Clustering ⇐ unsupervised method
dimensional reduction (입력만)

Inclass 10: K-Means Clustering

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

중간고사

- 10월 26일 (수) 오전 10시 (수업시간)
- P404호
- clustering까지, (이론, 실습, 과제)
- Cheet sheet 1장 (양면, 2쪽), A4용지.
- Closed book

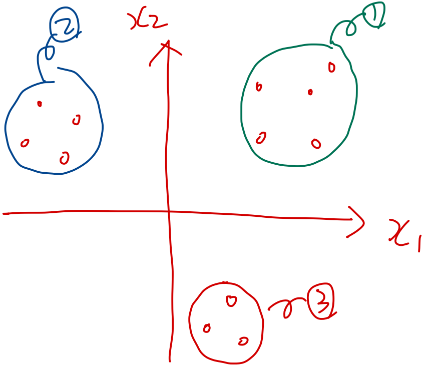
군집화

입력샘플 → Clustering → cluster label

γ ①

γ ②

γ ③



⇒ ① Kmeans clustering

② mixture of Gaussian.

입력

$x_1, x_2, \cdots, x_N$

→ K means clustering → one-of-K coding label

$\mathbb{R}^{N \times K}$

cluster의 개수 K

K개의 중심 = 평균

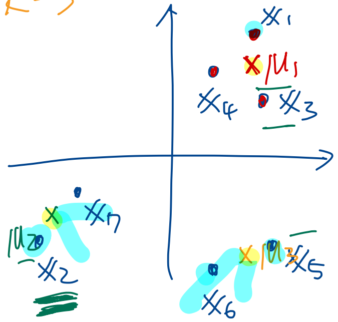
centroid mean

iterative method

주어진 샘플 → label을 계산

주어진 K개의 중심 ← 주어진 샘플

$K=3$



$$\boxed{\begin{matrix} R \\ N\times K \\ 7\times 3 \end{matrix}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\mu_1 \leftarrow \frac{1}{3}(x_1 + x_3 + x_4)$$

$$\mu_2 \leftarrow \frac{1}{2}(x_2 + x_7)$$

$$\mu_3 \leftarrow \frac{1}{2}(x_5 + x_6)$$

주어진 sample

$x_1 \cdots x_7$

주어진 label R

↓

중심 $\mu_1 \cdots \mu_3$ 갱신

0/1

$$\underset{N\times K}{D} = \| \underline{x_n} - \mu_k \|_2$$

$$\underset{5\times 3}{\underline{D}} = \begin{array}{ccc} \mu_1 & \mu_2 & \mu_3 \\ \end{array}$$

$$\begin{bmatrix} 1 & 4 & 9 \\ 2 & 5 & 6 \\ 8 & 2 & 7 \\ 6 & 1 & 9 \\ 10 & 5 & 1 \end{bmatrix} \rightarrow \underset{5\times 3}{R} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

K-means clustering

초기화.

샘플 + centroid → label

샘플 + label → centroid

K=2

↓

μ의 초기위치 결정

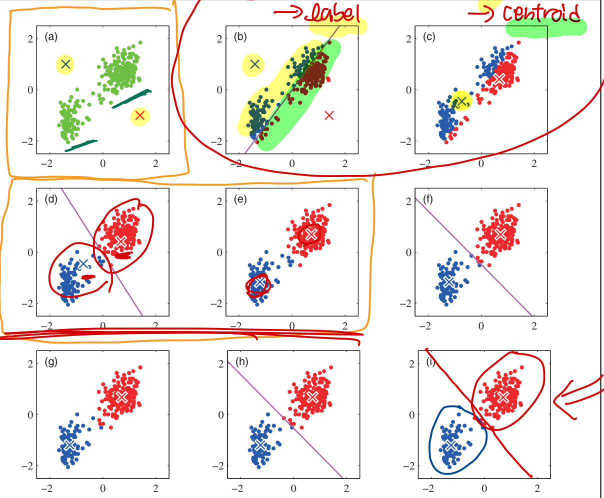


Figure 9.1 Illustration of the K-means algorithm using the re-scaled Old Faithful data set. (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on. (c) In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive E and M steps through to final convergence of

K-means: method

We begin by considering the problem of identifying groups, or clusters, of data points in a multidimensional space. Suppose we have a data set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ consisting of N observations of a random D-dimensional Euclidean variable $\mathbf{x}$. Our goal is to partition the data set into some number K of clusters, where we shall suppose for the moment that the value of K is given.

Intuitively, we might think of a cluster as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster. We can formalize this notion by first introducing a set of D-dimensional vectors $\boldsymbol{\mu}_k$, where $k = 1, ..., K$, in which $\boldsymbol{\mu}_k$ is a prototype associated with the k-th cluster.

K-means: method

As we shall see shortly, we can think of the $\boldsymbol{\mu}_k$ as representing the centres of the clusters. Our goal is then to find an assignment of data points to clusters, as well as a set of vectors $\{\boldsymbol{\mu}_k\}$, such that the sum of the squares of the distances of each data point to its closest vector $\boldsymbol{\mu}_k$, is a minimum.

It is convenient at this point to define some notation to describe the assignment of data points to clusters. For each data point $\mathbf{x}_n$, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, ..., K$ describing which of the K clusters the data point $\mathbf{x}_n$ is assigned to, so that if data point $\mathbf{x}_n$ is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$. This is known as **the 1-of-K coding scheme**.

K-means: method

We can then define an objective function, sometimes called a *distortion measure*, given by

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \qquad (1)$$

which represents the sum of the squares of the distances of each data point to its assigned vector $\boldsymbol{\mu}_k$.

Our goal is to find values for the $\{r_{nk}\}$ and the $\{\boldsymbol{\mu}_k\}$ so as to minimizes J.

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\sum_{n=1}^{3}\sum_{k=1}^{2} r_{nk}\|x_n - \mu_k\|^2$$

$n=1$ $$\sum_{k=1}^{2} r_{1k}\|x_1 - \mu_k\|^2$$

$$= 1\cdot\|x_1-\mu_1\|^2 + 0\cdot\cancel{\|x_1-\mu_2\|^2}$$

The terms involving different n are independent and so we can optimize for each n separately by choosing r_{nk} to be 1 for whichever value of k gives the minimum value of $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$. In other words, we simply assign the n-th data point to the closest cluster centre.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

R

$\|x_1 - \mu_1\|$ $\|x_1 - \mu_2\|$ $\|x_1 - \mu_3\|$

5 1 9

0 1 0

K-means: method

Now consider the optimization of the $\boldsymbol{\mu}_k$ with the r_{nk} held fixed. The objective function J is a quadratic function of $\boldsymbol{\mu}_k$, and it can be minimized by setting its derivative with respect to $\boldsymbol{\mu}_k$ to zero giving

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{3}$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} \tag{4}$$

k번째에 속하는 샘플의 합

k번째에 속하는 샘플개수

The denominator in this expression is equal to the number of points assigned to cluster k, and so this result has a simple interpretation, namely set $\boldsymbol{\mu}_k$ equal to the mean of all of the data points $\mathbf{x}_n$ assigned to cluster k.

The two phases of re-assigning data points to clusters and re-computing the cluster means are repeated in turn until there is no further change in the assignments (or until some maximum number of iterations is exceeded).

K-means: python example

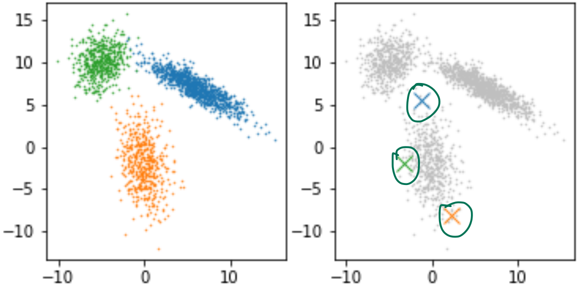


Figure 1: K-means algorithm python example: dataset (left) and the initialization (right).

K-means: python example

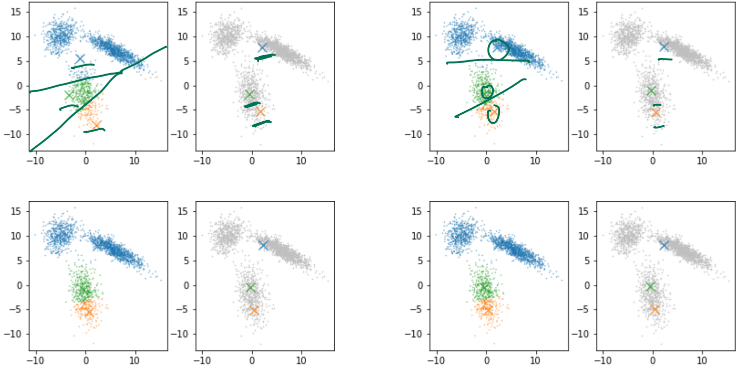


Figure 2: From the 1st to 4th iterations.

K-means: python example

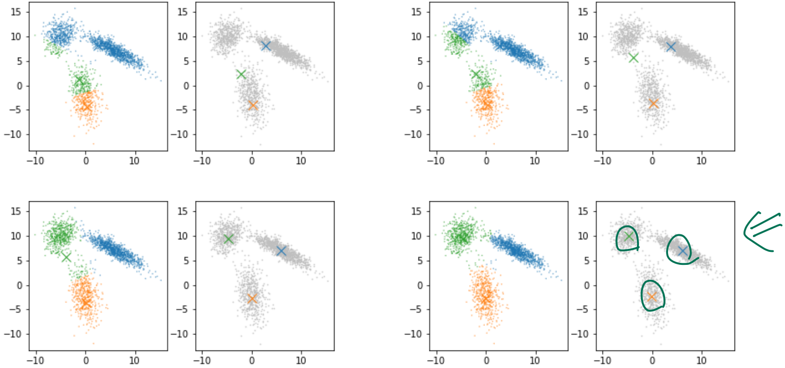


Figure 3: From the 9th to 12th iterations.

K-means: example

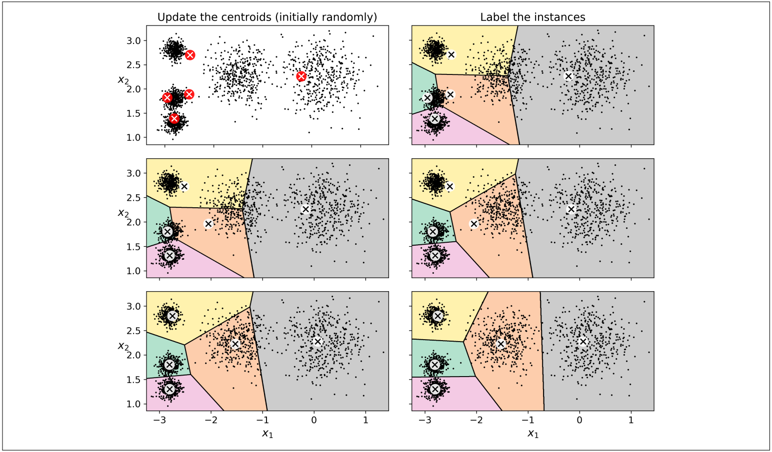
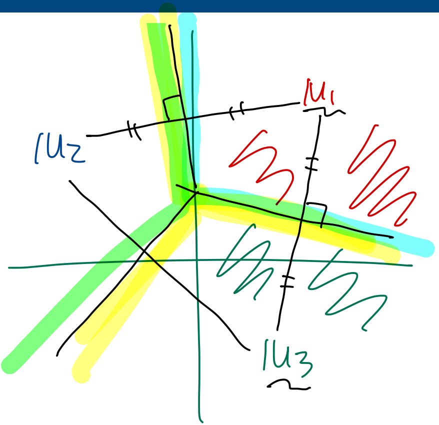


Figure 9-4. The K-Means algorithm



① convex

② single connect

$\not\uparrow$

수직이등분선

K-means: decision boundary

Decision boundary = Voronoi tessellation

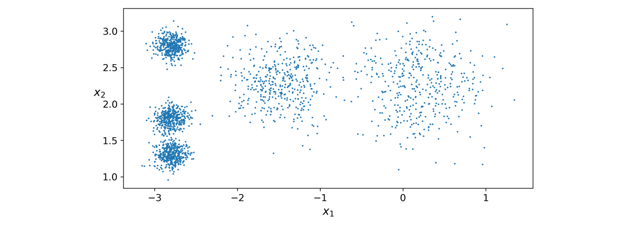


Figure 9-2. An unlabeled dataset composed of five blobs of instances

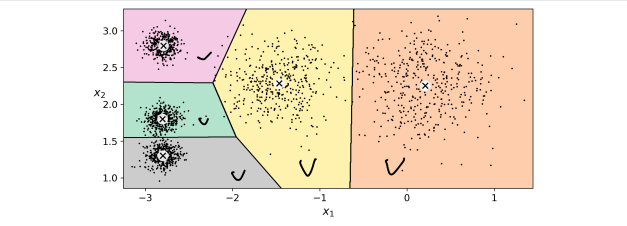


Figure 9-3. K-Means decision boundaries (Voronoi tessellation)

Appendix

Reference and further reading

- “Chap 9” of A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- “Chap 9” of C. Bishop, Pattern Recognition and Machine Learning
- Variational Bayesian mixtures of Gaussians