Clustering (Unsupervised learning)

Oliver W. Layton

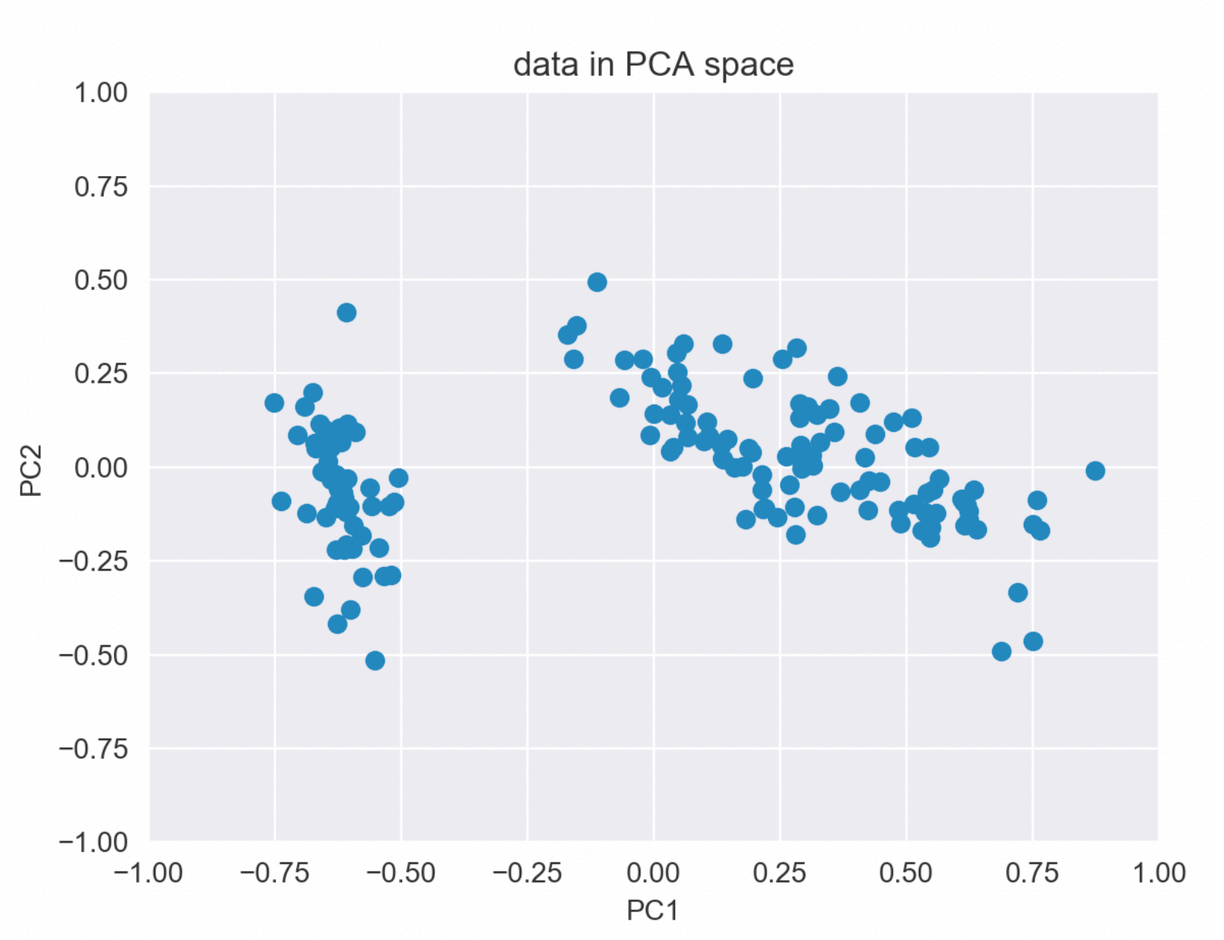
CS251: Data analysis and visualization

Lecture 21, Spring 2021

Wednesday March 31

Clustering (Unsupervised learning)

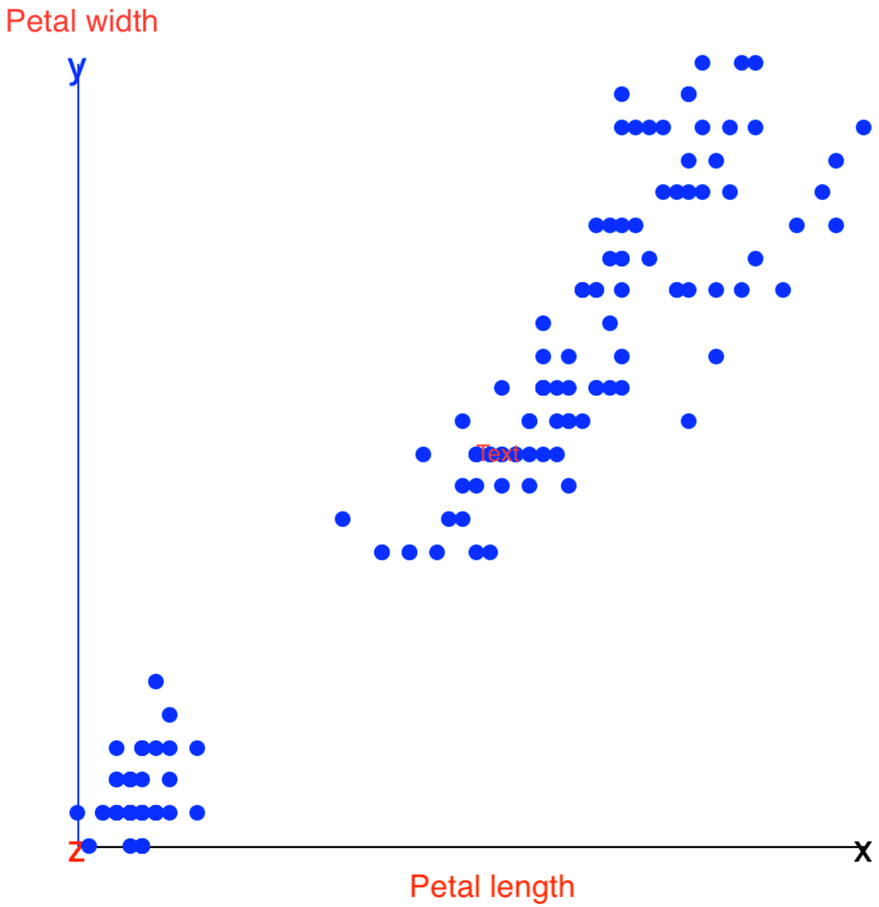
Iris data in PCA space



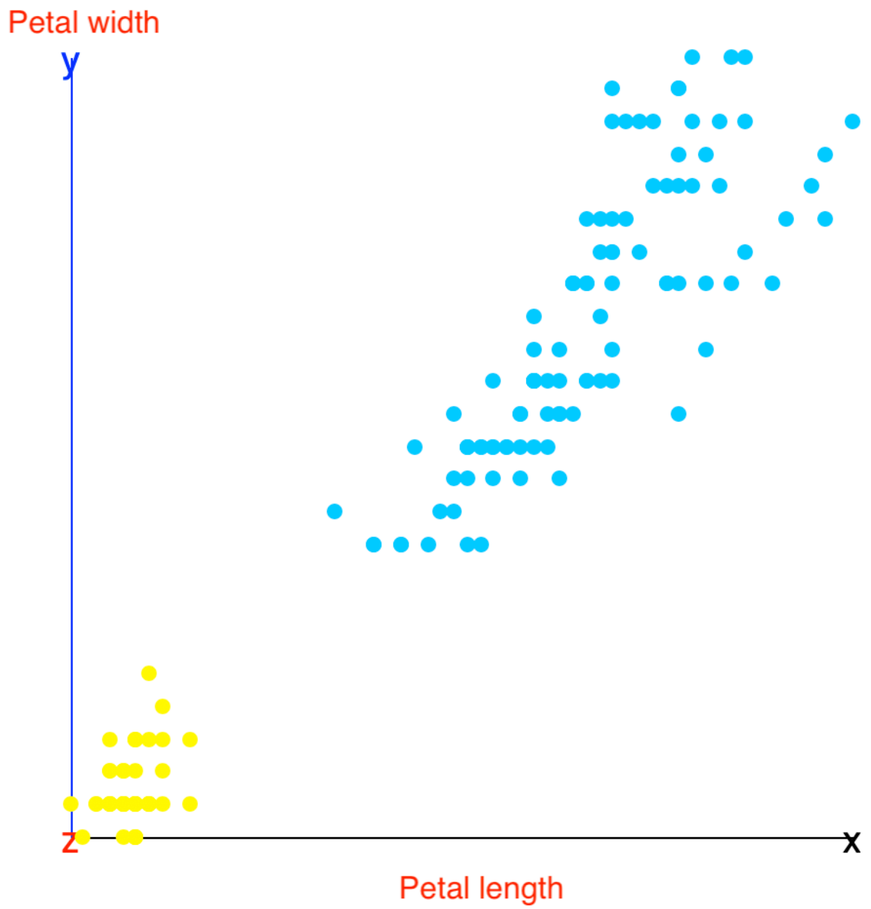
Identifying clusters in data

- *Goal*: Assign each data point to "groups" based on similarity to other data points (**clusters**).
  - Grouping means "tagging" or "coloring" each point. Like adding another feature (dimension) with an int that represents group membership (e.g. group 1, group 2, etc.).
- Algorithms differ in:
  - how **similarity** defined.
  - whether group membership is exclusive (1 group per point) or **fuzzy** (point belongs to multiple groups, to varying degrees).
  - how many groups/clusters to use.
  - whether grouping happens **top-down** (big picture first) or **bottom-up** (data point level).

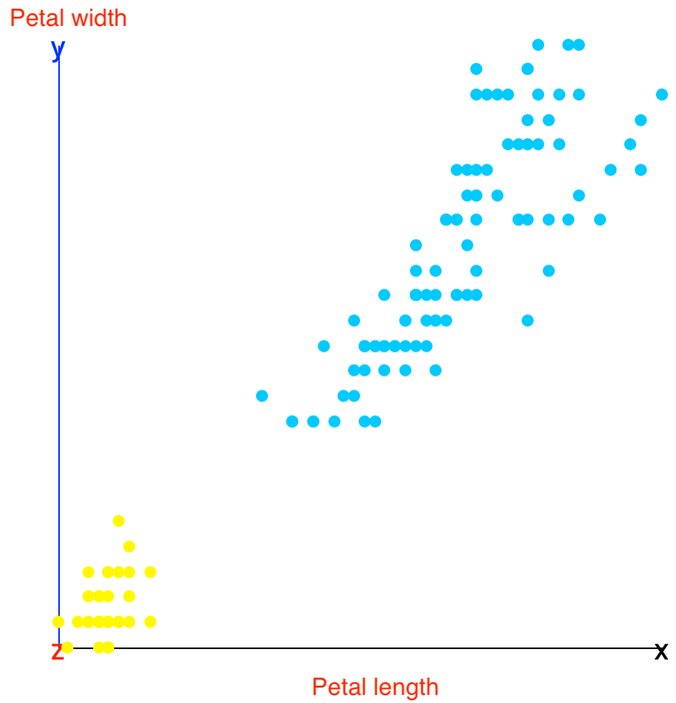
Input: Iris data



Output: Iris data assigned to 2 clusters



Iris data assigned to 2 clusters



- How do we define similarity (closeness) between two samples affects the clusters that get.
- Often controlled by **distance metric** — how do we measure distance between any 2 samples?

Distance metric (1/2)

- What's the most common way to measure distance?
  - Euclidean distance. But there are other ways...
- **Distance metric:** assign a *scalar value* between 2 pts $\vec{a}$ and $\vec{b}$ that represents how far they are from one another.
- Function that yields this scalar value: $d(\vec{a}, \vec{b})$.

Let's discuss distance metric properties and examples on the board...