Lab-11: Unsupervised machine learning: K-mean clustering

11.1 Objectives:

1. To learn and Implement K-mean clustering machine learning technique

11.2 Unsupervised machine learning

In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate the goodness of a potential solution. This distinguishes unsupervised from supervised learning. Unsupervised learning is defined as the task performed by algorithms that learn from a training set of unlabeled or unannotated examples, using the features of the inputs to categorize

them according to some geometric or statistical criteria. Unsupervised learning encompasses many techniques that seek to summarize and explain key features or structures of the data. Many methods employed in unsupervised learning are based on data mining methods used to preprocess data. Most unsupervised learning techniques can be summarized as those that tackle the following four groups of problems:

• **Clustering:** has as a goal to partition the set of examples into groups.

• **Dimensionality reduction:** aims to reduce the dimensionality of the data. Here, we encounter techniques such as Principal Component Analysis (PCA), independent component analysis, and nonnegative matrix factorization.

• **Outlier detection:** has as a purpose to find unusual events (e.g., a malfunction), that distinguish part of the data from the rest according to certain criteria.

• **Novelty detection:** deals with cases when changes occur in the data (e.g., in streaming data). The most common unsupervised task is clustering, which we focus on in this Lab.

11.2.1 Clustering

Clustering is a process of grouping similar objects together; i.e., to partition unlabeled examples into disjoint subsets of clusters, such that:

• Examples within a cluster are similar (in this case, we speak of high intraclass similarity).

• Examples in different clusters are different (in this case, we speak of low interclass similarity). When we denote data as similar and dissimilar, we should define a measure for this similarity/dissimilarity. Note that grouping similar data together can help in discovering new categories in an unsupervised manner, even when no sample category labels are provided. Moreover, two kinds of inputs can be used for grouping:

(a) in similarity-based clustering, the input to the algorithm is an n × n dissimilarity matrix or distance matrix;

(b) in feature-based clustering, the input to the algorithm is an n × D feature matrix or design matrix, where n is the number of examples in the dataset and D the dimensionality of each sample.

11.2.2 K-means

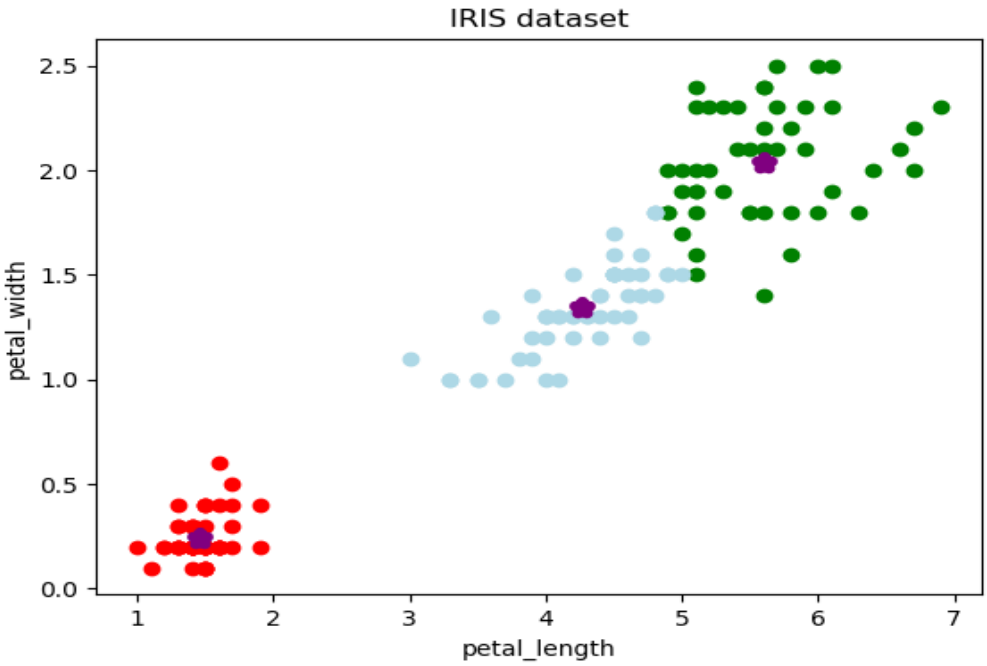
The K-means algorithm is a clustering method that is popular because of its speed and scalability. K-means is an iterative process of moving the centers of the clusters, called the **centroids**, to the mean position of their constituent instances and re-assigning instances to the clusters with the closest centroids. The titular k is a hyperparameter that specifies the number of clusters that should be created; K-means automatically assigns observations to clusters but cannot determine the appropriate number of clusters. k must be a positive integer that is less than the number of instances in the training set. Sometimes the number of clusters is specified by the clustering problem's context. For example, a company that manufactures shoes might know that it is able to support manufacturing three new models.

To understand what groups of customers to target with each model, it surveys customers and creates three clusters from the results, that is, the number of clusters specified by the problem's context. Other problems may not require a specific number of clusters, and the optimal number of clusters may be ambiguous.

The parameters of K-means are the positions of the clusters' centroids and the observations that are assigned to each cluster. Like generalized linear models and decision trees, the optimal values of K-means' parameters are found by minimizing a cost function. The cost function for K-means is given by the following equation:

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - \mu_k||^2$$

Here, μ_k is the centroid for cluster k. The cost function sums the distortions of the clusters. Each cluster's distortion is equal to the sum of the squared distances between its centroid and its constituent instances. The distortion is small for compact clusters and large for clusters that contain scattered instances. The parameters that minimize the cost function are learned through an iterative process of assigning observations to clusters and then moving the clusters. First, the clusters' centroids are initialized, often by randomly selecting instances. During each iteration, K-means assigns observations to the cluster that they are closest to and then moves the centroids to their assigned observations' mean location.



TASK: Use sepal_length and sepal_width as features from the iris_dataset and apply K-mean clustering. Attach the code and graph.