Week 9: Clustering

กิตติรักษ์ ม่วงมิ่งสุข (กิ๊)

Last Modify: Jan 28, 2026

Clustering

- คือการนำข้อมูลที่ "ไม่มีเฉลย" (No Label) มาให้คอมพิวเตอร์หาความเหมือนและความต่างเอง
- Unsupervised Learning
- K-Means clustering
  - K คือจำนวนกลุ่มที่อยากให้แบ่ง (เราเป็นคนกำหนด)
  - การทำงาน:
    - 1. สุ่มจุดศูนย์กลาง (Centroid) ขึ้นมา k จุด
    - 2. “วัดระยะห่าง” ข้อมูลจุดไหนอยู่ใกล้ Centroid ไหนที่สุด ให้ไปเข้ากลุ่มนั้น
    - 3. คำนวณหาค่าเฉลี่ยของกลุ่มเพื่อขยับ Centroid ไปที่ตำแหน่งใหม่
    - 4. ทำซ้ำจนกว่าจุด Centroid จะนิ่ง หรือ ครบจำนวนรอบที่กำหนด

K-Means on iris data

Iris setosa

Iris versicolor

Iris virginica

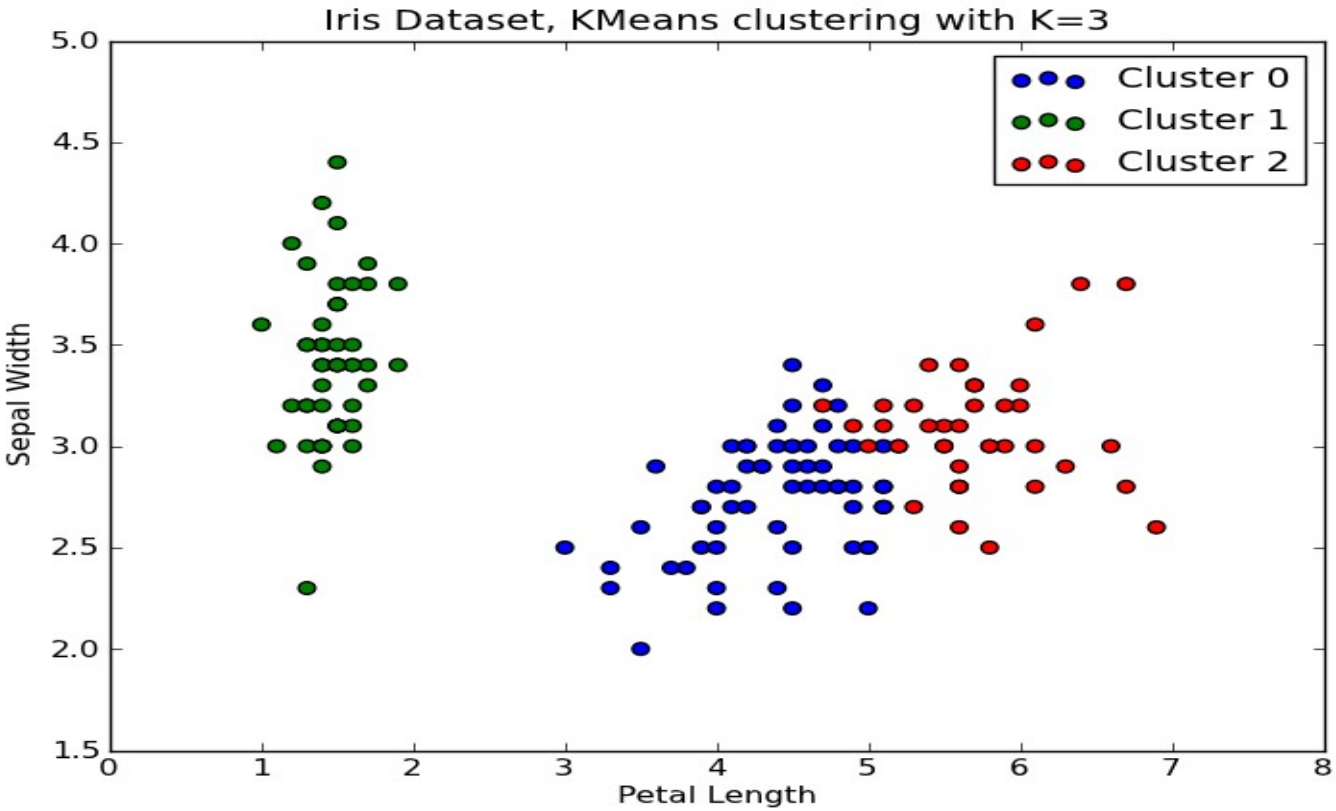
Ref: https://en.wikipedia.org/wiki/Iris_flower_data_set

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*

Iris data: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

SparkML kmean app:
https://github.com/apache/spark/tree/branch-1.5/examples/src/main/python/mllib

Ex. Iris K-means



Source: http://stackoverflow.com/questions/6645895/calculating-the-percentage-of-variance-measure-for-k-means

ระยะห่างแบบเส้นตรง

Euclidean distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d(x,y) = \sqrt{(5.1-4.9)^2 + (3.5-3.0)^2 + (1.4-1.4)^2 + (0.2-0.2)^2}$$

$$d(x,y) = 0.538$$

Unknow Class = [5.1 3.5 1.4 0.2]

Iris-setosa = [4.9 3.0 1.4 0.2]

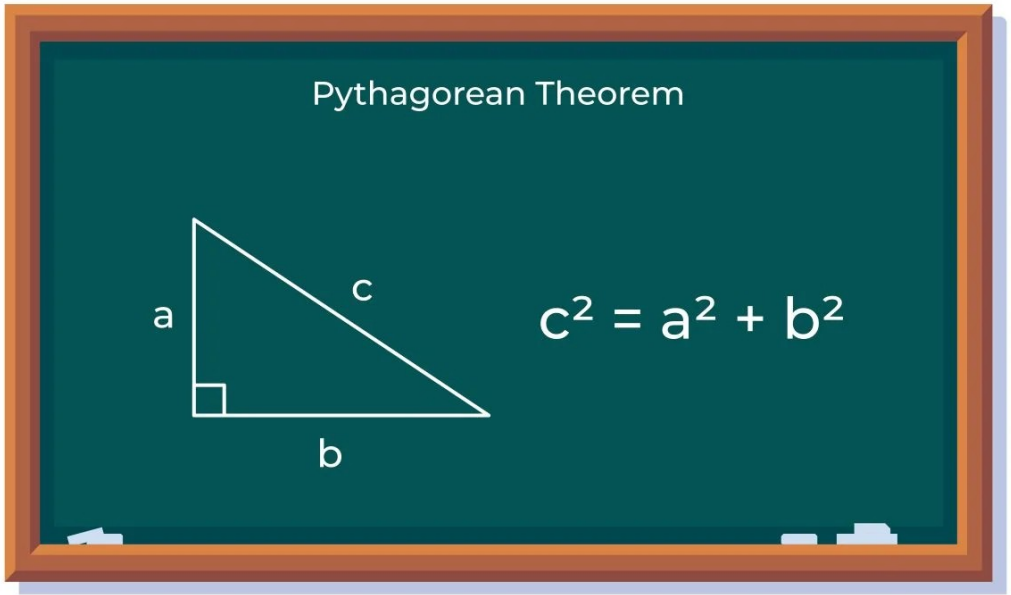
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

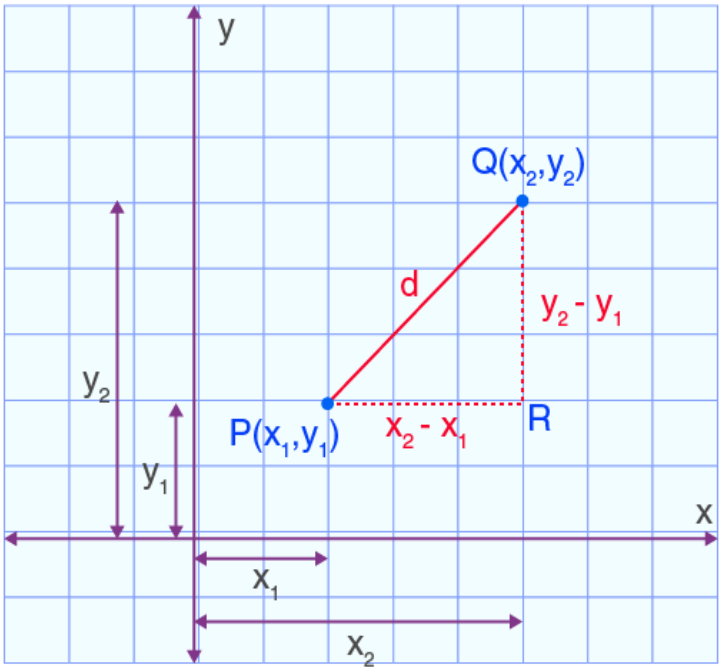
$$d(x,y) = \sqrt{(5.1-6.4)^2 + (3.5-3.2)^2 + (1.4-4.5)^2 + (0.2-1.5)^2}$$

$$d(x,y) = 3.6166$$

Iris-versicolor = [6.4 3.2 4.5 1.5]

พีทาโกรัส และ ยุคลิด

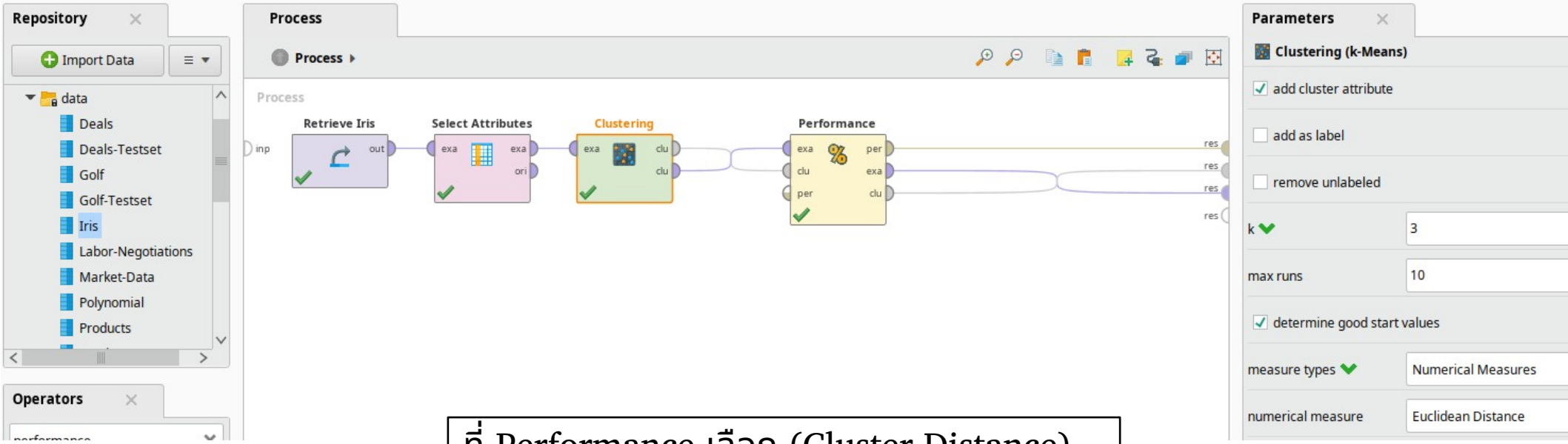




รูปแรกมาจาก: https://www.webythebrain.com/article/pythagoras-theorum-summary
รูปที่สองมาจาก https://byjus.com/maths/euclidean-distance/

Lab9-1 K-Mean with Iris data

- ใช้ข้อมูลตัวอย่าง Iris ทำ K-Means Clustering



ที่ Performance เลือก (Cluster Distance)

การวัดผล

- Davies-Bouldin Index: เป็นค่าที่บอกว่า กลุ่มแต่ละกลุ่มอยู่ห่างกันแค่ไหน และภายในกลุ่มเกาะกันแน่นไหม
    - ยิ่งค่าน้อย ยิ่งดี: หมายความว่าในกลุ่มเกาะกันแน่น และแต่ละกลุ่มแยกจากกันชัดเจน
- Elbow Method: การรันค่า k หลายๆ ค่า (เช่น 1-10) เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด

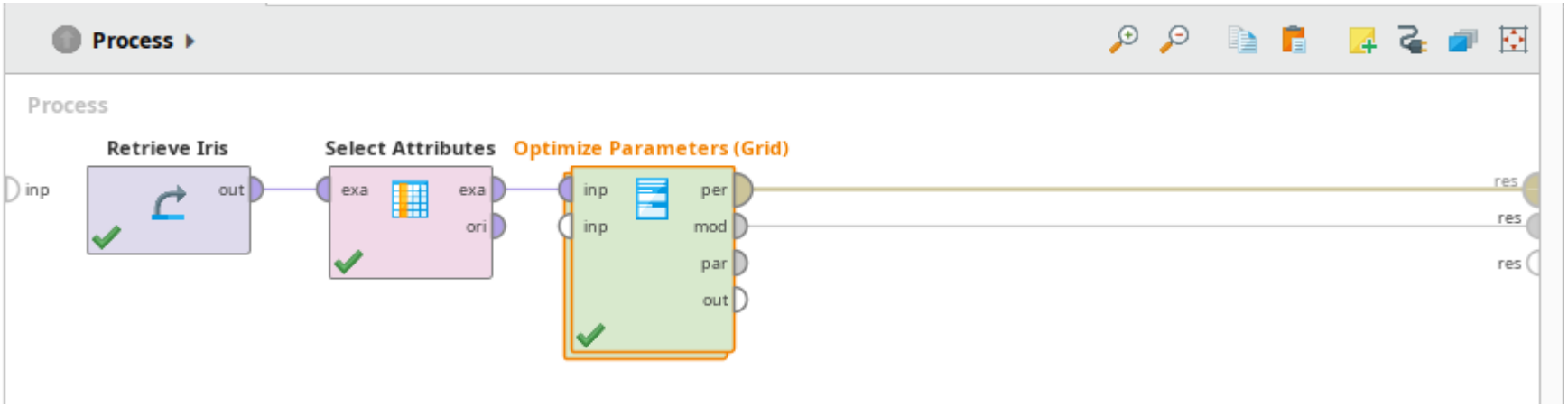
การนำไปใช้ในธุรกิจ (Business Use Case)

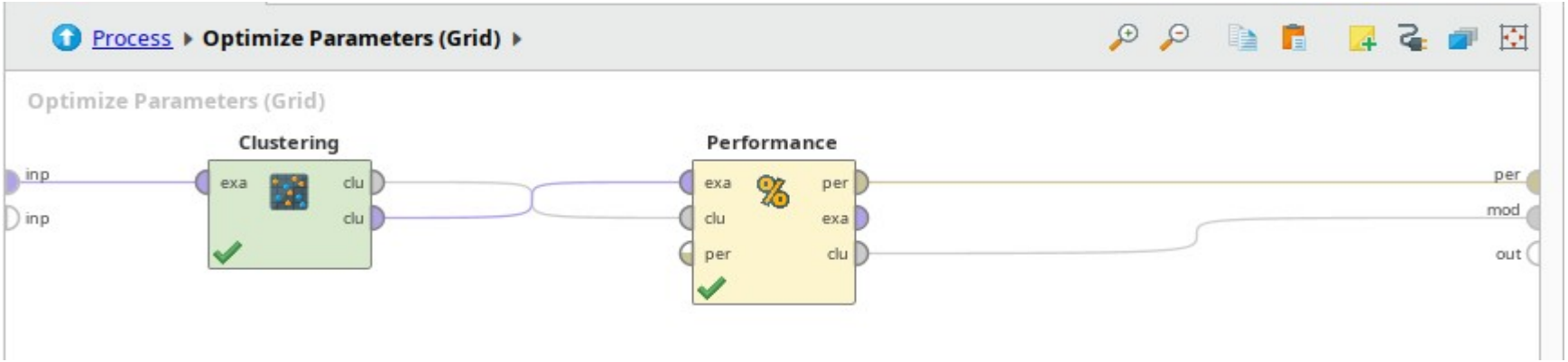
- **Customer Segmentation**: แบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อ (เช่น กลุ่มประหยัด, กลุ่มเปย์หนัก, กลุ่มขาประจำ)
- **Anomaly Detection**: ข้อมูลไหนที่ไม่เข้าพวกกับกลุ่มใดเลย อาจจะเป็น "การทุจริต" (Fraud)
- **Inventory Management**: จัดกลุ่มสินค้าที่มักจะถูกซื้อพร้อมกันหรือมีรอบการหมุนเวียนคล้ายกัน

เพิ่มเติม

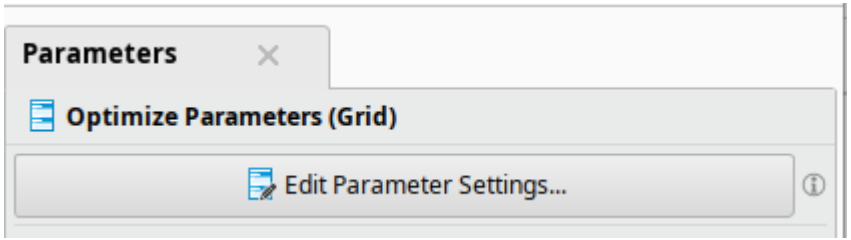
- ก่อนทำ Clustering ต้องทำ Normalize ข้อมูลเสมอ เพื่อให้ทุกตัวแปร (เช่น รายได้หลักแสน กับ อายุหลักสิบ) มีน้ำหนักเท่ากัน ไม่ให้ตัวแปรที่มีเลขเยอะๆ ข่มตัวแปรอื่น

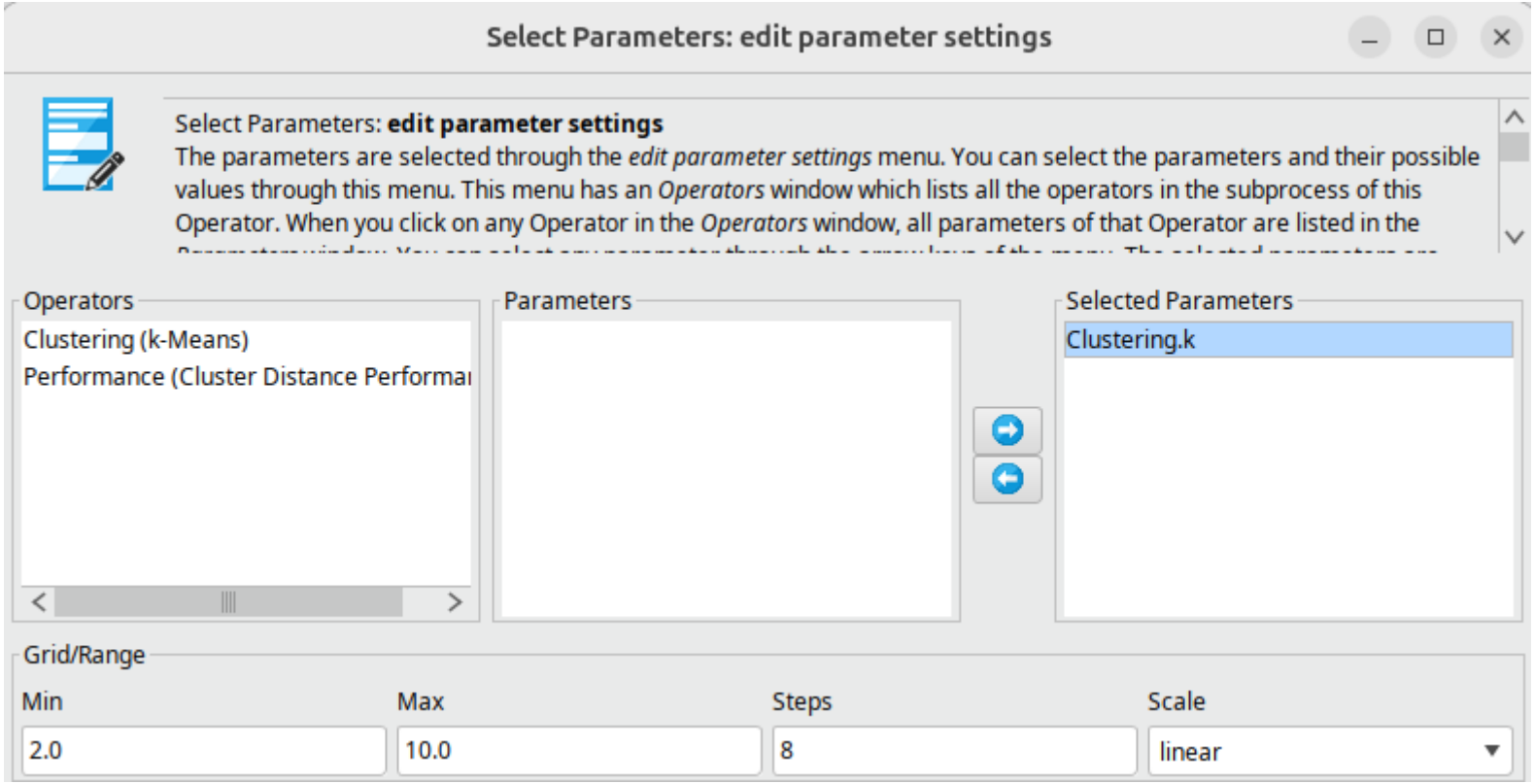
Lab9-2: Optimize Parameters





Optimize Parameters
Edit Parameter Settings





Cluster Distance Performance

- Avg. Distance (Elbow Method) (Elbow แปลว่า ข้อศอก ~ หักศอก) Elbow Method: การรันค่า k หลายๆ ค่า (เช่น 1-10) แล้วดูจุดที่กราฟความผิดพลาดเริ่มนิ่ง เหมือน "ข้อศอก" เพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด
- Davies-Bouldin Index (DBI) เป็นค่าที่บอกว่า กลุ่มแต่ละกลุ่มอยู่ห่างกันแค่ไหน และภายในกลุ่มเกาะกันแน่นไหม

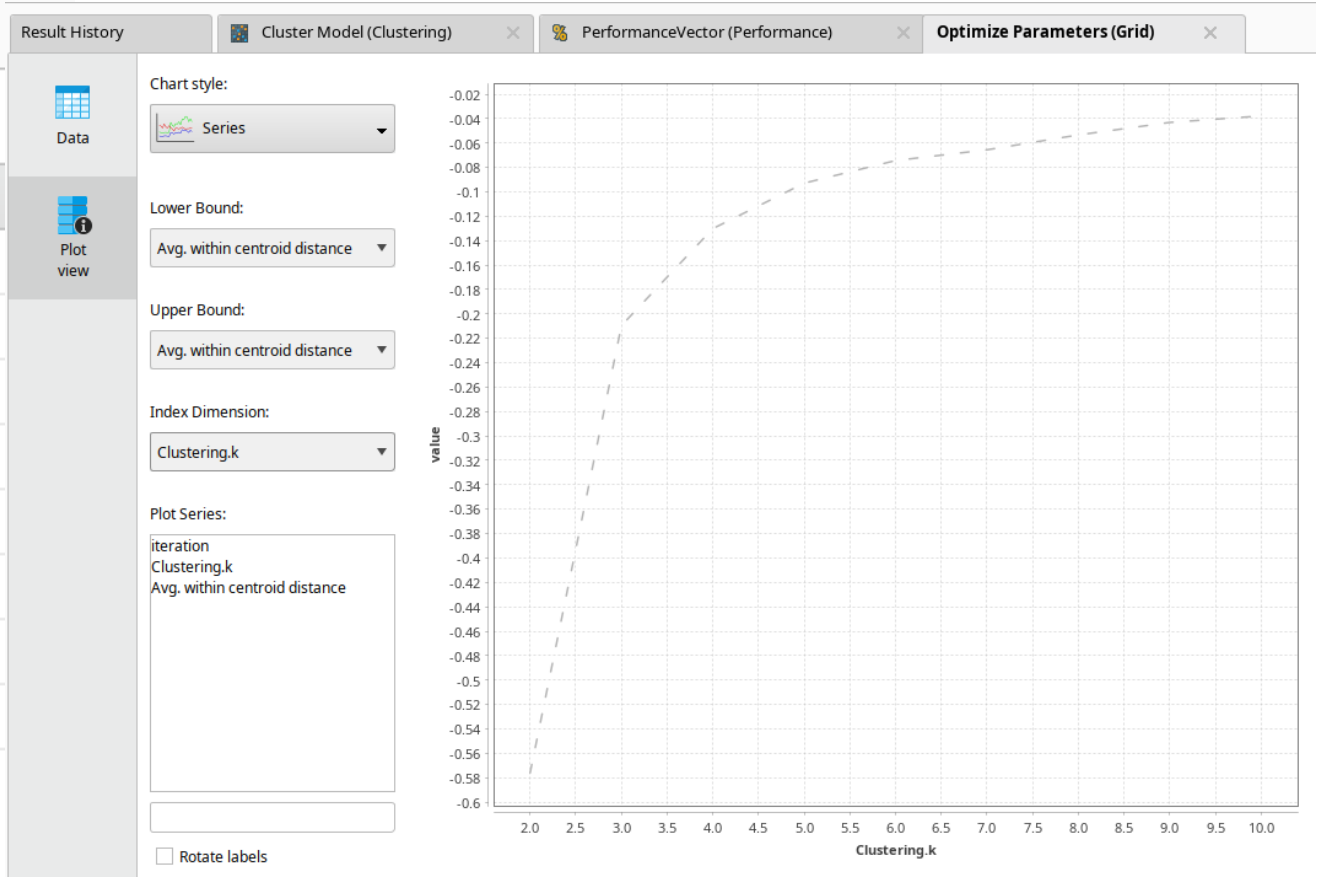
Avg. Distance (Elbow Method)

(Elbow แปลว่า ข้อศอก ~ หักศอก)

Optimize Parameters (Grid)

Optimize Parameters (Grid) (9 rows, 3 columns)

iteration	Cluster...	Avg. within centroid distance
1	2	-0.576
2	3	-0.210
3	4	-0.130
4	5	-0.093
5	6	-0.074
6	7	-0.066
7	8	-0.053
8	9	-0.043
9	10	-0.038



Davies-Bouldin Index (DBI)

Optimize Parameters (Grid) (9 rows, 3 columns)

iteration	Clustering.k	Davies Bouldin
1	2	-0.265
2	3	-0.486
3	4	-0.561
4	5	-0.626
5	6	-0.689
6	7	-0.648
7	8	-0.628
8	9	-0.637
9	10	-0.741

ดูค่าบวกที่น้อยที่สุด
สำหรับ Rapidminer ดูค่าที่ใกล้ 0

ClusterKit

Q & A