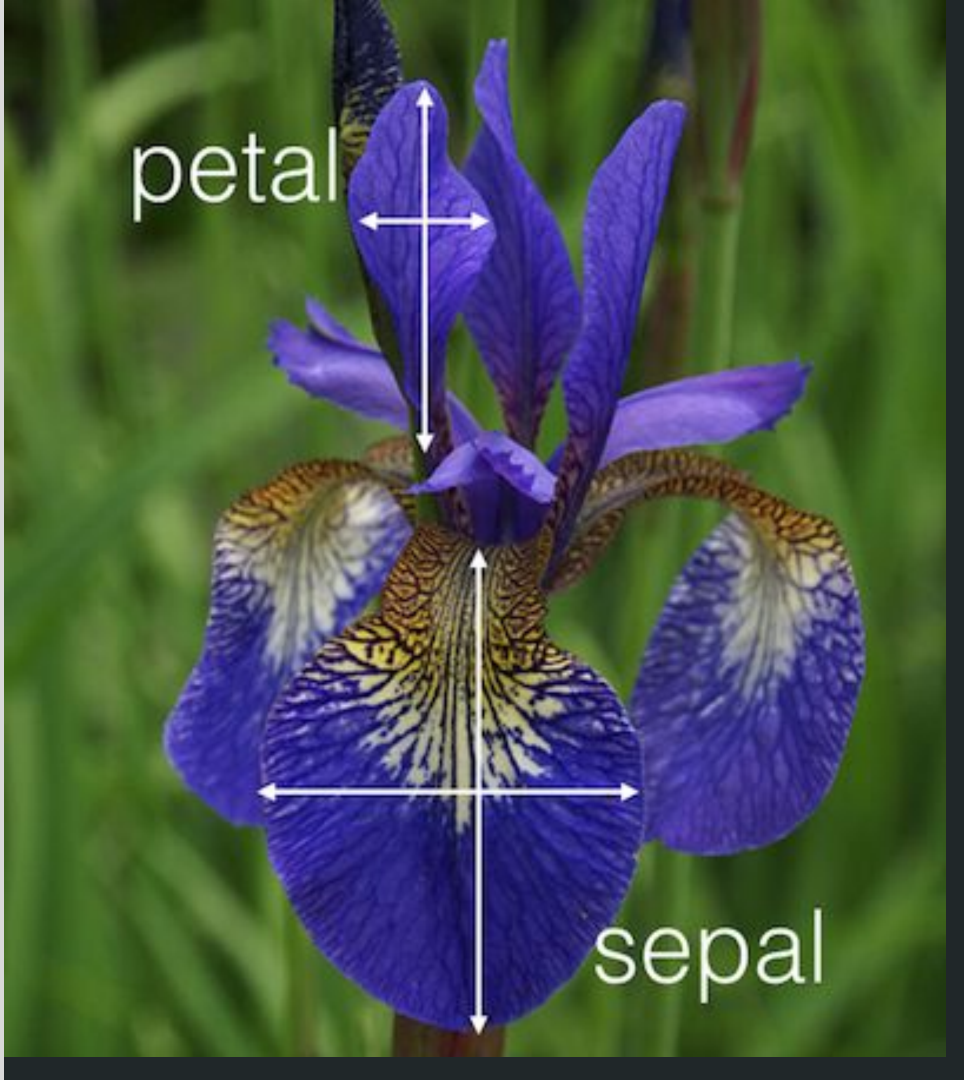
Lights Camera → Action

Iris dataset :

- Measurements of Iris Plants
- 3 species in dataset
- 4 features → sepal length, sepal width, petal length, petal width

Iris flower

Properties



So what are we gonna do

- Load the dataset into numpy arrays
- See the data
- Use some magic to identify the pattern in data
- Does it find the pattern?
- Learn the magic trick

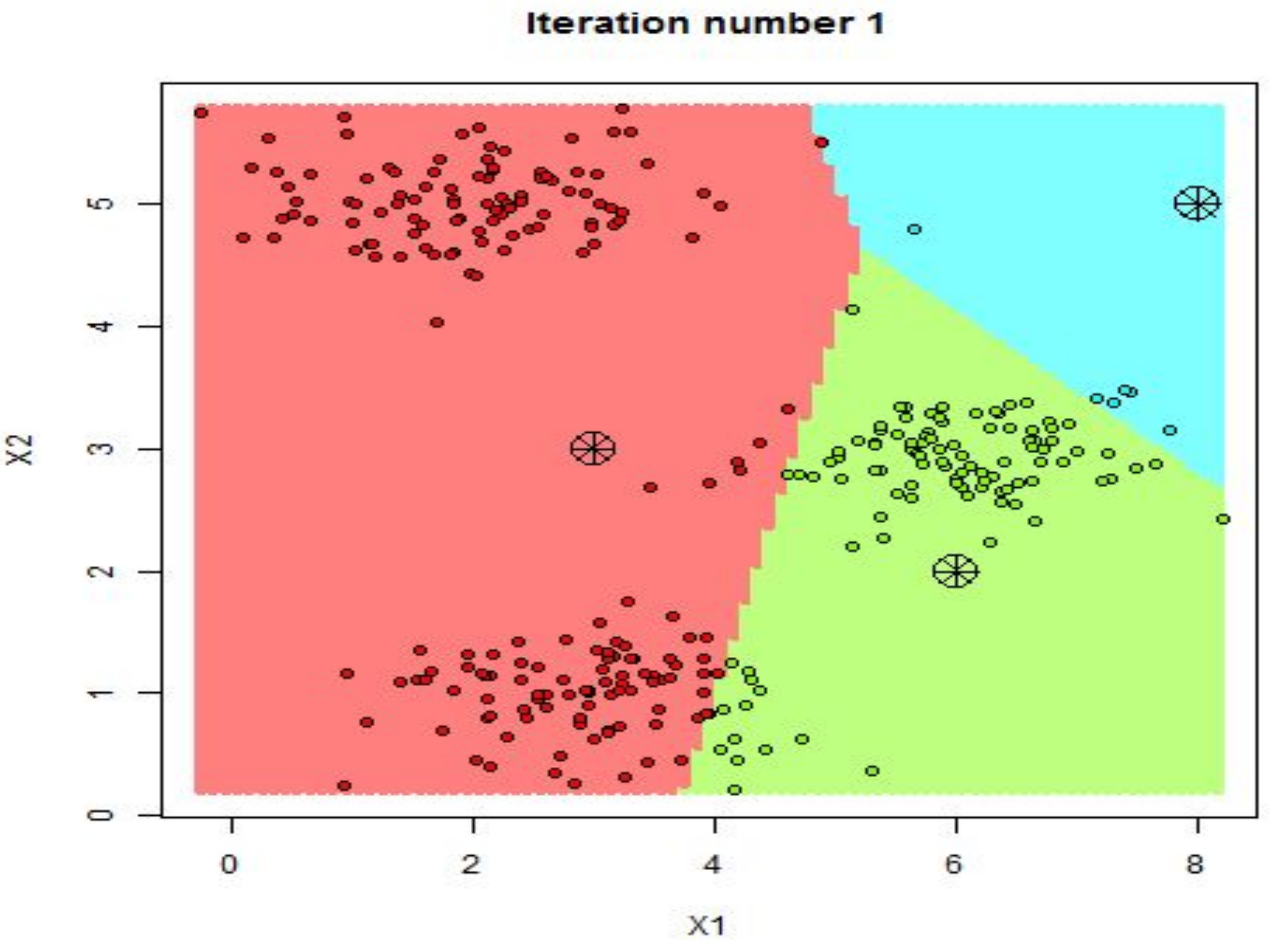
To Code and Beyond

MAGIC CODE : What was that KMeans?

- It’s a clustering algorithm
- Clustering ? Grouping data based on some metric
- Most commonly used
- Very Easy to understand

K-Means Algorithm

1. Choose n points randomly as cluster centroids.
2. Data Assignment :
    a. Calculate the distance between each point and cluster centres
    b. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
3. Centre Update :
    a. Take mean of all points and assign it as new cluster centre
4. Goto Step 2 until no new assignment is made


Iteration number 1
X2
0
1
2
3
4
5
0
2
4
6
8
X1

Issues with k-means

- Number of clusters should be known before hand
- A random initilization of cluster can fail to do the cluster correctly
- Outliers and noise can change results
- Fail for non-linear data

How to Evaluate a clustering ?

When we know about the data

- Check correspondence with the labels
  - Eg : we know there are 3 iris species so how many were correctly clustered
  - Create a cross tabulation with original cluster and algo cluster

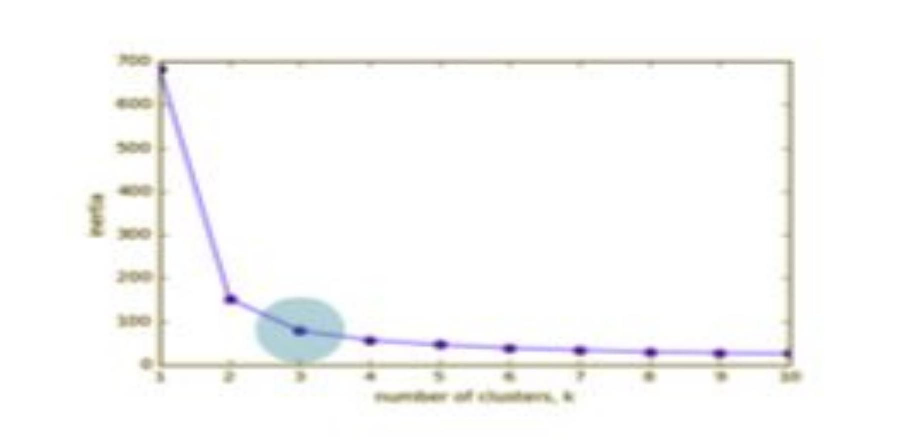
How to Evaluate a clustering ?

When we don’t know about the data

- Cluster should be spread
- The points inside the cluster should be tightly closed
- Inertia : Distance from each sample to centroid of cluster (should be low)
- More clusters means low inertia

How low we can go ? (or how to choose # of clusters)

- Cluster with different number of values
- Create an inertia plot for each of them
- Choose the elbow point from the plot
  - Choose that point after which inertia decreases slowly
- Inertia Plot



To Code and Beyond

Where else K-Means fails?

- Data is not in proper format
  - Some features have more variance in values and some less
- What can we do about it?
  - Transform the data to standard values of mean 0 and variance 1
  - This always help

Data transformation

- Data having mean = 0 and variance = 1 is called as standardized

- How do we do that ?
  - Write own code as (x-mean) / standard deviation
  - Or use inbuild library **StandardScaler()\**

- Now fit the clusters again and get the clustering

To Code and Beyond

Assignment

- Cluster stocks using k-means
- Find which stocks move together