Clustering

March 13, 2018

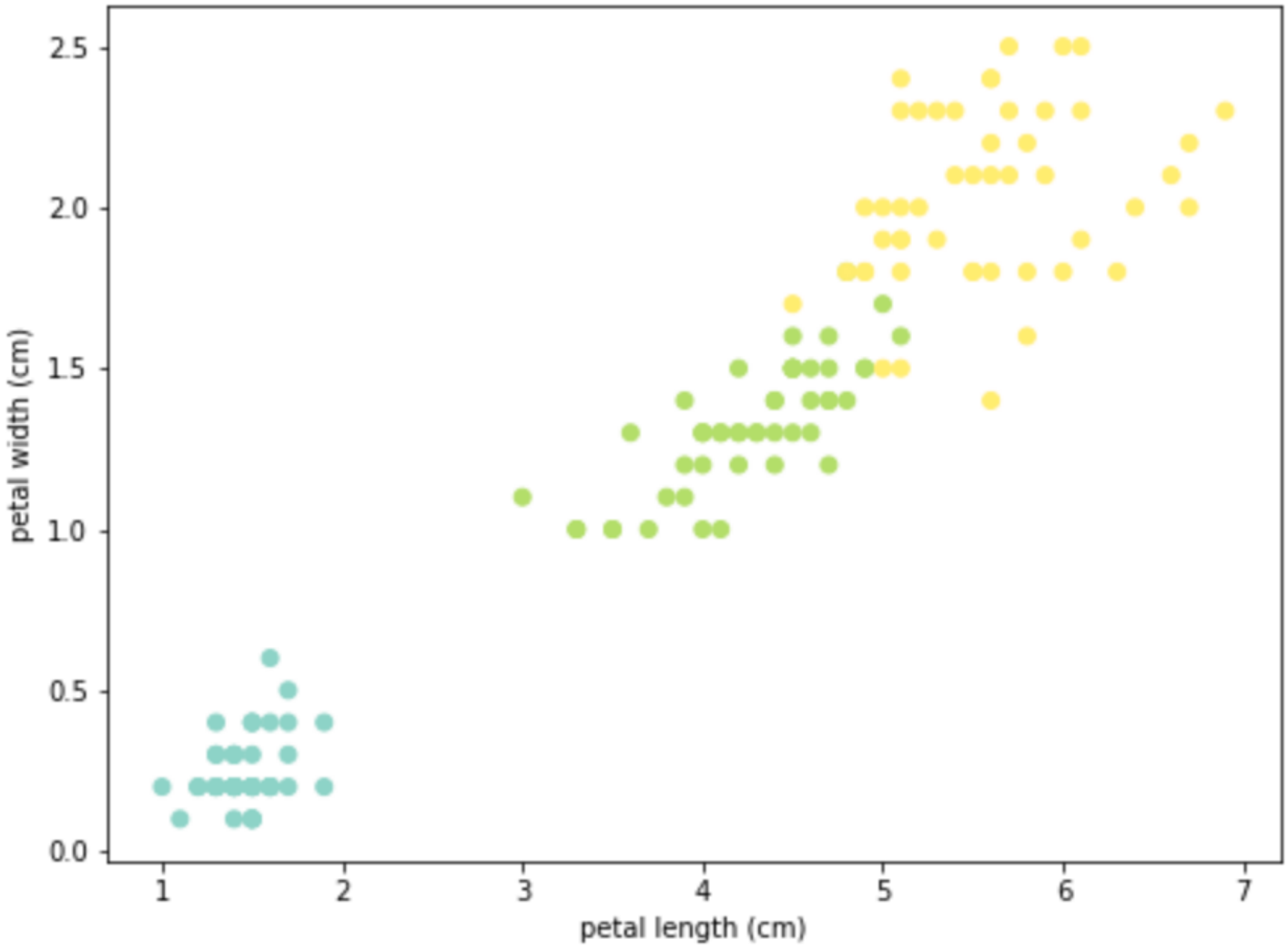
@priska

Plan for today

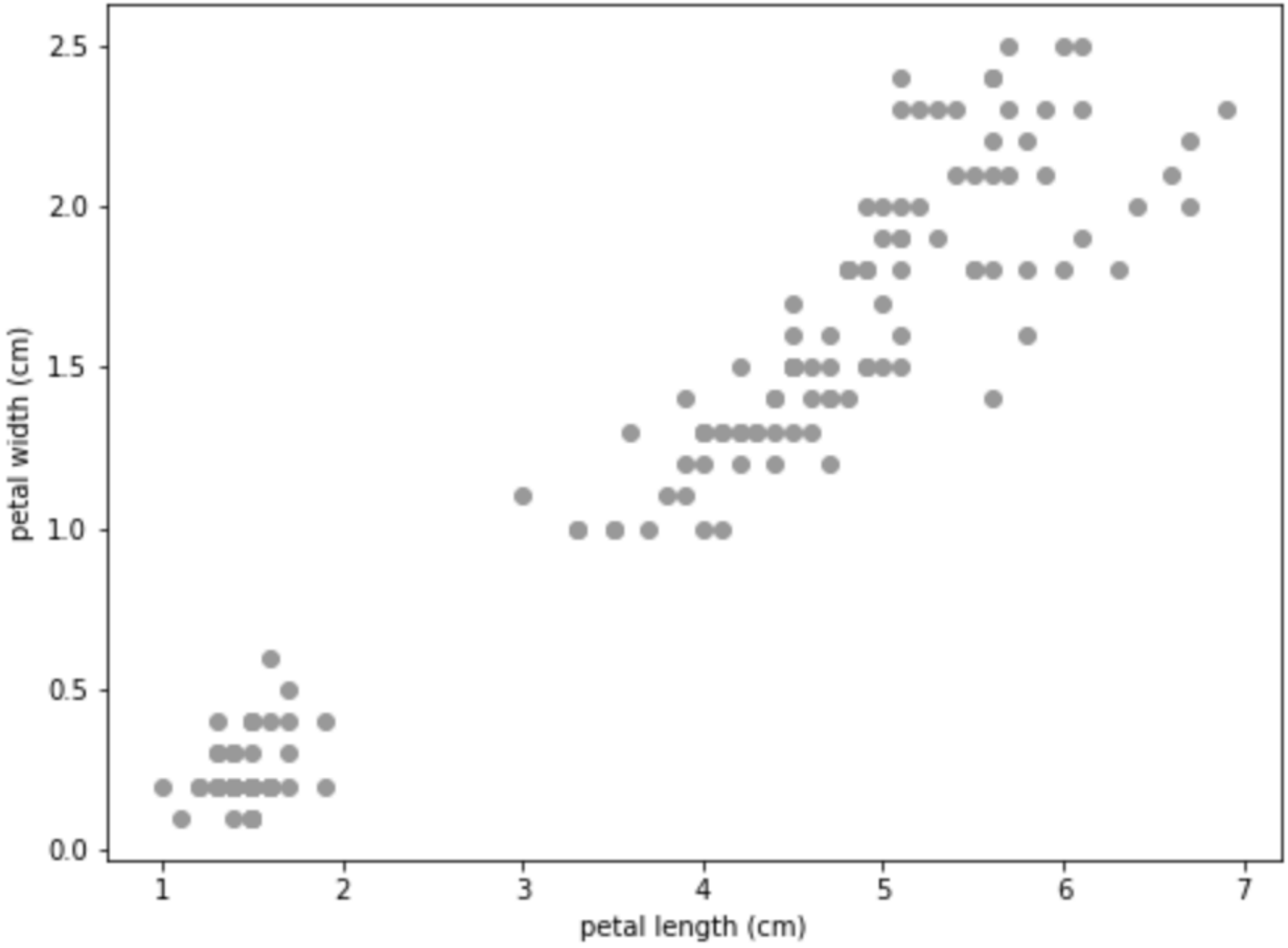
- Some slides (big picture and some references)
- Some code (practice)

(blueish text is links)

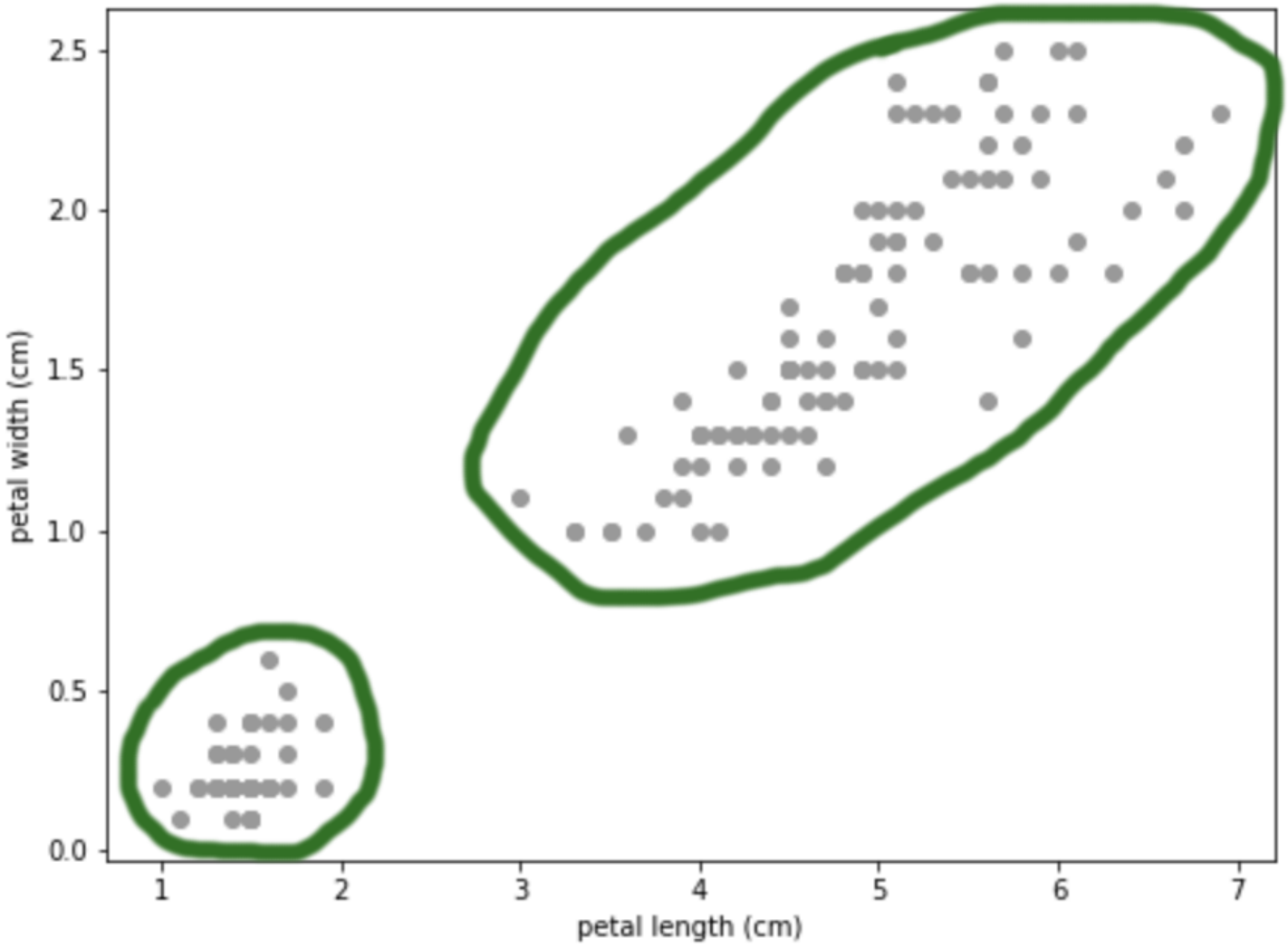
Supervised Learning



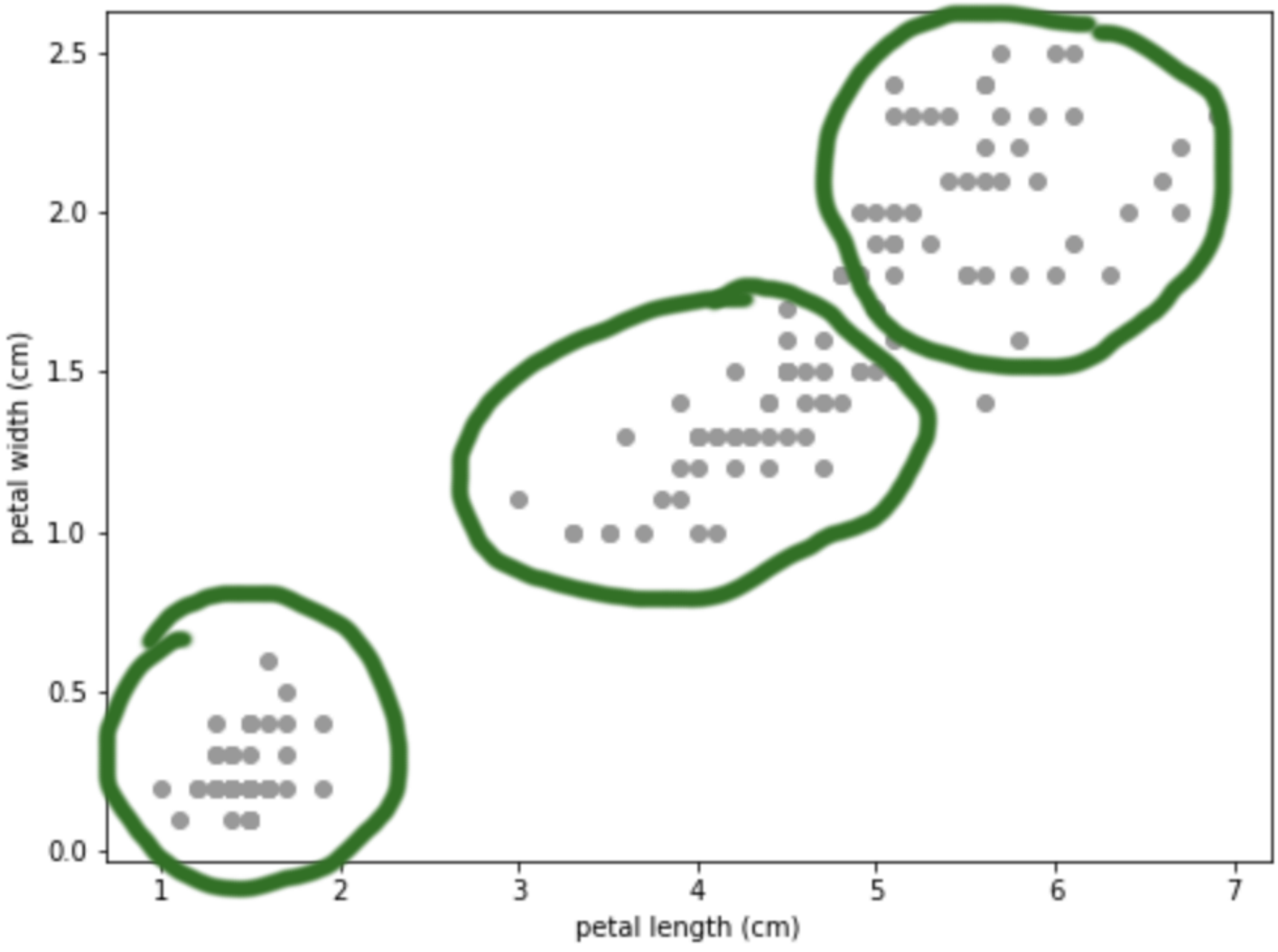
Unsupervised Learning



Clustering



Clustering



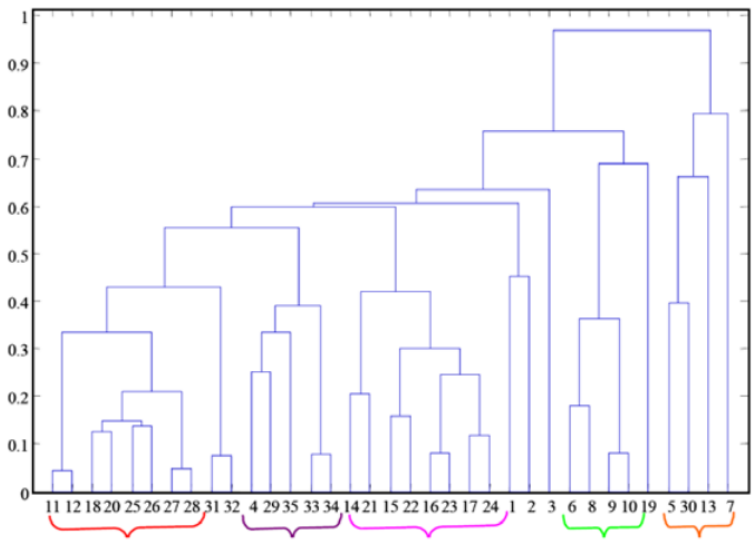
Clustering

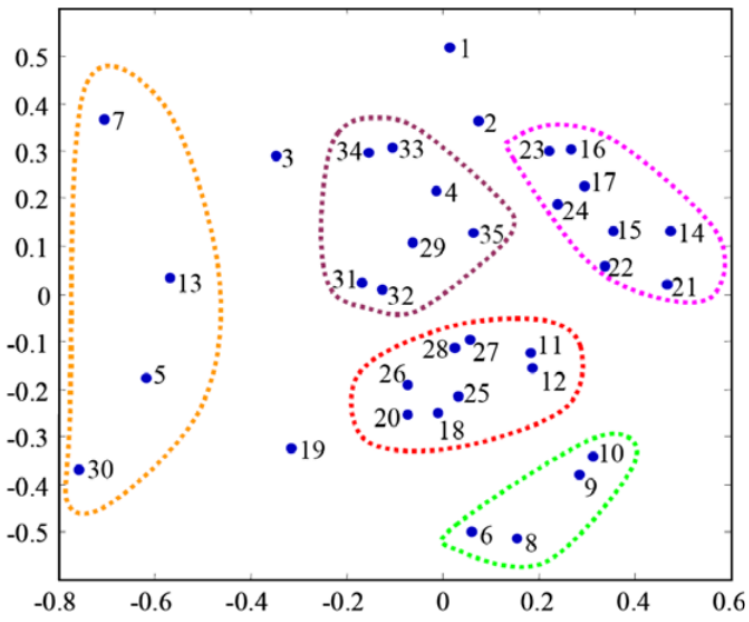
is "the process of grouping similar objects together." Murphy (2012)

But... how is similarity defined?

Clustering

- Flat clustering
- Hierarchical clustering





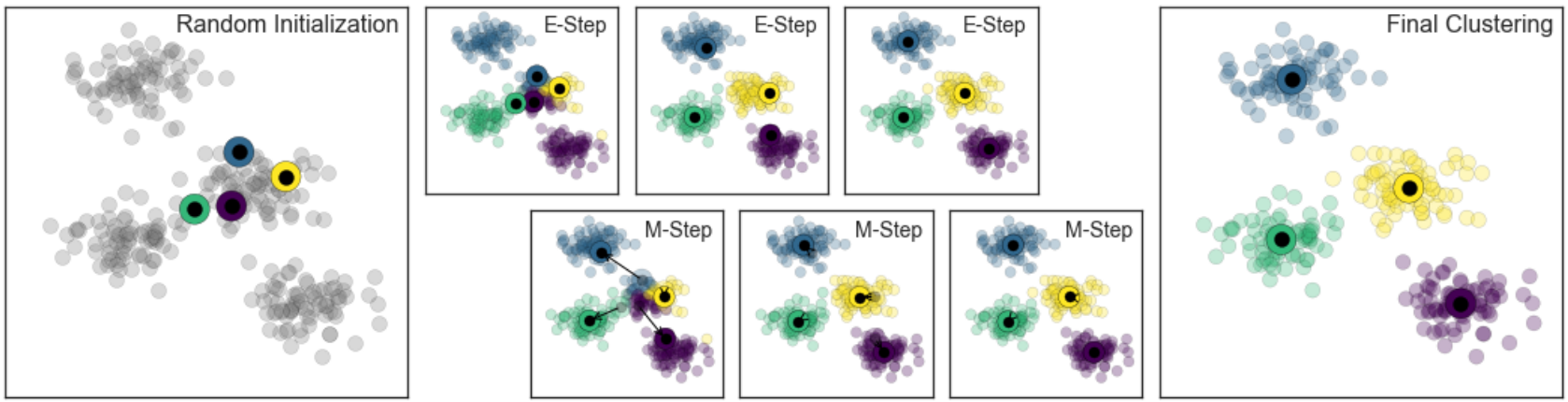
Jain (2010) ← a review paper on clustering

k-means clustering

The algorithm:

1. Choose k points/centroids randomly
2. Assign each data point to its nearest (read: most similar) centroid based on the Euclidean distance
3. Recompute the k centroids by taking the arithmetic mean of all data points assigned to them
4. Repeat steps 2 and 3 until a stopping criterion is reached

k-means clustering



from the Python Data Science Handbook by Jake VanderPlas

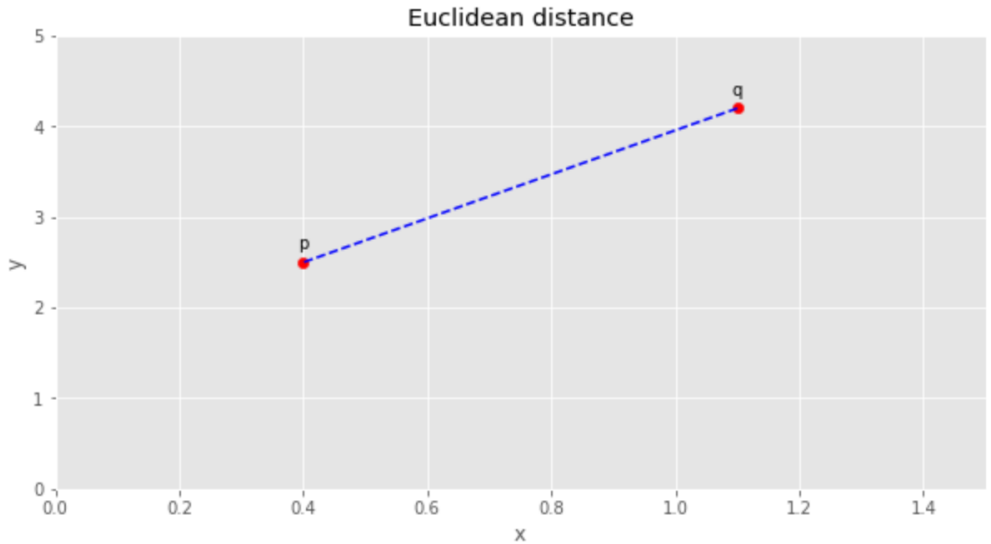
k-means clustering

Stopping criteria, e.g.:

- No data points change cluster anymore.
- The sum of within-cluster distances is minimized.
- A maximum number of iterations has been reached.

Distance measures

The Euclidean distance measures the length of the direct/straight path connecting two points in space.



$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}$$

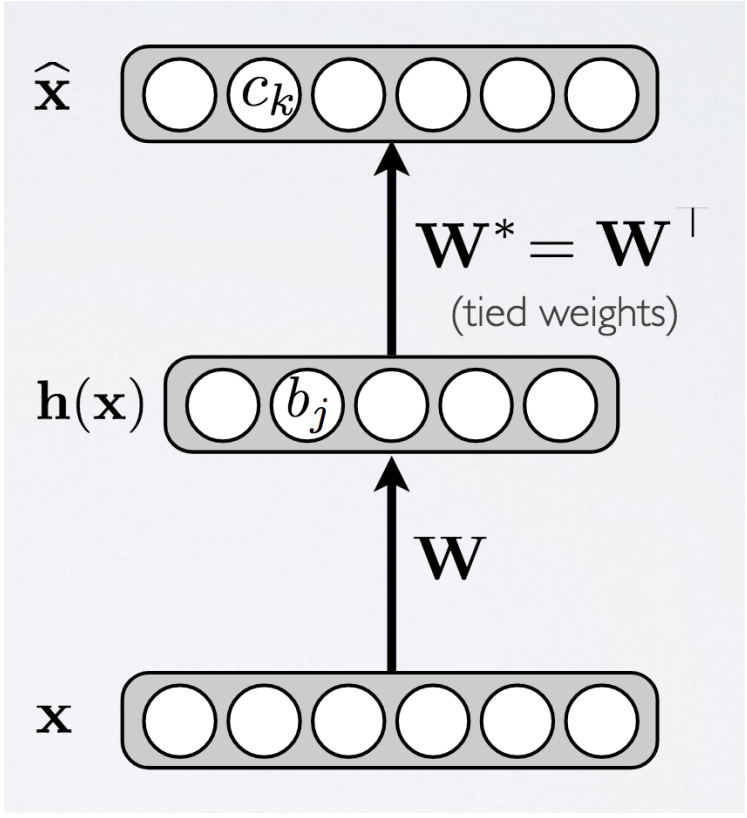
$$= \sqrt{(1.1 - 0.4)^2 + (4.2 - 2.5)^2}$$

Unsupervised Learning

Apart from clustering there are other techniques of unsupervised learning:

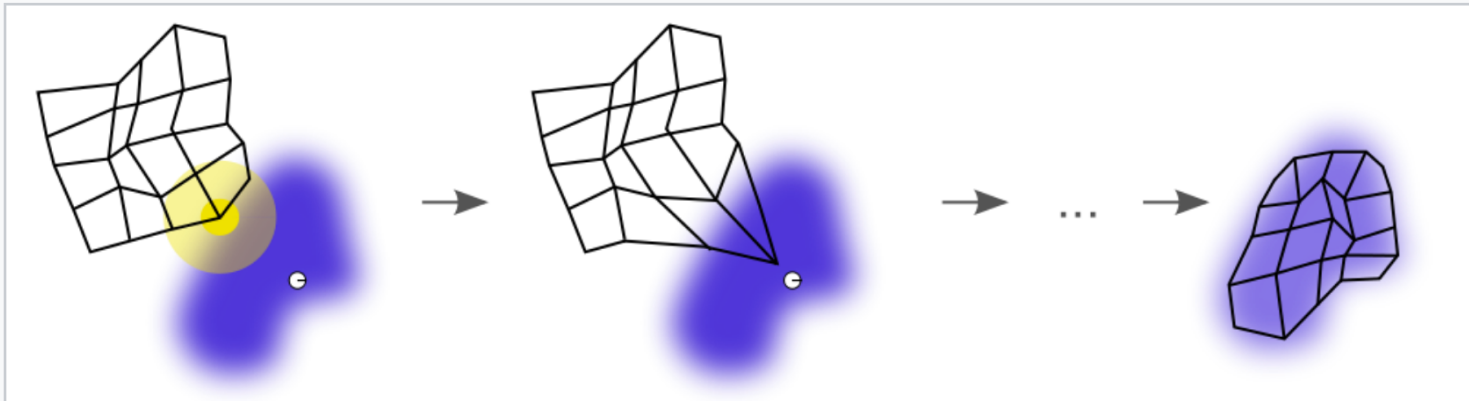
- self-organizing maps
- autoencoders
- hidden Markov models
- dimensionality reduction
- blind source separation

Autoencoders



Hugo Larochelle's course on Neural Networks

Self-organizing Maps



An illustration of the training of a self-organizing map. The blue blob is the distribution of the training data, and the small white disc is the current training datum drawn from that distribution. At first (left) the SOM nodes are arbitrarily positioned in the data space. The node (highlighted in yellow) which is nearest to the training datum is selected. It is moved towards the training datum, as (to a lesser extent) are its neighbors on the grid. After many iterations the grid tends to approximate the data distribution (right).

Wikipedia article on self-organizing maps

→ clustering.ipynb