Clustering

Instructor: Xiaoqian Wang
10/14/2024

Motivation of Clustering

- Assumption in most learning methods: Similar data has similar behavior.

Example of clustering: image compression

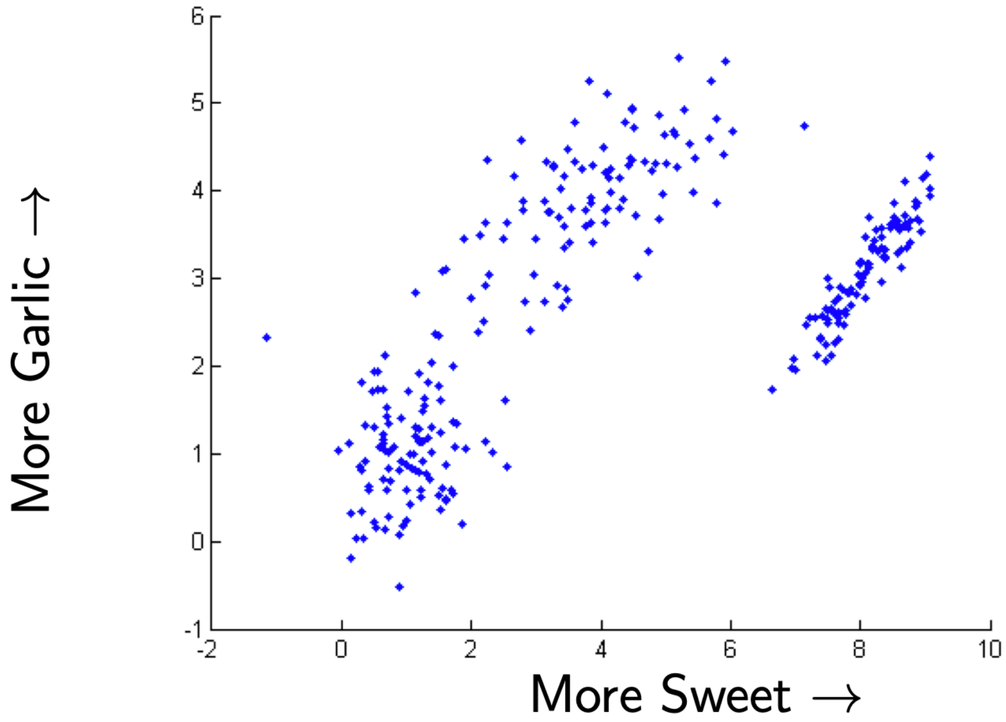
Image credit: https://towardsdatascience.com/image-compression-using-k-means-clustering-aa0c91bb0eeb

Example of clustering: unsupervised image segmentation

Image credit: R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2282, Nov. 2012, doi: 10.1109/TPAMI.2012.120.

Example of clustering: grouping customers

- A major tomato sauce company wants to tailor their brands to sauces to suit their customers
- They run a market survey where the test subject rates different sauces



Example from https://www.cs.toronto.edu/~urtasun/courses/CSC411/tutorial7.pdf

Example of clustering: precision medicine

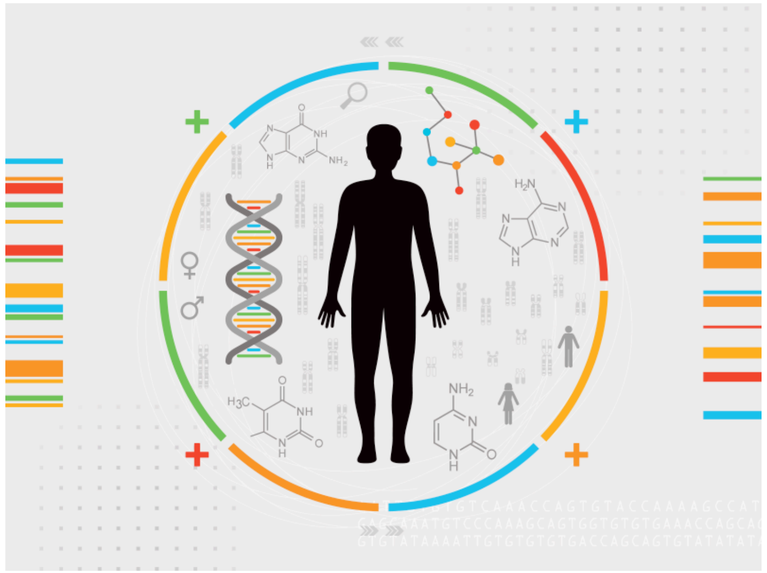


Image credit: https://www.kaufmanhall.com/insights/article/precision-medicine-future-now

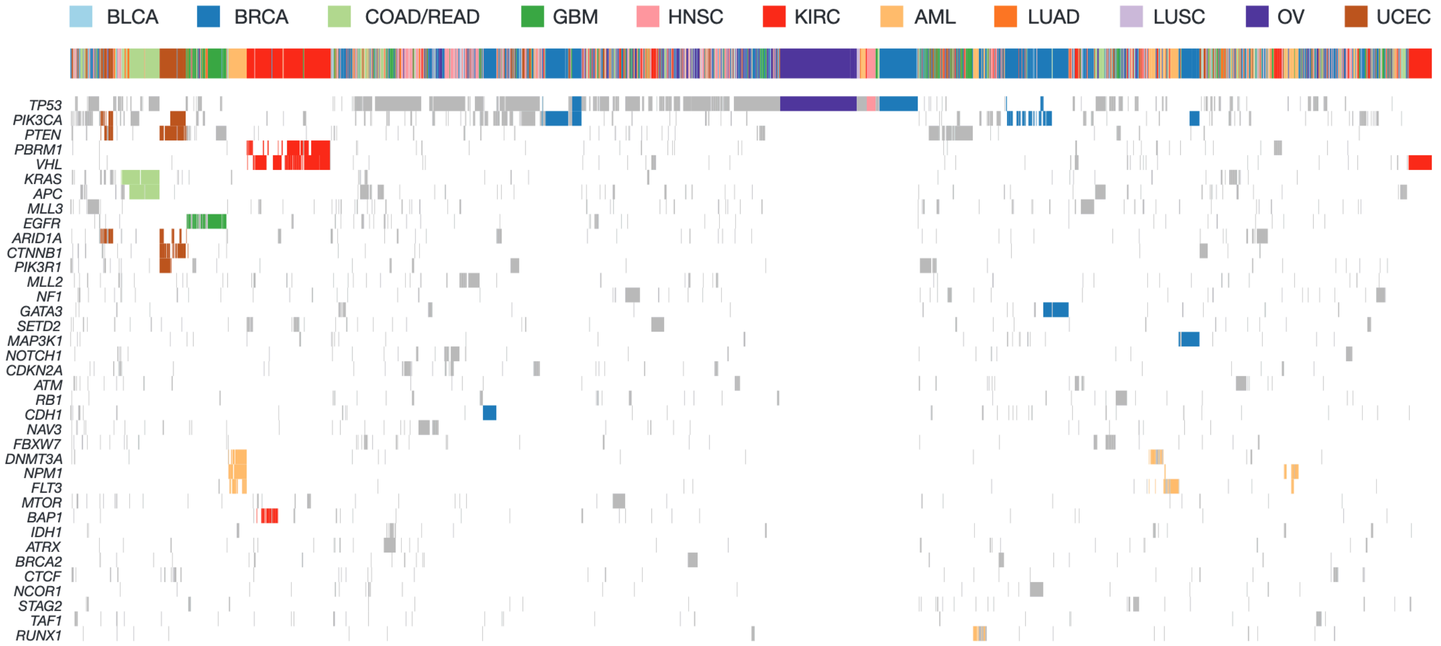
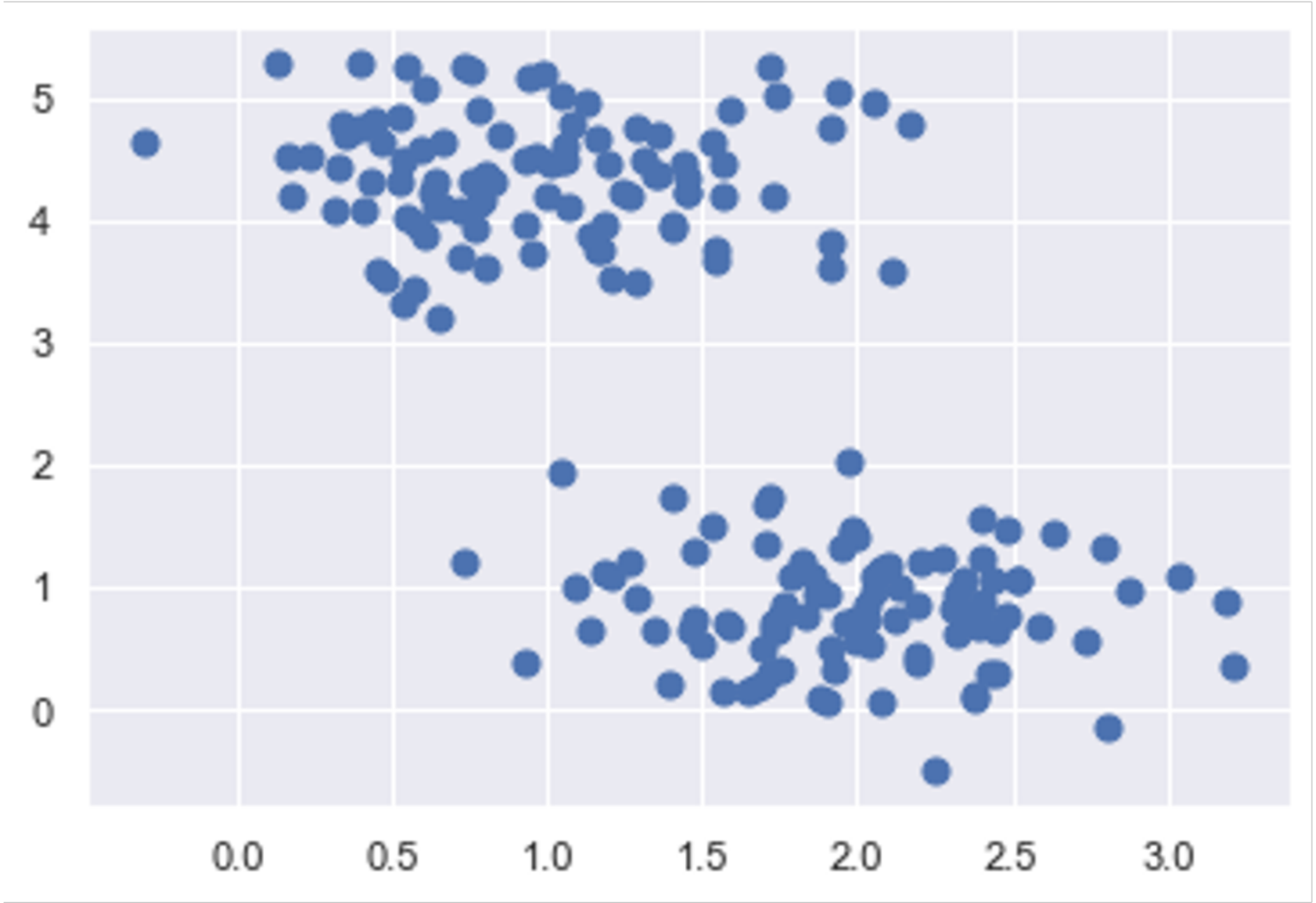


Image credit: Kandoth, Cyriac, et al. "Mutational landscape and significance across 12 major cancer types." Nature 502.7471 (2013): 333-339.

Demo

- How to put these points into two clusters?



K-Means Clustering

- Objective function

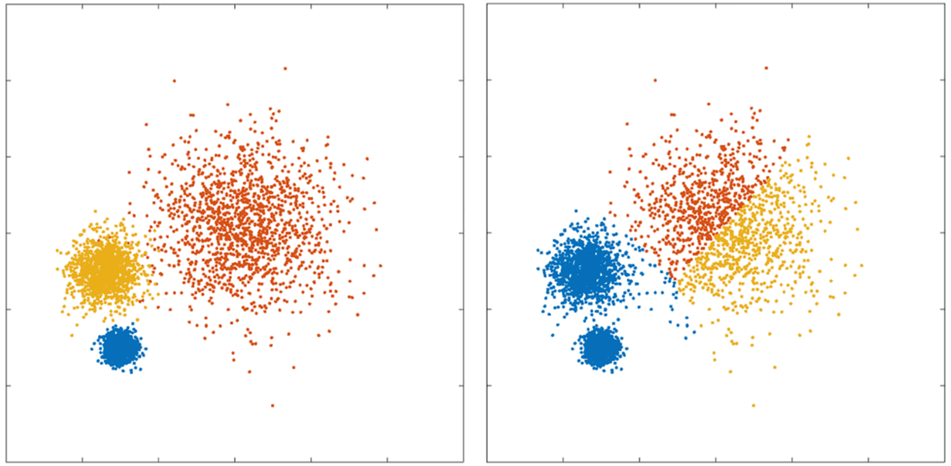
$$\min_{C_1, C_2, \ldots, C_K, \mu_1, \mu_2, \ldots, \mu_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mu_k\|^2$$

K-Means Algorithm

- Initialize K centroids
- Iterate until convergence
    - Assignment of clusters
    - Update of K centroids

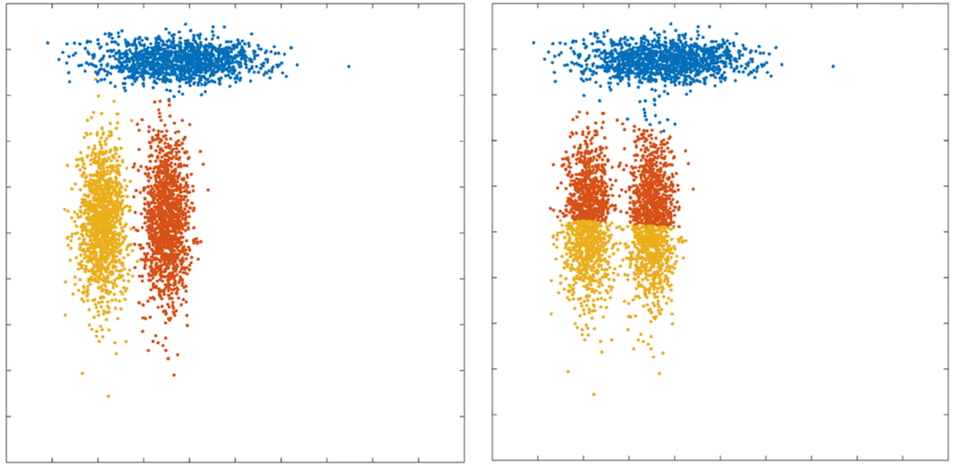
K-mean is an efficient algorithm and usually provides good solutions in practice

Demo of caveats with K-Means



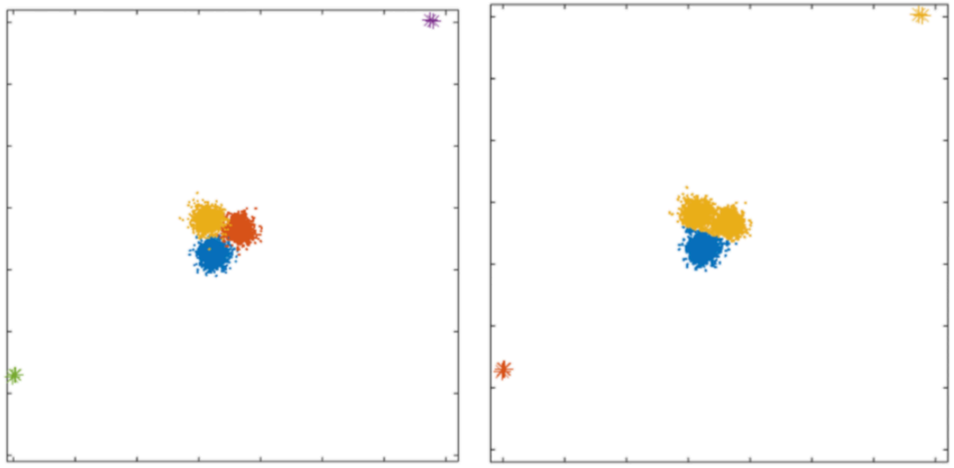
(a) Generated synthetic data (b) K-means

1. When clusters have different size and density



(a) Generated synthetic data (b) K-means

2. When clusters are non-spherical



(a) Generated synthetic data (b) K-means

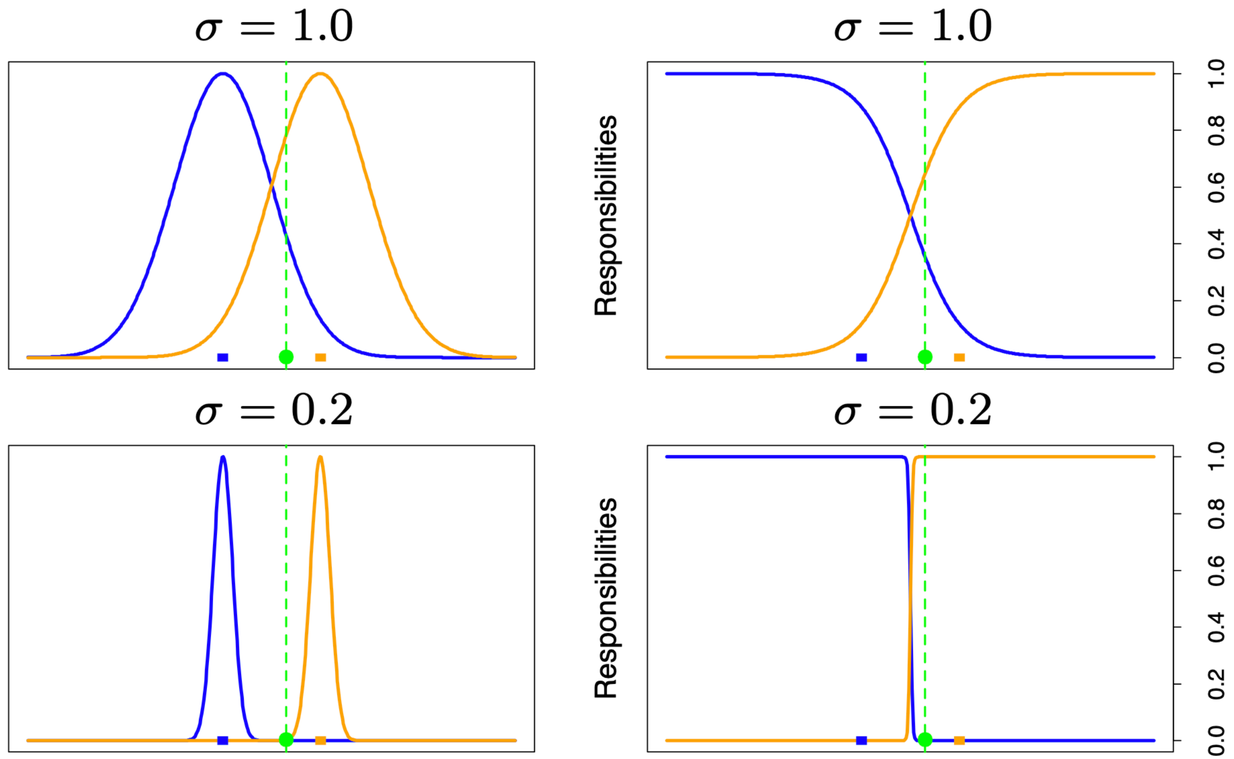
3. When outliers are in the data

4. Hard clustering assignments

Image credit: Raykov, Yordan P., et al. "What to do when k-means clustering fails: a simple yet principled alternative algorithm." *PloS one* 11.9 (2016): e0162259.

Gaussian Mixture Models

- Mixture Model $f(x) = \sum_{j=1}^{k} \pi_j g_j(x)$
- Gaussian Mixture $g_j(x) = \phi_{\theta_j}(x), \theta_j = (\mu_j, \sigma_j^2)$



Content from http://statweb.stanford.edu/~tibs/stat315a/LECTURES/em.pdf

Gaussian Mixture Models

- Suppose $(\mu, \mathbf{\Sigma}, \pi)$ is given, we can evaluate clustering assignment $\mathbf{z}$ as:

$$p(z_j = 1 \mid \mathbf{x}) = \frac{\pi_j \phi_{\theta_j}(\mathbf{x})}{\sum_j \pi_j \phi_{\theta_j}(\mathbf{x})}$$

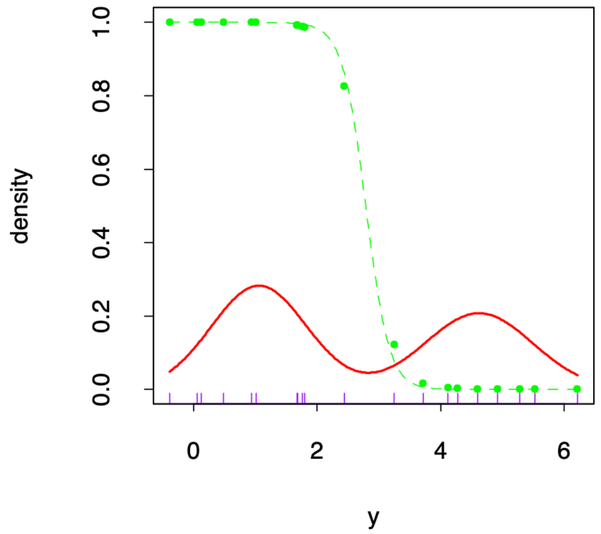


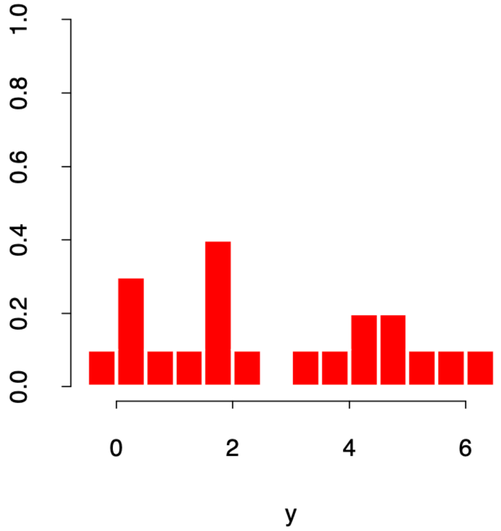
Figure from http://statweb.stanford.edu/~tibs/stat315a/LECTURES/em.pdf

Gaussian Mixture Models

Density of X is $g_X(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \phi_{\theta_j}(\mathbf{x})$

Given n training samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the log-likelihood is

$$l(\theta, X) = \sum_{i=1}^{n} \log[\sum_{j=1}^{k} \pi_j \phi_{\theta_j}(\mathbf{x}_i)]$$



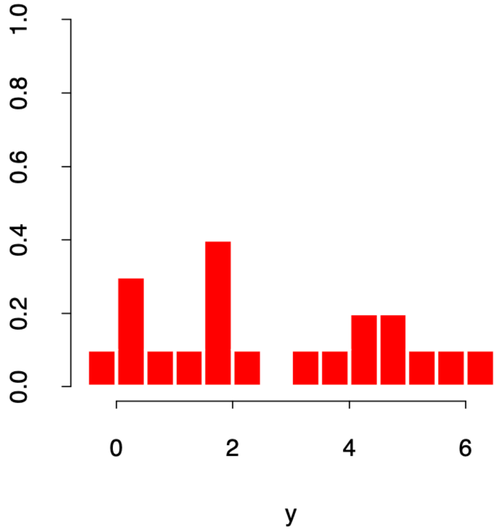
Content based on http://statweb.stanford.edu/~tibs/stat315a/LECTURES/em.pdf

Gaussian Mixture Models

Density of X is $g_X(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \phi_{\theta_j}(\mathbf{x})$

Given n training samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, the log-likelihood is

$$l(\theta, X) = \sum_{i=1}^{n} \log[\sum_{j=1}^{k} \pi_j \phi_{\theta_j}(\mathbf{x}_i)]$$



Suppose $\mathbf{z}$ is given, the log-likelihood can be rewritten as

$$l(\theta, X, \mathbf{z}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}[\log \pi_j + \log \phi_{\theta_j}(\mathbf{x}_i)]$$

Content based on http://statweb.stanford.edu/~tibs/stat315a/LECTURES/em.pdf

Expectation Maximization Algorithm

- Initialize $\hat{\mu}, \hat{\sigma}^2, \hat{\pi}$
- Expectation Step

$$\gamma_{ij} = p(z_j = 1 \,|\, x = \mathbf{x}_i) = \frac{\pi_j \phi_{\theta_j}(\mathbf{x}_i)}{\sum_j \pi_j \phi_{\theta_j}(\mathbf{x}_i)}$$

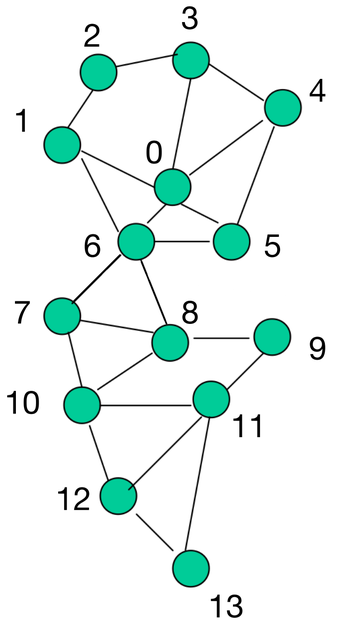
- Maximization Step

$$\hat{\mu}_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}, \hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \gamma_{ij} (\mathbf{x}_i - \hat{\mu}_j)^2}{\sum_{i=1}^n \gamma_{ij}}, \hat{\pi}_j = \frac{\sum_{i=1}^n \gamma_{ij}}{n}$$

Graph-based clustering with demo

A **graph** G = (V, E) consists two components:

- nodes (vertices) V
- edges E between the nodes



A graph with n nodes can be represented by an **adjacency matrix** $A \in \mathbb{R}^{n \times n}$

- Adjacency matrix for unweighted edges E: all edges have weight 1
- Adjacency matrix for weighted edges E:

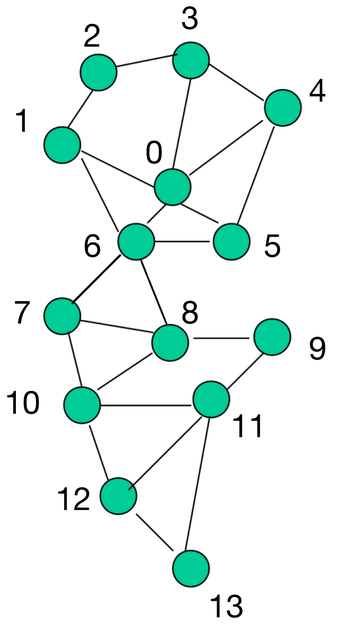
$$\mathrm{A}(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Figure from http://ranger.uta.edu/~chqding/Spectral/spectralA.pdf

Graph-based clustering with demo

A **graph** G = (V, E) consists two components:

- nodes (vertices) V
- weighted edges E between the nodes



A graph with n nodes can be represented by an **adjacency matrix** $A \in \mathbb{R}^{n \times n}$

The **degree matrix** defined as a diagonal matrix whose elements are the sum of rows of A:

$D = \text{diag}(A1)$

Graph Laplacian is defined as L = D - A

Figure from http://ranger.uta.edu/~chqding/Spectral/spectralA.pdf

Graph-based clustering

The number of 0 eigenvalues of the Laplacian is the number of connected components

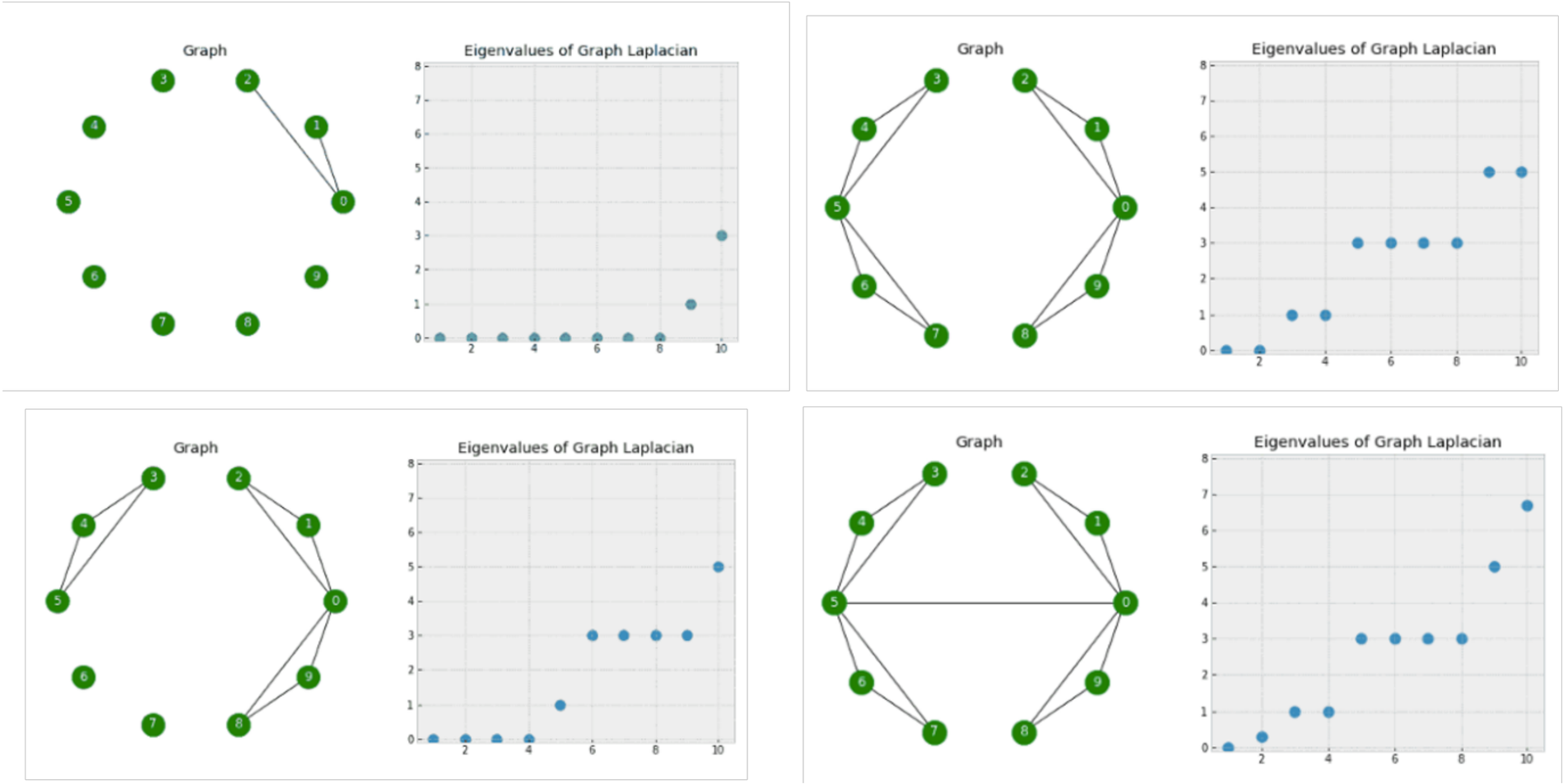
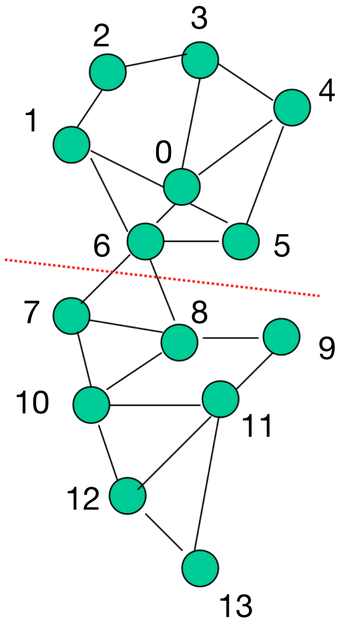


Figure from https://towardsdatascience.com/spectral-clustering-aba2640c0d5b

Graph-based clustering

Motivation of spectral clustering is to minimize cut size

Ratio Cut: $\frac{s(C_1, C_2)}{|C_1|} + \frac{s(C_1, C_2)}{|C_2|}$, where $s(C_1, C_2) = \sum_{i \in C_1} \sum_{j \in C_2} a_{ij}$



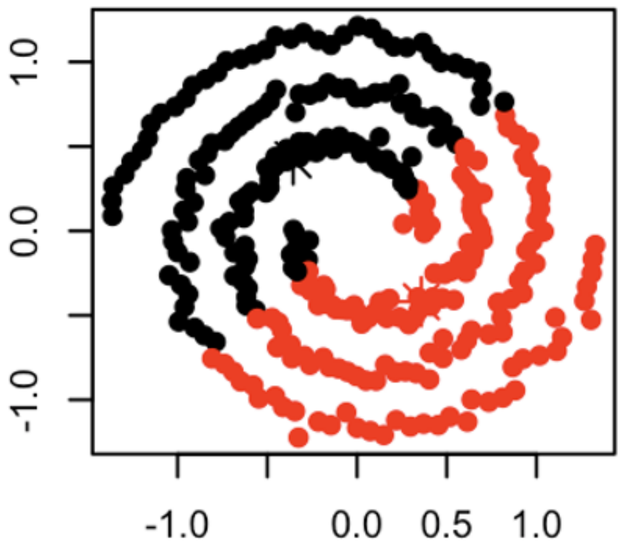
Spectral clustering generalizes to $k>2$ clusters:

- Take the 2nd to k-th lowest eigenvectors as a new node representation
- Then run K-means on the new node representation

Content based on http://ranger.uta.edu/~chqding/Spectral/spectralA.pdf

K-means focuses on “compactness” of each cluster

K-means assumes that points in a cluster are close to each other, and is only applicable to linear boundaries and circular clusters



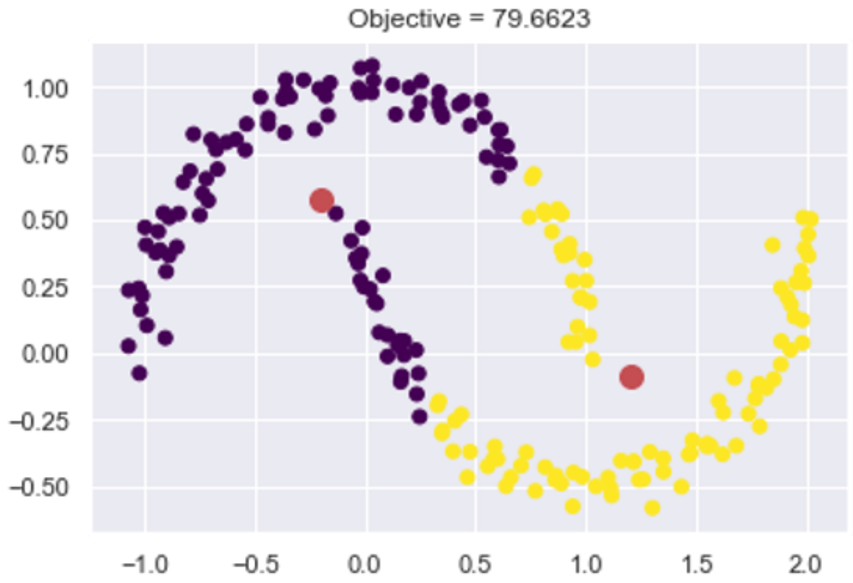
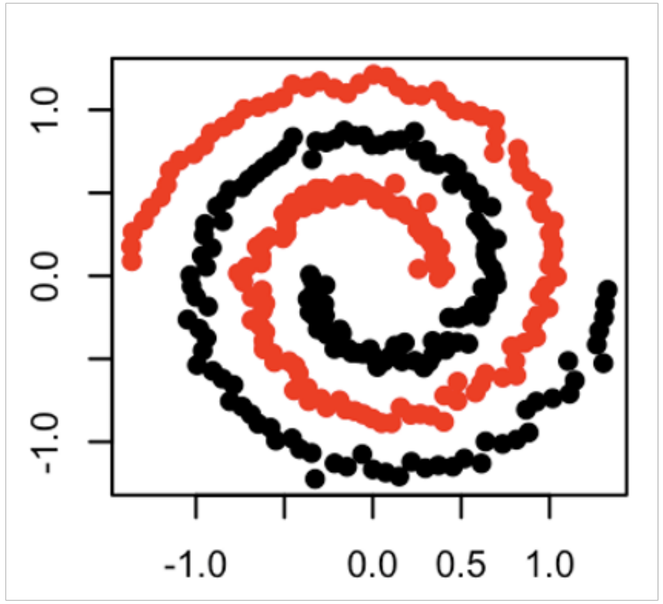


Figure from https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7

Spectral clustering focuses on “connectivity” of each cluster

Spectral clustering assumes that points in a cluster are connected to each other, and is applicable non-circular clusters



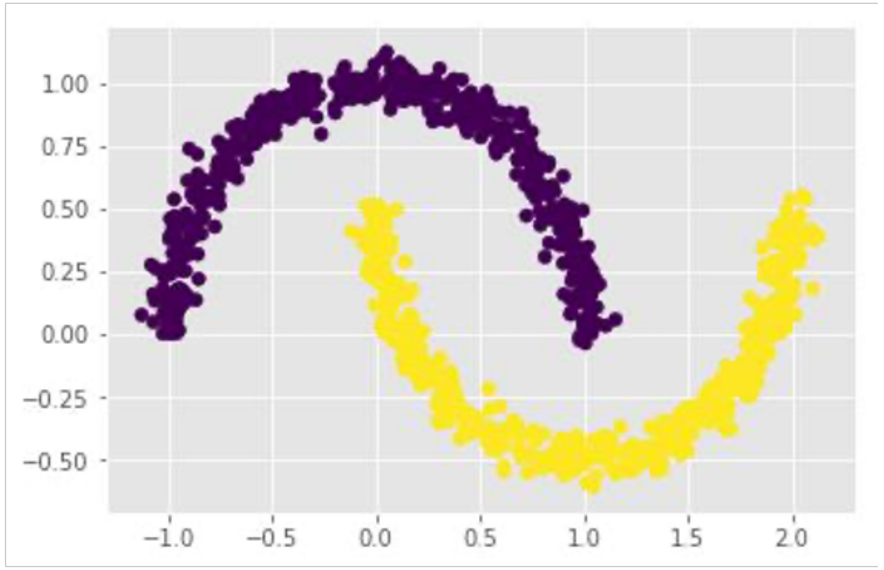


Figure from https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7