Clustering validation: Silhouette Coefficient

CSCI 347

Adiesha Liyana Ralalage

F-score

- Balances precision and recall.
- For cluster i:
  - $F_i = \frac{2\left(precision(C_i)\right)\left(recall(C_i)\right)}{\left(precision(C_i)+recall(C_i)\right)}$
- For clustering:
  - $F = \frac{1}{r}\sum_{i=1}^{r} F_i$
- Why does F-score use harmonic mean instead of arithmetic mean?

Different types of means

- Given set of numbers $x_1, x_2, \dots, x_n$
- Arithmetic mean
  - $AM = \frac{x_1+x_2+\cdots+x_n}{n}$
- Geometric mean
  - $GM = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$
- Harmonic mean
  - $HM = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}$
- Quadratic mean (Root mean squared)
  - $RMS = \sqrt{\frac{x_1^2+x_2^2\cdots x_n^2}{n}}$
- Weighted mean
  - $WD = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n}$
- There are other types of mean as well.
- Each of these means serve a different purpose.

MONTANA STATE UNIVERSITY

Different types of means

- In F-score, you have two components, precision and recall.
- We should give a high F-score when both precision and recall is high, and low scores when both precision and recall is low or one of the precision or recall is low.
- But if we use arithmetic mean we cannot do this.
  - Ex: $precision_i = 0.1, recall_i = 0.9$, then if we use the arithmetic mean average would be $\frac{(0.1+0.9)}{2} = 0.5$, but we should not output this as the result.
- If we use Harmonic mean we could avoid this.
  - $HM = \frac{2\cdot(0.1)\cdot(0.9)}{0.1+0.9} = 0.18$
- Balances precision and recall well—penalizes imbalance.

MONTANA STATE UNIVERSITY

SILHOUETTE COEFFICIENT

- Internal measure.
- Useful when you do not have the ground-truth (which is typically the case)

SILHOUETTE COEFFICIENT

- SILHOUETTE COEFFICIENT
- Measure of both cohesion and separation of clusters.
  - Compares mean distance of points to their cluster's points and to other clusters' points.
  - $[-1, +1]$
  - $+1$ indicates that points are in general closer to their cluster's mean than to other clusters' mean.
  - For point i:
    - $s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$
  - Where:
    - $\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}}$

For all points

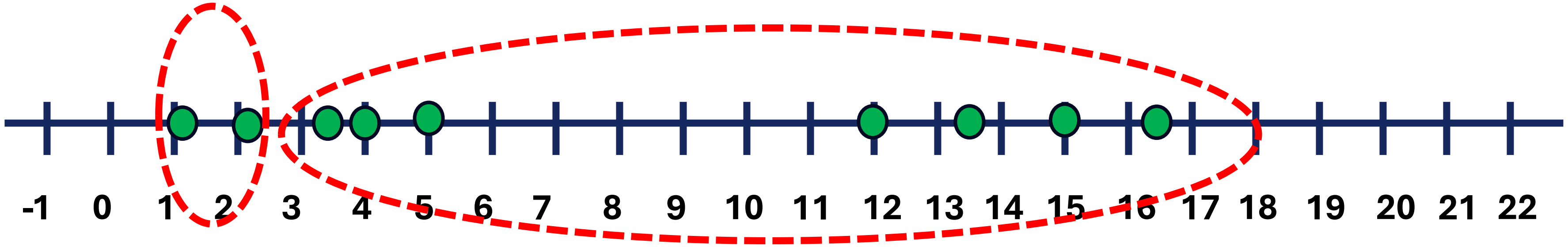
$$SC = \frac{1}{n} \sum_{i=1}^{n} s_i$$

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j}$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$

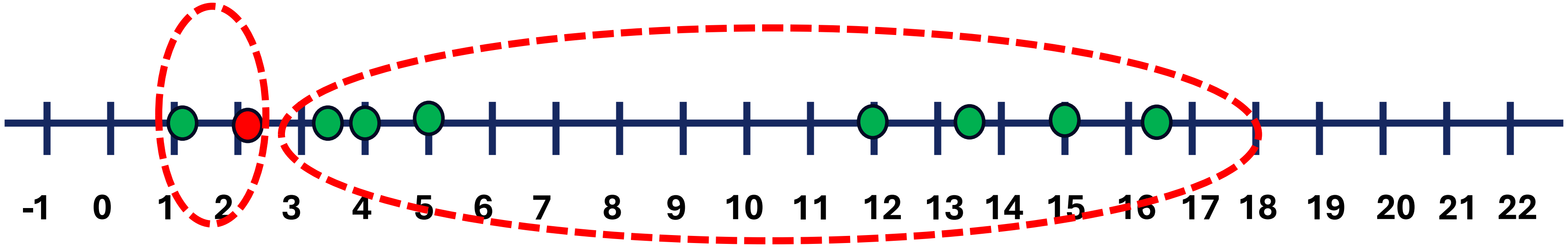


$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \qquad \mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}} \qquad \mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right)$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \left\{\{x_2, \textcolor{red}{x_5}\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\right\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

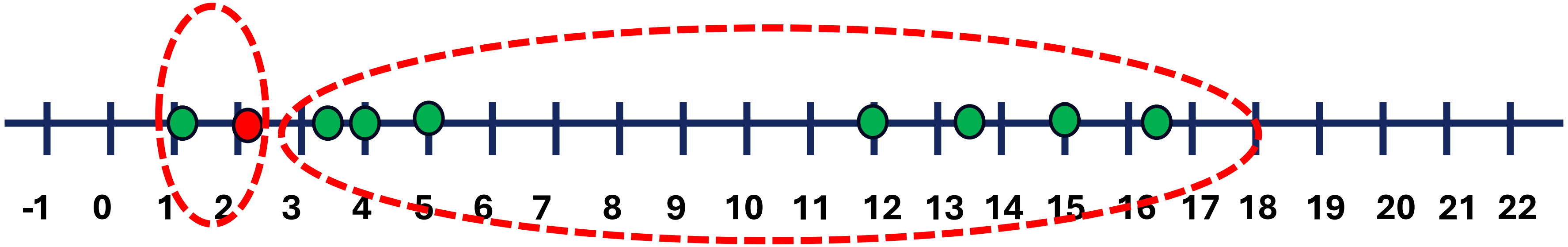
$$\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}}$$

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right)$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

$$\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}}$$

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right)$$

$$s_5 = \frac{\mu_{out}^{min}(x_5) - \mu_{in}(x_5)}{\max\{\mu_{out}^{min}(x_5), \mu_{in}(x_5)\}}$$

$$\mu_{out}^{min}(x_5) = \frac{\delta(x_5, x_1) + \delta(x_5, x_3) + \delta(x_5, x_4) + \delta(x_5, x_6) + \delta(x_5, x_7) + \delta(x_5, x_8) + \delta(x_5, x_9)}{7}$$

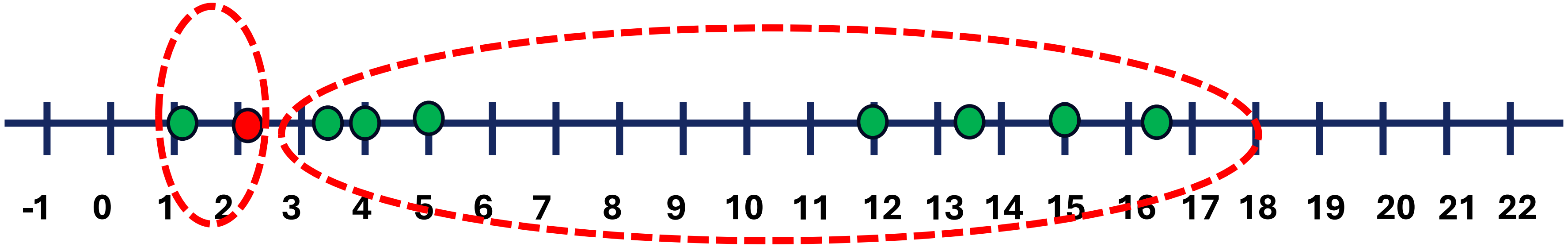
$$\mu_{out}^{min}(x_5) = \frac{1.7 + 9.7 + 14.1 + 2.7 + 12.7 + 11.4 + 1.2}{7} = 7.64$$

MONTANA STATE UNIVERSITY

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

$$\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}}$$

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right)$$

$$\mu_{out}^{min}(x_5) = 7.64$$

$$s_5 = \frac{\mu_{out}^{min}(x_5) - \mu_{in}(x_5)}{\max\{\mu_{out}^{min}(x_5), \mu_{in}(x_5)\}}$$

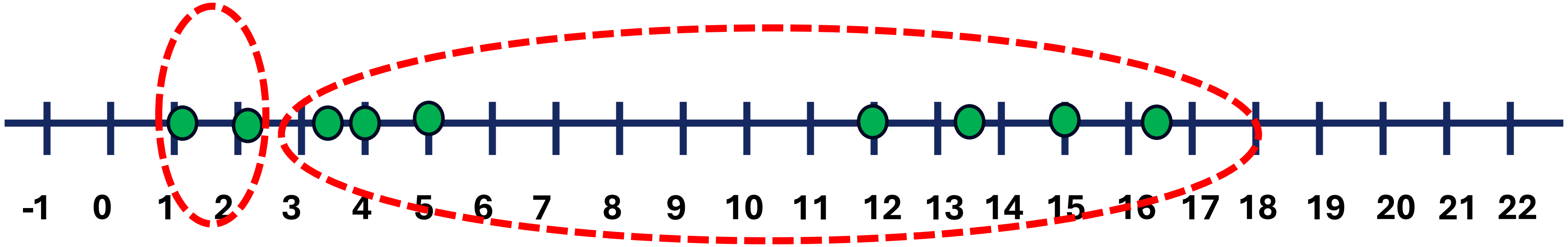
$$\mu_{in}(x_5) = \frac{\sum_{x_j \in C_1, j \neq i} \delta(x_5, x_j)}{n_1 - 1} = \frac{\delta(x_5, x_2)}{1} = 1.2$$

$$s_5 = \frac{7.64 - 1.2}{\max\{7.64, 1.2\}} = \frac{6.44}{7.64} = 0.84$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

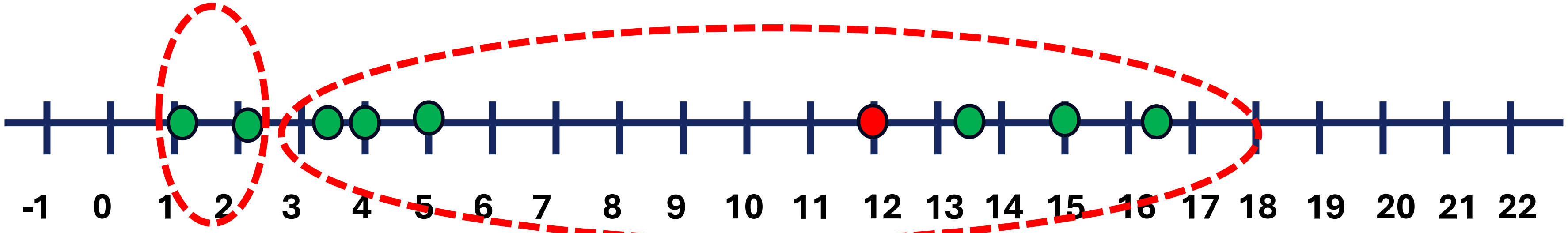
$$\mu_{in}(x_i) = \frac{\sum_{x_j \in C_{\hat{y}_i}, j \neq i} \delta(x_i, x_j)}{n_{\hat{y}_i - 1}}$$

$$\mu_{out}^{min}(x_i) = \min_{j \neq \hat{y}_i} \left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j} \right)$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

$$\mu_{out}^{min}(x_3) = \min_{j \neq \hat{y}_i}\left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j}\right) = \frac{\delta(x_3, x_2) + \delta(x_3, x_5)}{n_1} = \frac{10.9 + 9.7}{2} = 10.3$$

$$\mu_{in}(x_3) = \frac{\sum_{x_j \in C_2, j \neq i} \delta(x_3, x_j)}{n_2 - 1} = \frac{\delta(x_3, x_1) + \delta(x_3, x_4) + \delta(x_3, x_6) + \delta(x_3, x_7) + \delta(x_3, x_8) + \delta(x_3, x_9)}{6}$$

$$\mu_{in}(x_3) = \frac{8 + 4.4 + 7 + 3 + 1.7 + 8.5}{6} = 5.43$$

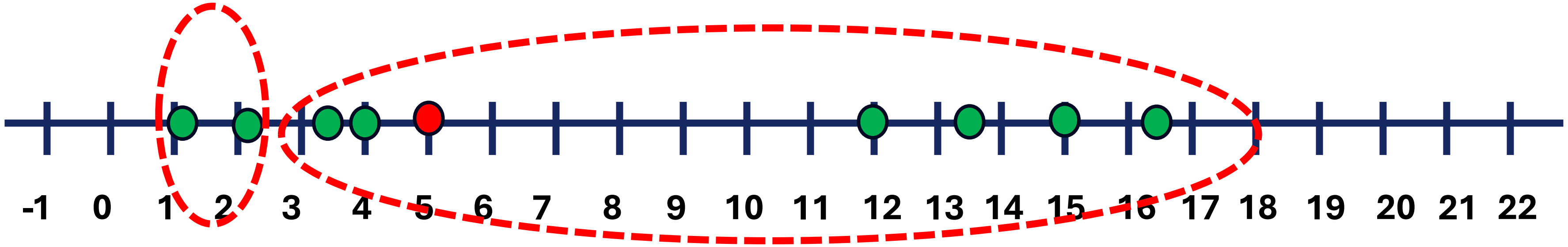
$$s_3 = \frac{\mu_{out}^{min}(x_3) - \mu_{in}(x_3)}{\max\{\mu_{out}^{min}(x_3), \mu_{in}(x_3)\}} = \frac{10.3 - 5.43}{10.3} = 0.47$$

MONTANA STATE UNIVERSITY

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

$$\mu_{out}^{min}(x_6) = \min_{j \neq \hat{y}_i}\left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j}\right) = \frac{\delta(x_6, x_2) + \delta(x_6, x_5)}{n_1} = \frac{3.9 + 2.7}{2} = 3.3$$

$$\mu_{in}(x_3) = \frac{\sum_{x_j \in C_2, j \neq i} \delta(x_6, x_j)}{n_2 - 1} = \frac{\delta(x_6, x_1) + \delta(x_3, x_6) + \delta(x_6, x_4) + \delta(x_6, x_7) + \delta(x_6, x_8) + \delta(x_6, x_9)}{6}$$

$$\mu_{in}(x_3) = \frac{1 + 7 + 11.4 + 10 + 8.7 + 1.5}{6} = 6.6$$

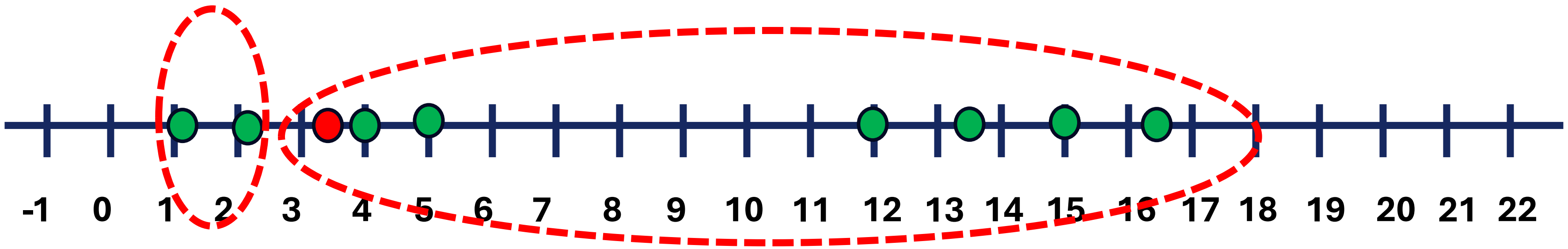
$$s_3 = \frac{\mu_{out}^{min}(x_6) - \mu_{in}(x_6)}{\max\{\mu_{out}^{min}(x_6), \mu_{in}(x_6)\}} = \frac{3.3 - 6.6}{6.6} = -0.5$$

MONTANA STATE UNIVERSITY

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}}$$

$$\mu_{out}^{min}(x_1) = \min_{j \neq \hat{y}_i}\left(\frac{\sum_{y \in C_j} \delta(x_i, y)}{n_j}\right) = \frac{\delta(x_9, x_2) + \delta(x_9, x_5)}{n_1} = \frac{2.4 + 1.2}{2} = 1.8$$

$$\mu_{in}(x_3) = \frac{\sum_{x_j \in C_2, j \neq i} \delta(x_6, x_j)}{n_2 - 1} = \frac{\delta(x_9, x_1) + \delta(x_9, x_3) + \delta(x_9, x_4) + \delta(x_9, x_6) + \delta(x_9, x_7) + \delta(x_9, x_8)}{6}$$

$$\mu_{in}(x_3) = \frac{0.5 + 8.5 + 12.9 + 1.5 + 11.5 + 10.2}{6} = 7.51$$

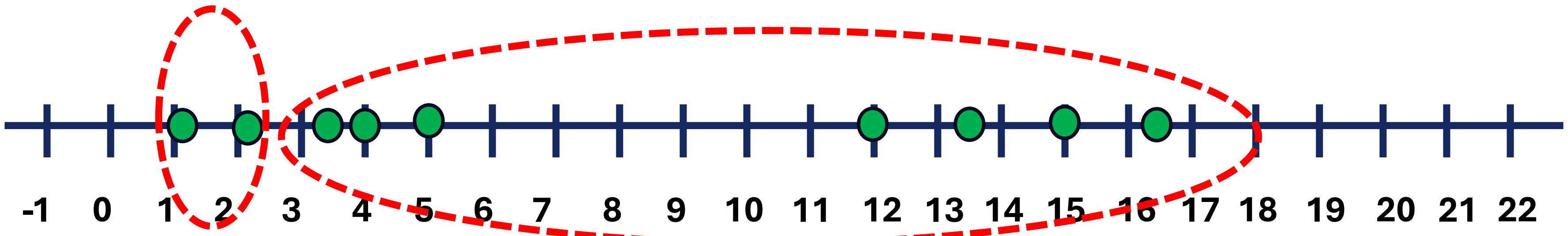
$$s_3 = \frac{\mu_{out}^{min}(x_6) - \mu_{in}(x_6)}{\max\{\mu_{out}^{min}(x_6), \mu_{in}(x_6)\}} = \frac{1.8 - 7.51}{7.51} = -0.76$$

MONTANA STATE UNIVERSITY

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_2, x_5\}, \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}\}$$



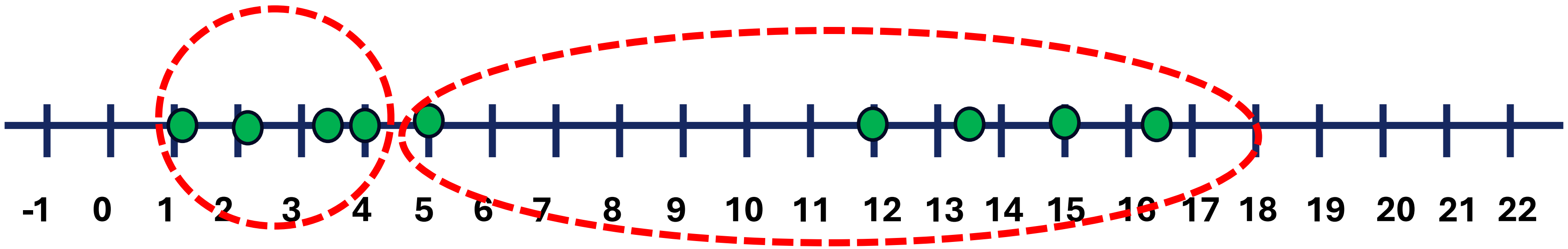
$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \qquad SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

$$In\ this\ example, SC = 0.2$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_1, x_2, x_5, x_9\}, \{x_3, x_4, x_6, x_7, x_8\}\}$$



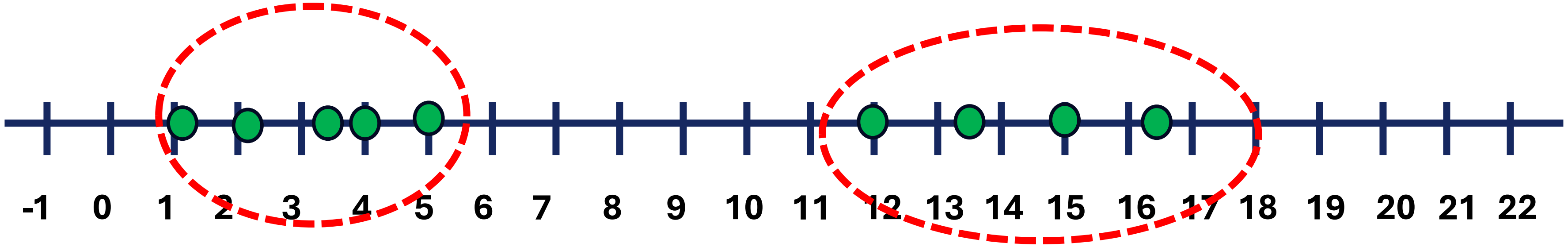
$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \qquad SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

$$In\ this\ example, SC = 0.57$$

Silhouette coefficient (Ground truth is not known)

	X_1
x_1	4
x_2	1.1
x_3	12
x_4	16.4
x_5	2.3
x_6	5
x_7	15
x_8	13.7
x_9	3.5

$$\mathcal{C} = \{C_1, C_2\} = \{\{x_1, x_2, x_5, x_6, x_9\}, \{x_3, x_4, x_7, x_8\}\}$$



$$s_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{\max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)\}} \qquad SC = \frac{1}{n}\sum_{i=1}^{n} s_i$$

$$In\ this\ example, SC = 0.80$$

Question:

- Can you think of an example dataset and a clustering where silhouette coefficient is 1 or very close to 1?
  - $D = \{(0,0), (0.01, 0.02), (100,100), (100.1, 100.2)\}$
  - $\mathcal{C} = \{C_1, C_2\}, where,$
  - $C_1 = \{(0,0), (0.01, 0.02)\}, C_2 = \{(100,100), (100.1, 100.2)\}$

- Can you think of an example dataset and a clustering where silhouette coefficient is -1?
  - $C_1 = \{(0,0), (100.01, 100.02)\}, C_2 = \{(0.01,0.02), (100, 100)\}$