INTELLIGENT SYSTEMS LABORATORY

Lesson 8
Cluster Analysis

Mathematics and Statistics for Data Science

Tapanan Yeophantong
Vincent Mary School of Science and Technology
Assumption University

Content

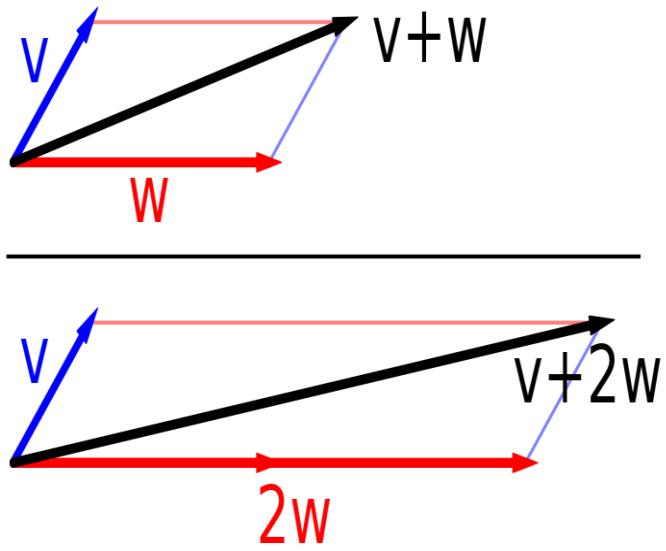
- Clustering techniques & their applications
- Metrics for evaluating clusters

Vector Arithmetic

Let's first look at the basics ...

Vector Space

- Vector spaces are the subject of linear algebra.
- Their dimension specifies the number of independent directions in the space.
- Objects are represented in a vector space as vectors, which may be added with another vector or multiplied by numbers called scalars.

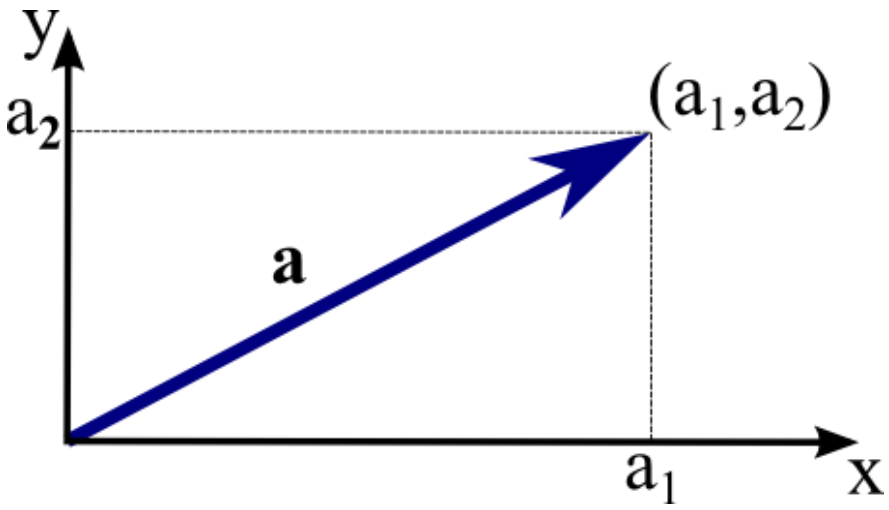


Vector Basics

- Vector addition & subtraction
- Scalar multiplication
- Scalar (dot) product
- Vector (cross) product

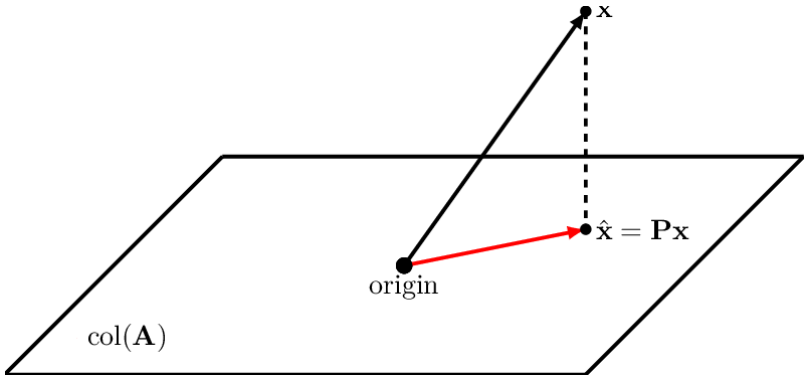
Vector in a Cartesian Plane

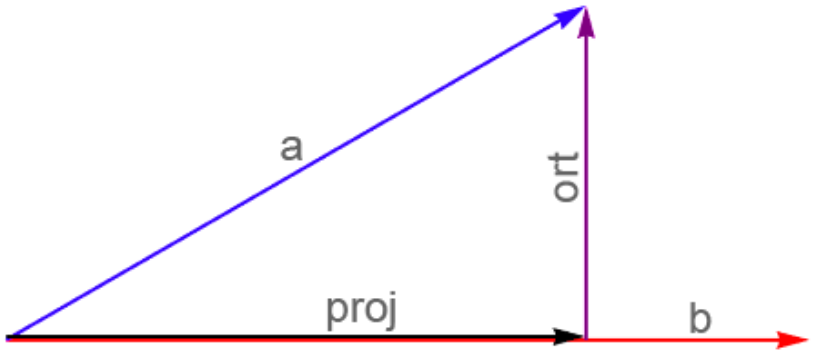
- Vectors can easily represented in a Cartesian plane, where each point in the plane is identified by its x and y components.
- To determine the coordinate, translate the vector so that tail is at the origin.
- Then, the head of the vector will be at some point (a_1, a_2). We call this the coordinate of the vector.
- We denote $\mathbf{a} \in \Re^2$ to say that the vector can be described in real vector space.



Vector Projection

- A projection is a linear transformation from a vector space to itself.
- That is, "projecting" a vector space in $\mathfrak{R}^n$ to a vector space in $\mathfrak{R}^{n-1}$.
  - For example, a function that maps (x, y, z) in $\mathfrak{R}^3$ to (x, y, z) in $\mathfrak{R}^2$.
- A very important type of projection is called an orthogonal projection.
  - Decomposition (e.g. PCA)
  - Least-squares regression (e.g. OLS)

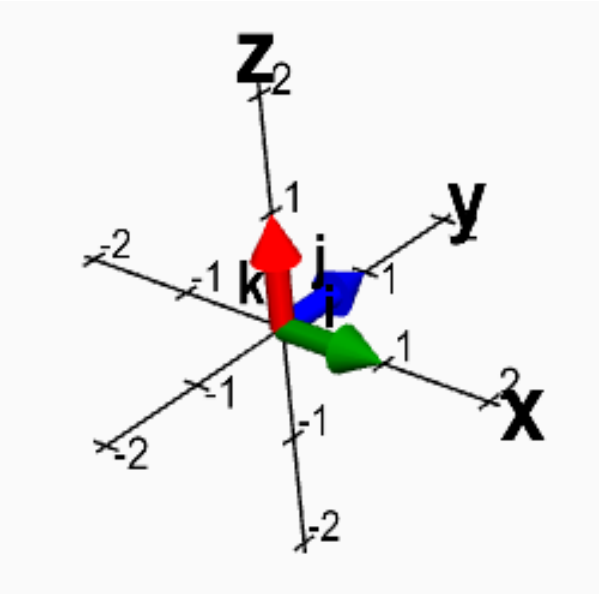




Vectors in 3D Space

- 3D vectors have standard unit vectors **(i, j, k)**.
- These are unit vectors in the positive **(x, y, z)** direction, respectively.
- Everything else works the same way as 2D.

- Up to $\Re^3$, we can still relatively visualize the vectors graphically.
- At higher dimensions, we need to represent them using list of numbers instead.



Vectors & Matrices

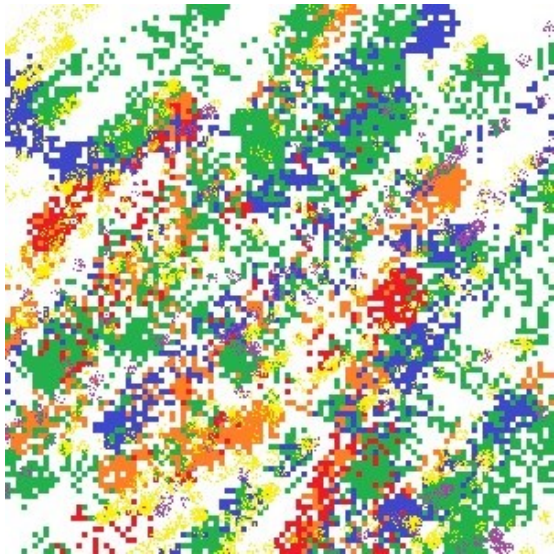
- A vector can be represented as a list of numbers.
- A matrix is an array of numbers, having one or more rows and one or more columns.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ a_{31} & a_{32} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \end{array}$$

Vector Representation of Data

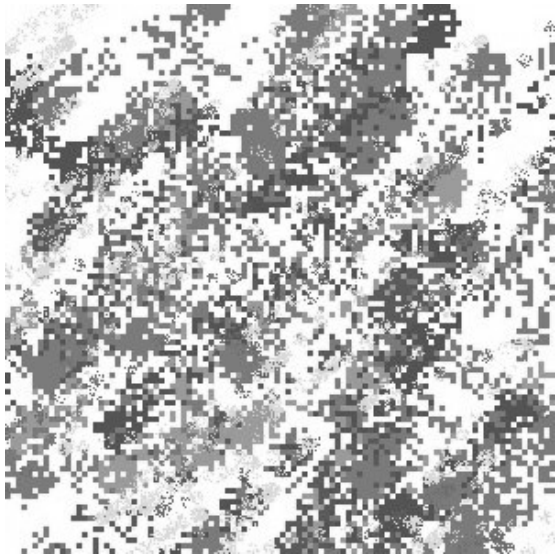
- Many statistical and machine learning techniques work with data represented as vectors of numerical values.
- Common representations:
  - Images: vectors of pixel values
  - Text: vectors of term occurrences/frequencies (also known as bags of words)
  - Categorical data: encoded into numeric array (often binary)
- Vectors of random variables are known as multivariate random variables, or random vectors.

Images as Vectors



Resized to 4x4 pixels

```
[[[ 87 170  48]
  [253 255 254]
  [240 255 231]
  [ 79 171  52]]
 [[ 48 135 245]
  [ 85 171  47]
  [ 87 164  53]
  [198 204 245]]
 [[ 76 177  38]
  [255 241 255]
  [ 76 176  51]
  [241 251 255]]
 [[ 84 159  55]
  [238 255 252]
  [243 255 239]
  [225 255 249]]]
```



[124 254 246 125 158 124 122 216 124 247 127 251 119 252 249 250]
Grayscaled & resized to 4x4

Texts as Vectors

["computer science", "digital technology", "information technology", "software engineering", "information systems", "computer engineering"]

['computer', 'digital', 'engineering', 'information', 'science', 'software', 'systems', 'technology']

[1.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00]
[0.00 1.00 0.00 0.00 0.00 0.00 0.00 1.00]
[0.00 0.00 0.00 1.00 0.00 0.00 0.00 1.00]
[0.00 0.00 1.00 0.00 0.00 1.00 0.00 0.00]
[0.00 0.00 0.00 1.00 0.00 0.00 1.00 0.00]
[1.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00]

(Count Vectors)

[0.63 0.00 0.00 0.00 0.77 0.00 0.00 0.00]
[0.00 0.77 0.00 0.00 0.00 0.00 0.00 0.63]
[0.00 0.00 0.00 0.71 0.00 0.00 0.00 0.71]
[0.00 0.00 0.63 0.00 0.00 0.77 0.00 0.00]
[0.00 0.00 0.00 0.63 0.00 0.00 0.77 0.00]
[0.71 0.00 0.71 0.00 0.00 0.00 0.00 0.00]

(TF-IDF Vectors)

Data as Vectors

```
[['Male','Senior’],['Female','Adult'],['Female','Child’]]
```

```
['Female' 'Male' 'Adult' 'Child' 'Senior']
[0.00   1.00   0.00   0.00   1.00]
[1.00   0.00   1.00   0.00   0.00]
[1.00   0.00   0.00   1.00   0.00]
```

(Sklearn’s OneHotEncoder)

Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

Cosine Similarity

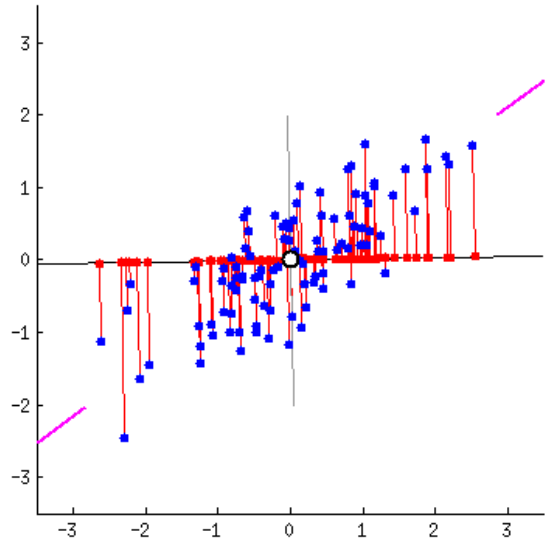
- If d1 and d2 are two vectors, then

$$\cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \|d_2\|}$$

  - where • indicates dot product and ||d|| is the length of vector d.

Principal Component Analysis

- Principal components of a collection of points in real coordinate space are a sequence of unit vectors, where the i^{th} vector is the direction of a line that best fits the data while being orthogonal to the first i-1 vectors.

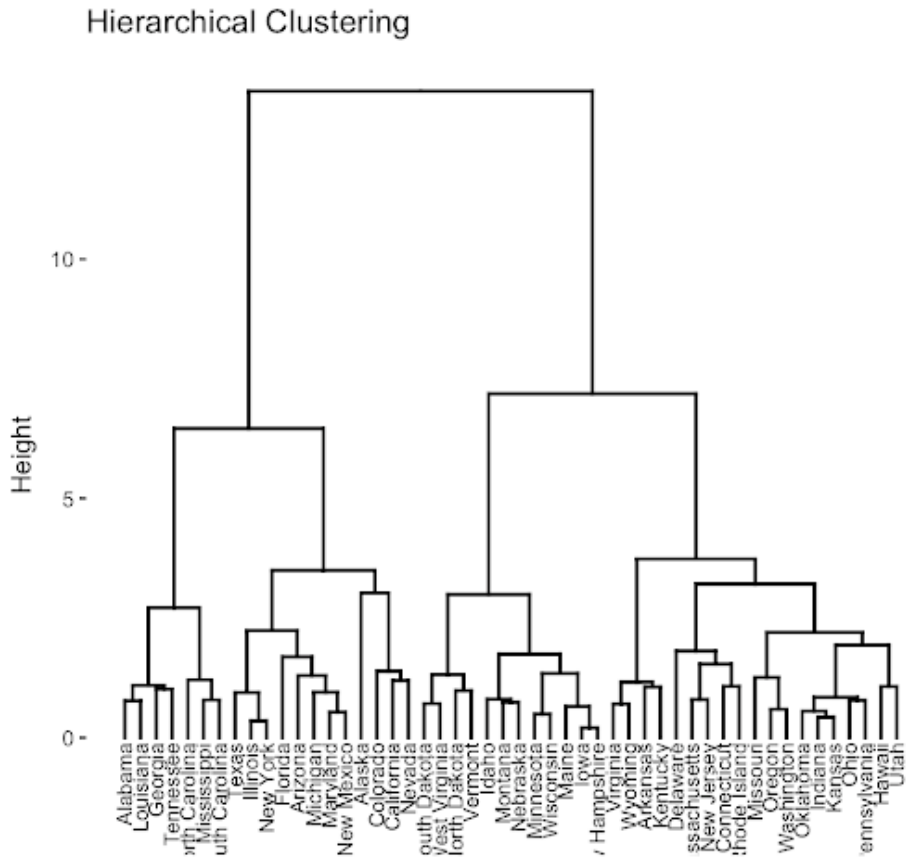


Cluster Analysis

- The task of grouping a set of objects in such a way that objects in the same group (a cluster) are more "similar" to each other than to those in other groups/clusters.
- Common algorithms include:
  - Hierarchical clustering
  - Centroid-based clustering
  - Distribution-based clustering
  - Density-based clustering

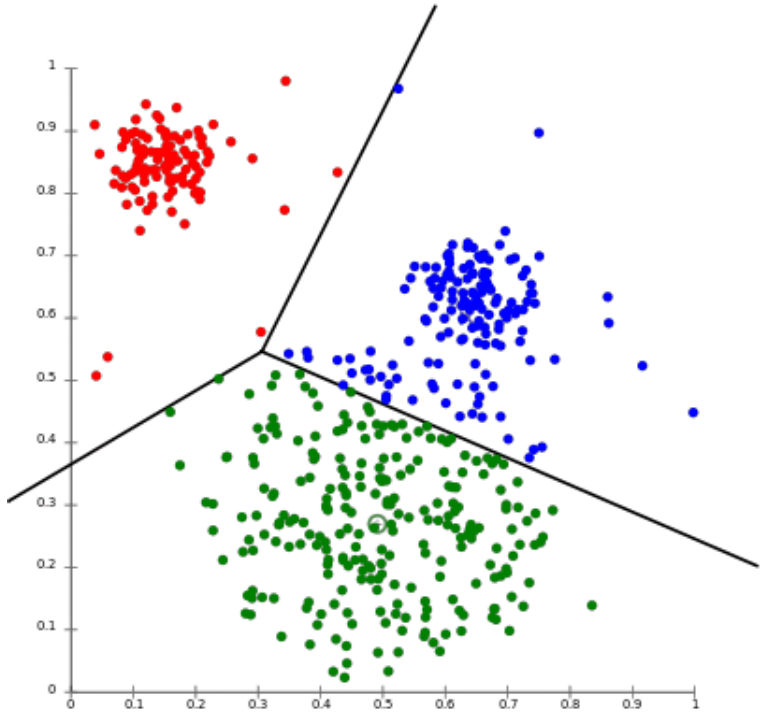
Hierarchical Clustering

- Build nested clusters by merging (agglomerative) or splitting (divisive) them successively.
- Linkage criteria determines the metric used for merging:
  - Single linkage
  - Complete linkage
  - Average linkage
  - Ward



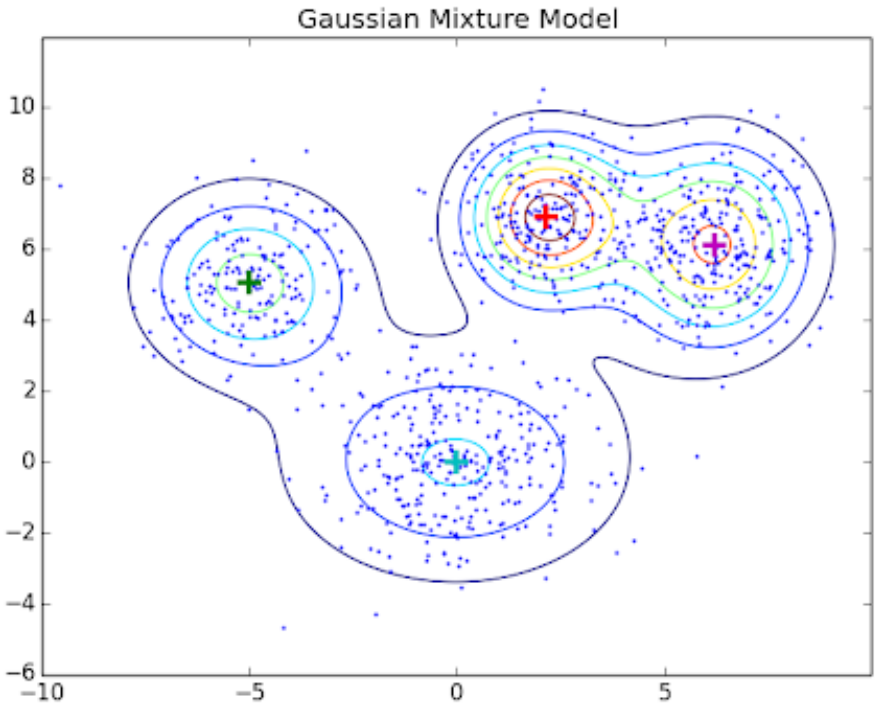
Centroid-based Clustering

- Most common: *k*-means.
- Basic idea: find *k* cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.
- Drawbacks:
  - Need to specify *k*.
  - Does not guarantee optimality.



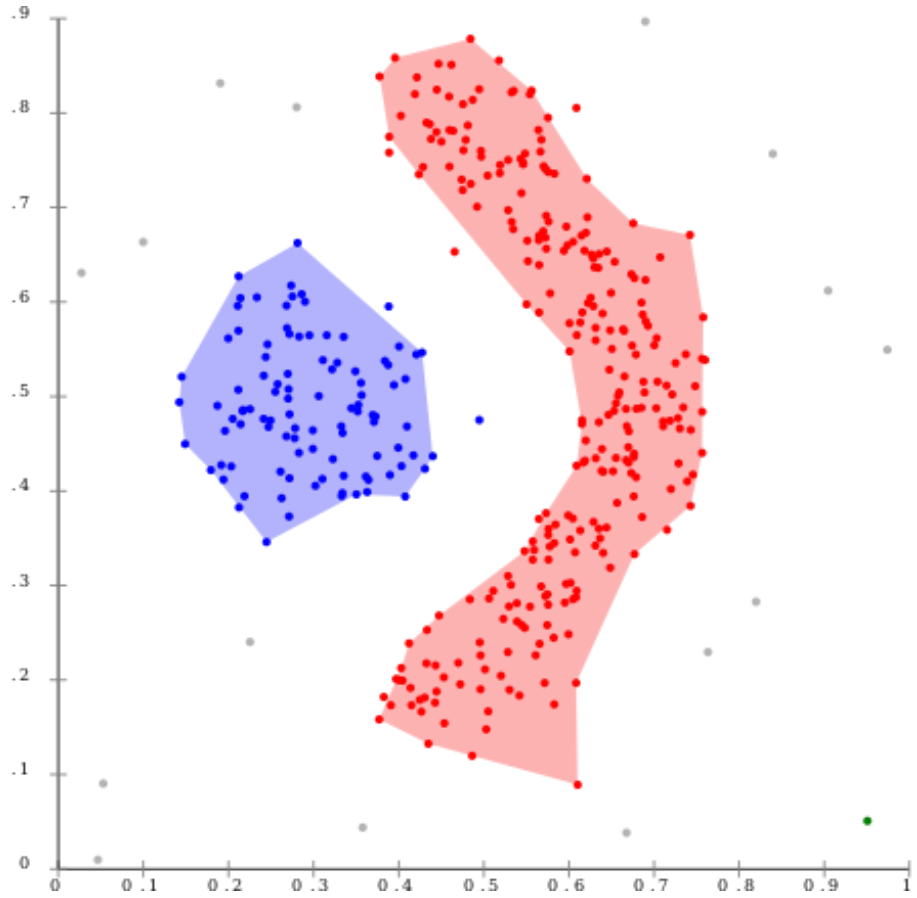
Distribution-based Clustering

- Clusters are defined as objects belonging most likely to the same distribution.
- Common approach is known as Gaussian mixture models.
- It assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.



Density-based Clustering

- Clusters are defined as areas of higher density when compared to the rest of the dataset.
- Objects in sparse areas are considered noise.
- The most popular technique is called DBSCAN.
- Parameters: epsilon & minPts



Homogeneity and Completeness

- Homogeneity
  - Each cluster contains only members of a single class.
- Completeness
  - All members of a given class are assigned to the same cluster.
  - Symmetrical to homogeneity.

V-measure

- Harmonic mean of homogeneity and completeness.

$$v = \frac{(1+\beta) \times \text{homogeneity} \times \text{completeness}}{(\beta \times \text{homogeneity} + \text{completeness})}$$

Silhouette Coefficient

- If the ground truth labels are not known, evaluation must be performed using the model itself.
- The Silhouette Coefficient is defined by:
  - a: Mean distance between a sample and all others in the same class.
  - b: Mean distance between a sample and all others in the next nearest cluster.

$$s = \frac{b - a}{max(a, b)}$$