K-Means Clustering on Seeds Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

Linear regression is useful model where the target function is real valued. Here we have divided the data into three classes and we have tried to classify it with K means algotithm.

We drew some clusters and tried to see if the classification is good enough or not. We have clustered the data into three types of seeds correctly.

1. Introduction

Unsuperwised learning is the learning from the data where the target value is not given. K-Means is the algorithm in which the data is divided into K clusters. First the K centreroids are defined randomly. In case of K Means++ it centeroids are taken as far as possible. Then each points are assigned to the nearest centeroids. And again the centeroid is calculated of those clustered data. Then again several times the procedure is repeated until no points are altered.

This method is very useful for the distinctive data which can be visualized as groups.

2. Methodology

We can define similarity as the opposite of distance, and a commonly used distance for clustering samples with continuous features is the squared Euclidean distance between two points x and y in m-dimensional space:

$$d(x,y)^2=\Sigma(x_j-y_j)^2 = ||x-y||^2_2$$

Note that, in the preceding equation, the index j refers to the jth dimension (feature column) of the sample points x and y. In the rest of this section, we will use the superscripts i and j to refer to the sample index and cluster index, respectively. Based on this Euclidean distance metric, we can describe the k-means algorithm as a simple optimization problem, an iterative approach for minimizing the withincluster sum of squared errors (SSE), which is sometimes also called cluster inertia:

$$SSE=\Sigma\Sigma w^{(i,j)}|| x^i - \mu^j ||^2_2$$

Here, μ(j) is the representative point (centroid) for cluster j, and $w^{(ij)}$, =1 if the sample $x^{(i)}$ is in cluster j; $w^{(ij)}$, = 0 otherwise.

3. Data

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected forthe experiment.

Attribute Information:

To construct the data, seven geometric parameters of wheat kernels were measured:
1. area A,
2. perimeter P,
3. compactness C = 4*pi*A/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

4. Algorithm

k-means algorithm that can be summarized by the following four steps:

1. Randomly pick k centroids from the sample points as initial cluster centers.

2. Assign each sample to the nearest centroid μ( ) j , j k ∈{ } 1,…, .

3. Move the centroids to the center of the samples that were assigned to it.

4. Repeat the steps 2 and 3 until the cluster assignment do not change or a user-defined tolerance or a maximum number of iterations is reached.

The initialization in k-means++ can be summarized as follows:

1. Initialize an empty set M to store the k centroids being selected.

2. Randomly choose the first centroid $\mu^{(j)}$ from the input samples and assign it to M .

3. For each sample $x^{(i)}$ that is not in M , find the minimum squared distance $d^{(x(i),M)2}$ to any of the centroids in M .

4. To randomly select the next centroid $\mu^{(p)}$, use a weighted probability distribution equal to

$$d(\mu^{(p)},M)^2/\Sigma d(x^{(i)},M)^2$$

5. Repeat steps 2 and 3 until k centroids are chosen.

6. Proceed with the classic k-means algorithm.

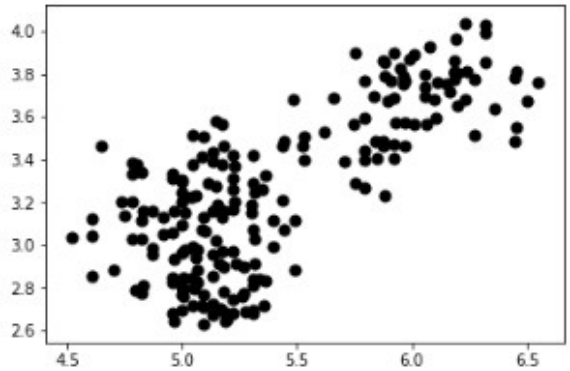
5. Experiments

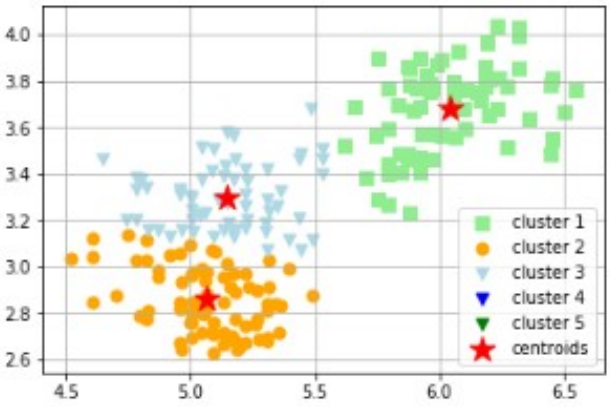
First we have took the seed data and plotted the pair plot.

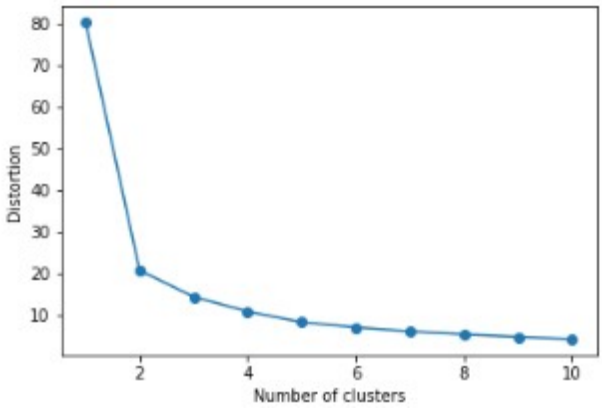
Then I took the graph between length of kernel groove and width of kernel and plotted the scatter plot. After that I have applied the K means algorithm to divide the data points between three clusters. And plotted the graph. Then we analyzed the graph between the number of clusters and distortion of points. Also found the Silhouette coefficient of each cluster.

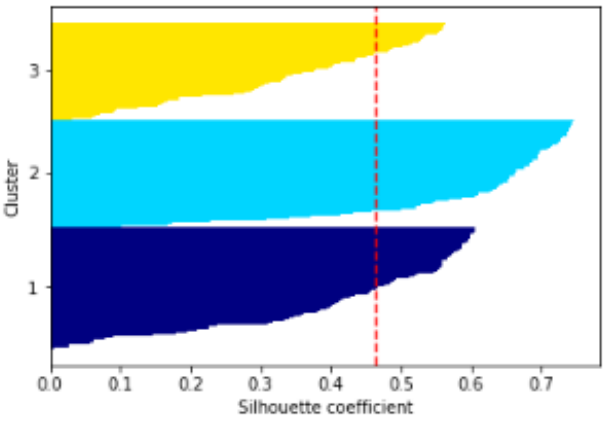
Similarly we have done for the length of kernel vs length of kernel groove. And for asyymetic coefficient vs length of kernel groove.

1) For length of kernel groove and width of kernel

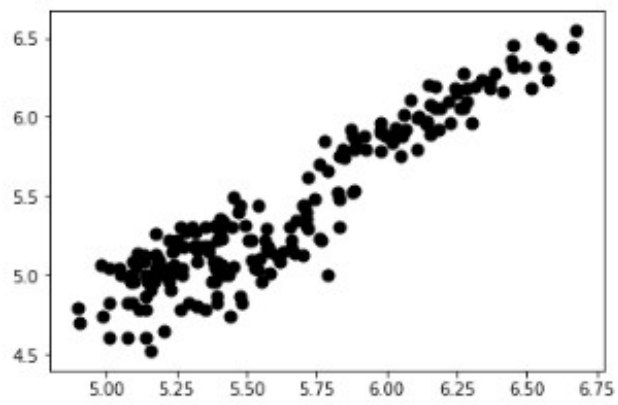


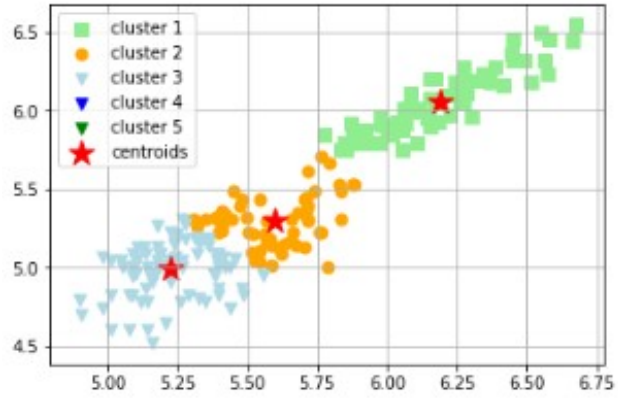






2) length of kernel vs length of kernel groove





6. Conclusion

From the results for the clustering of graph of length of kernel groove and width of kernel we can see the three types of seeds. Also we can see that after 10 clusters distortion of data becomes very less.

In another graph also we can see two of the classes clearly before clustering but after allying K means we can see that two of three clusters are merged with each other so that we cannot distinguish them clearly. Same goes for the first graph also.

7. Referances

1) Python Machine Learning by Sebastian Raschka Chapter 10