Machine Learning Engineer Course
Day 14

- Clustering (k-means) -

DIVE INTO CODE

Thursday June 10, 2021

DIOP Mouhamed

Agenda

1. Check-in
2. How to proceed
3. Quick Review
4. Clustering (K-means)
5. K-means Class of Scikit-learn
6. Assignment
7. Scratch Clustering– Sample Code
8. Check-out

Check-in

3 minutes Please post the following point to Zoom chat.

Q. What did you learn in the previous week?
(Anything is fine.)

How to proceed - Objective

What is the purpose?

1. Know the clustering algorithm and understand k-means through scratching

2. Use Principal Component Analysis, an algorithm introduced with knowledge of linear algebra

3. Use cluster analysis

Let’s learn the basics of Clustering

Quick Review (Scratch Decision Tree)

Knowing the Decision Tree Problem Setting

① Use an arbitrary value of the feature value as the "threshold" for segmentation (done for each feature value)

② Divide the sample by the threshold value in ① above (done for each feature)

③ Find the difference between the sum of the gini impurity of the samples in each group after the split and the gini impurity of all samples before the split (calculate the information gain).

④ The one that maximizes the difference (information gain) is used as the decision criterion for splitting the root node.

Target audience for this assignment

① Able to write code to train and estimate using scikit-learn clustering model.

② Experience with supervised learning algorithms.

Supervised and unsupervised learning

What is an unsupervised learning algorithm?

Unsupervised learning algorithms are a convenient name used to distinguish machine learning algorithms from supervised learning algorithms based on the differences in the actual values acquired from a dataset during the learning process. It is a name used for convenience to distinguish machine learning algorithms from supervised learning algorithms [1].

[1] See Ian Goodfellow et al. 'Deep Learning (Adaptive Computation and Machine Learning series)' 2016. Chapter 5.8. There is no formal and rigorous definition, as there is no meta-test to determine whether a measurement is an explanatory variable or an objective variable (note; the book does not give specific cases where it cannot be determined, but it could be taken to mean that there is no standard to strictly distinguish between "supervised data" (correct labels) and those that are not). Roughly speaking, unsupervised learning can be explained as an attempt to extract information from a distribution that does not need to be manually annotated.

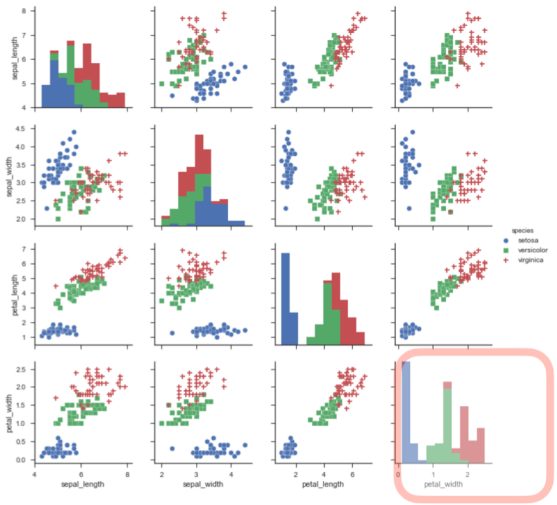
Supervised and unsupervised learning

Supervised learning vs. unsupervised learning

Supervised learning can be thought of as an algorithm that learns to make predictions based on a probability vector (obtained from a feature vector) and a set of examples of related vectors (elements of which are, for example, 0 or 1), by learning to estimate a conditional probability distribution [2].

To explain it in another way, for example, in the case of the iris dataset, supervised learning means that given a set of objective variables representing the species to which the iris belongs, and a set of explanatory variables (features) consisting of measurements of each part of the iris associated with each of these variables, the algorithm learns while being evaluated using the objective variables, and at the time of estimation, it is able to predict the species from the unknown explanatory variables. The objective is to predict the species.

[2] Machine learning algorithm expressed in terms of maximum likelihood estimation method. The conditional probability distribution p(y|x) represents the probability distribution of y when we know that the value of x is a specific value for x and y.



Histogram consisting of feature vectors of petal_width. When this is scaled and the sum of the total area is equal to 1, this vector is the probability vector x.

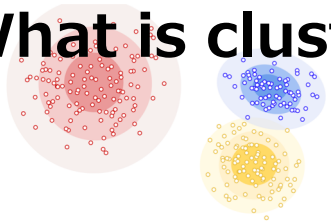
Supervised and unsupervised learning

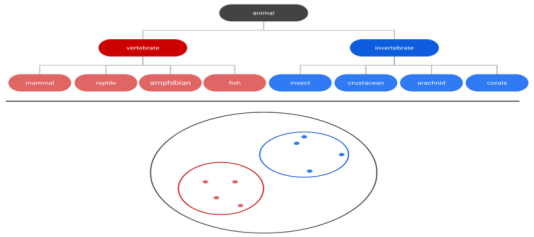
Supervised learning vs. unsupervised learning

In unsupervised learning, the probability distribution and key characteristics are learned from the probability vector alone.
An example of explicitly learning a distribution is density estimation [3], and its extension tasks include denoising and image generation, which implicitly learn a distribution.
Furthermore, there are algorithms for learning certain characteristics simpler than distributions from data sets, such as clustering and principal component analysis. These are among the classics of unsupervised learning.
A method of estimating from data the probability density function from which it is assumed to have been sampled.

Clustering (k-means method)

What is clustering?



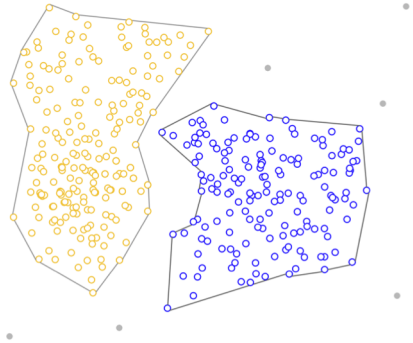


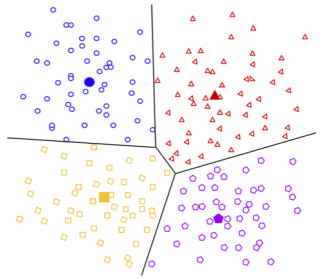
Clustering is an algorithm that collects data sets that are close to each other and divides them into clusters [4]. In here, we will focus on k-means, one of the clustering methods [5]. There are four typical clustering methods: center-of-gravity-based (lower right), density-based (lower left), distribution-based (upper left), and hierarchical clustering (upper right) [6]. k-means is included in the center-of-gravity-based method.

A cluster is a set of data points that are aggregated for a particular similarity [4].

[5] Here is a survey paper that provides a comprehensive survey of clustering.
https://link.springer.com/article/10.1007/s40745-015-0040-1

[6] https://developers.google.com/machine-learning/clustering/clustering-algorithms





k-means method

What is the k-means method?

The k-means algorithm is a mechanism that identifies k (fixed number) centers of gravity, performs iterative calculations to optimize the center of gravity location (minimize the mean), and assigns all data points to the closest cluster. By reducing the sum of error squares [7] in the clusters, all data points are assigned to each cluster.

[7] See slide p13.

k-means method

What are the given conditions?

In the k-means method, the following assumptions are made

① Input data is only feature matrix X (unsupervised learning)
② Enter a fixed value k as a hyperparameter

k-means method

Preparing for Sprint

Geometric explanation of k-means

① Let k data points randomly sampled from the data distribution be the center of gravity of the cluster (k is a hyperparameter)
② Calculate the Euclidean distance to all data points for each center of gravity.
③ The data point county with the smallest distance to each center of gravity is the cluster attributed to that center of gravity.
④ For each of the k clusters, find the point that is the average of the data and make it the new center of gravity.
⑤ Return to ②.

k-means method

Preparing for Sprint

Review the implementation steps.

① Assign random initial labels for k classes to the index of the number of samples.

② Create clusters by grouping data points for each label

③ Find the average value of the data points for each cluster and use it as the center of gravity for that cluster.

④ Calculate the distance from its center of gravity to the data points of all samples.

⑤ Assign that data point to the cluster of centers of gravity with the smallest distance from each data point.

⑥ Repeat steps ③ to ⑤.

⑦ Exit when the convergence conditions (value does not change, the defined number of iterations has been reached, etc.) are met.

⑧ Change the initial value, repeat ① to ⑦ n times, and select the one with the smallest SSE.

k-means method

(Coordinates of the data point) - (Coordinates of the center of gravity)

$$SSE = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \| X_n - \mu_k \|^2$$

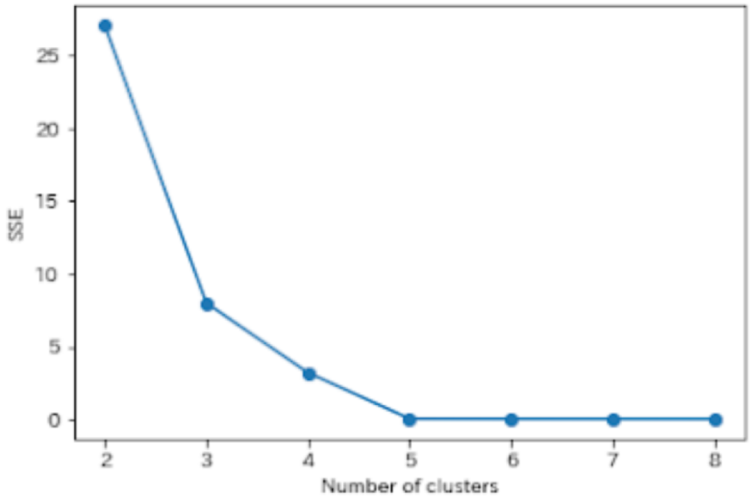
1 if it is a cluster you belong to
0 if it is a cluster to which you do not belong to

About SSE

Sum of Squared Errors (SSE) is a performance evaluation function for clustering.

About the Elbow Method

This is one of the methods to determine the number of clusters. Make a graph with SSE on the vertical axis and the number of clusters on the horizontal axis as shown in the figure on the right, and determine the number of clusters by looking at the points that are bent like elbows.



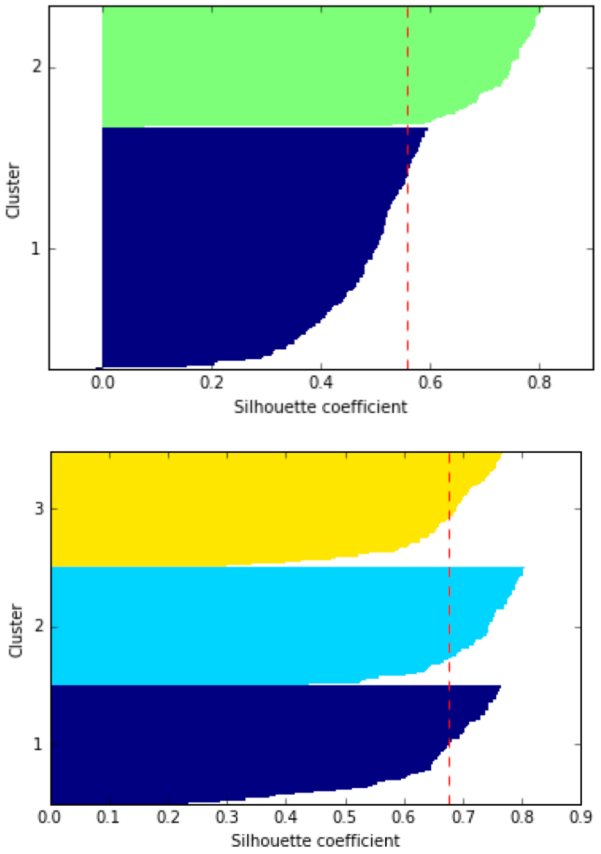
k-means method

About Silhouette Analysis

One of the methods to determine the number of clusters. Create a silhouette diagram with the vertical axis as the index of the sample (sorted by cluster) and the horizontal axis as the silhouette coefficient. The dotted line represents the average of the silhouette coefficients. Create several silhouette diagrams with different numbers of clusters, and select the following ones

- thickness is approximately equal
- Every cluster has a certain amount of samples that exceed the dotted line.

For example, in the example on the right, k=3 (bottom) is preferable to k=2 (top).



Sample code

How to solve problems
“Scratch Clustering”

[Problem 1] Determine the initial value of the center point
[Problem 2] Creation of a function to obtain SSE
[Problem 3] Allocation to cluster
[Problem 4] Movement of the center point
[Problem 5] Repeat
[Problem 6] Calculate with different initial values
[Problem 7] Estimate
[Problem 8] Implementation of elbow method
[Problem 9] (Advance assignment) Silhouette diagram
[Problem 10] Selection of the number of clusters k
[Problem 11] Comparison with known groups
[Problem 12] Useful information for wholesalers
[Problem 13] (Advance assignment) Investigation of other methods
[Problem 14] (Advance assignment) Use of t-SNE and DBSCAN

K-means Class of scikit-learn

Let’s first have a look at the one used until now with the help of the scikit-learn library.

Scikit-learn’s K-means Class

Sprint 7 – Scratch Clustering

Explanation about this Sprint is given but please try it on your own first.

Sprint 7 – Scratch Clustering

Please work on your own after class and submit your assignments on DIVER.

Sprint 7 – Scratch Clustering

A Sample Code of this Sprint is given but please try it on your own.

Sprint 7 – Scratch Clustering

Please work on your own after class and submit your assignments on DIVER.

ToDo by next class

Next class will be Zoom : Thursday 17 June 2021 19:30~20:30

🚨ToDo: Sprint Ensemble Learning
https://diver.diveintocode.jp/curriculums/1868

Check-out

3 minutes Please post the following point to Zoom chat.

Q. What do you do to graduate on time

Thank You For Your Attention