EECE 5644: Unsupervised Clustering

Mark Zolotas

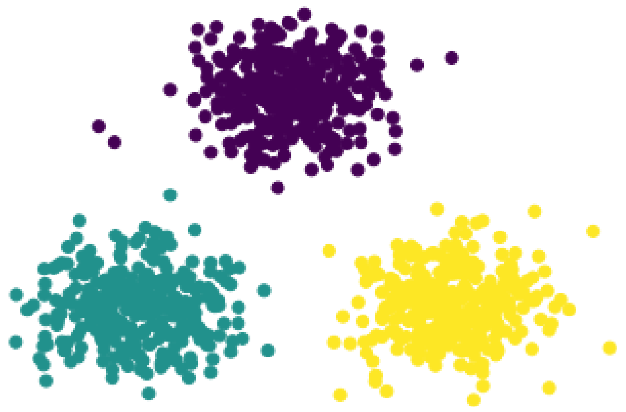
E-mail: m.zolotas@northeastern.edu
Webpage: https://coe.northeastern.edu/people/zolotas-mark/

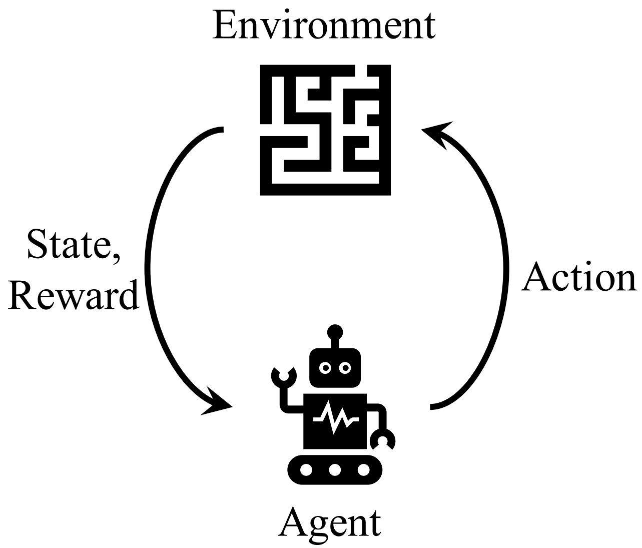
Tentative Course Outline (Wks. 5-6*)

Topics	Dates	Assignments	Additional Reading
~~Neural Networks: Multilayer Perceptrons & Backpropagation~~	~~08/01-03~~	**Homework 3** released on Canvas on 08/01 **Due 08/10**	~~Chpts. 13.1-13.5 Murphy 2022~~
~~*HW1 Review*~~	~~08/02~~		~~N/A~~
Clustering: K-means, Gaussian Mixture Models (GMMs)	08/04		Chpt. 21 Murphy 2022
Support Vector Machine (SVM)	08/08	**Homework 4** released on Canvas on 08/08 **Due 08/17**	Burges Tutorial
Reinforcement Learning	08/09		N/A

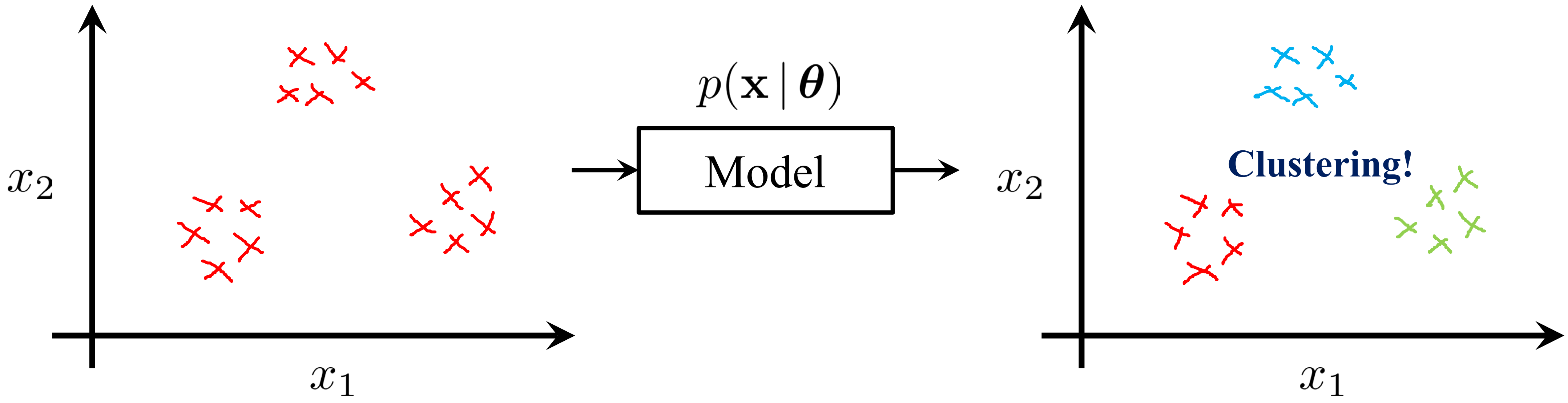
Types of Machine Learning

- **Supervised Learning:** Train or "teach" an algorithm using input-output pairs (labelled/categorized data)
  - Classification
  - Regression
- **Unsupervised Learning:** No feedback, "make sense" of structure in the data (*knowledge discovery*)
  - **Clustering**
  - Dimensionality Reduction (e.g., PCA)
  - Feature Learning (e.g., Autoencoders)
- **Reinforcement Learning:** Equip intelligent agents with reward-maximizing decision-making (action-taking)





Unsupervised Learning



No need to collect large labeled datasets (time-consuming and expensive)

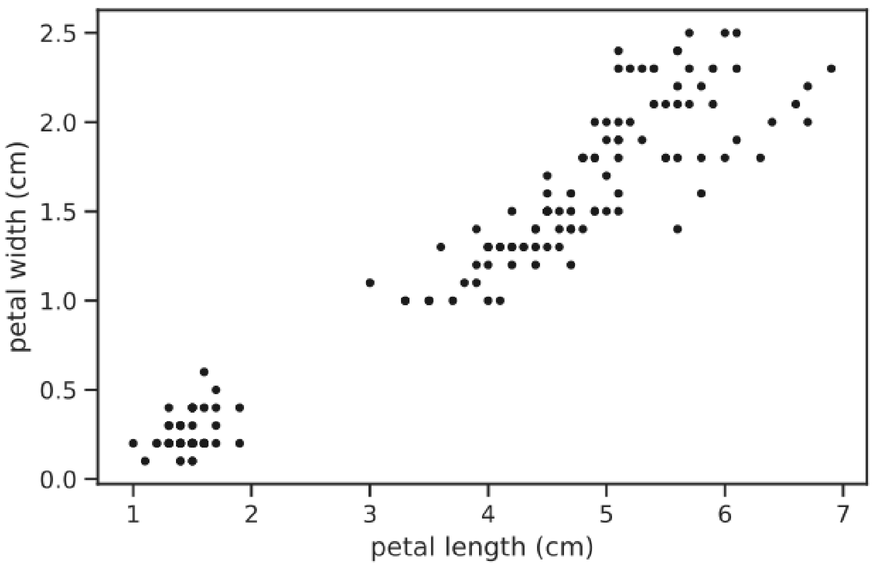
Avoid categorizing data, e.g. ambiguous situations like labeling an action – **better for ill-defined tasks**

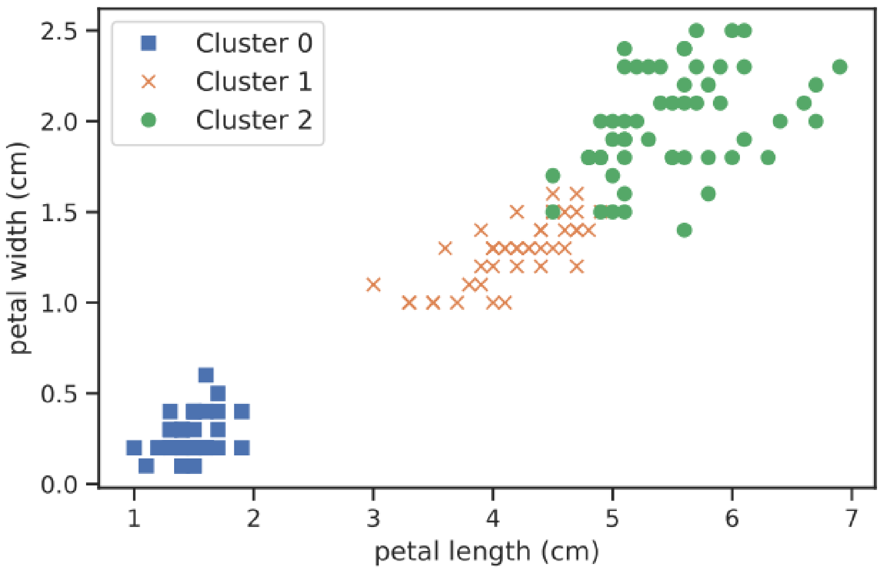
"Explain" high-dimensional inputs (data), rather than just low-dimensional outputs

Clustering – Basic Idea

- **Goal:** Automatically group data into coherent subsets or regions (**clusters**) of "similar" points

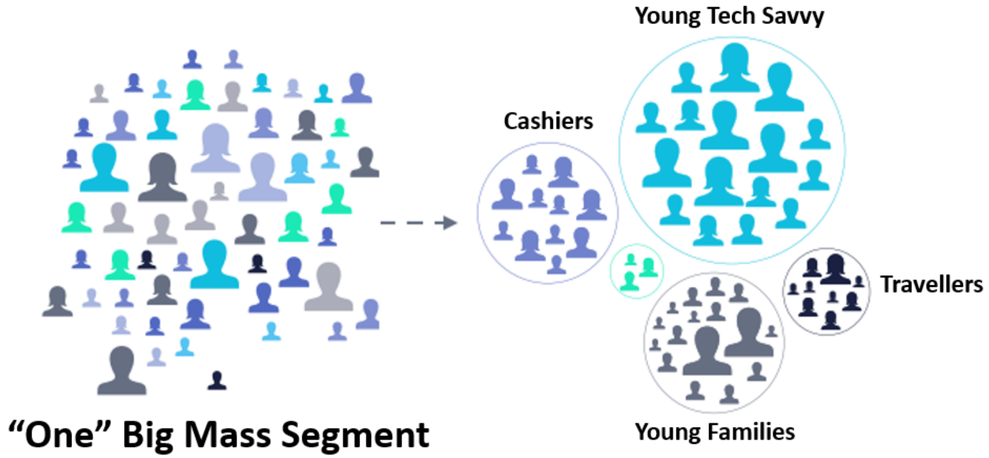
Murphy, *"Probabilistic Machine Learning: An Introduction"*, 2022

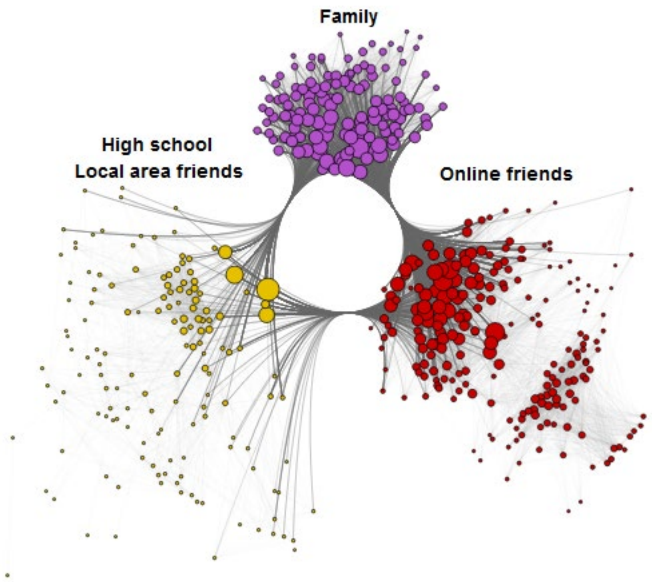


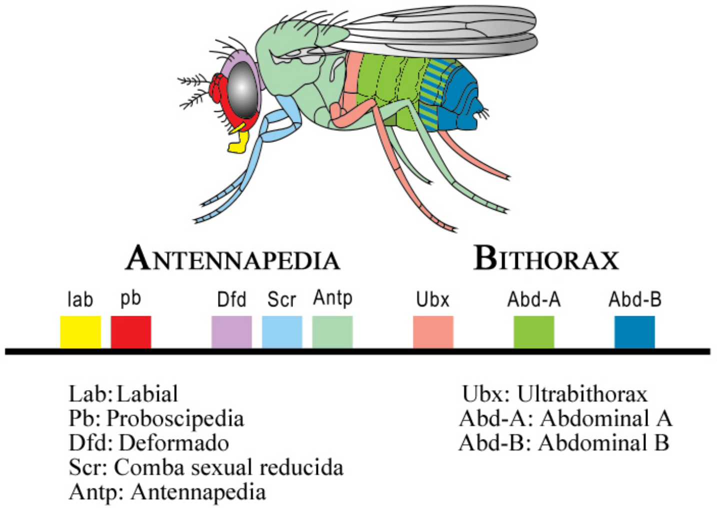


- No "correct" number of clusters (K): can be anything from $1 \rightarrow N$

Clustering – When & Why?







Sources: Smartera3s (customer segmentation), Wolfram (social network), Wikipedia (gene clustering), Max Planck Institute (astronomical data analysis)

Clustering Algorithm Categories

- Hierarchical
  - Connect objects according to a dissimilarity matrix in a tree structure
  - Used typically in biogenetics, e.g., dissect plant/insect down to gene granularity
- Centroid-based, e.g., **K-means**
  - Objective-oriented where each cluster is represented by a **centroid**
  - Assign objects to nearest cluster center, e.g., via Euclidean distance
- Distribution-based, e.g., **GMM clustering**
  - Clusters defined based on **likelihood** of belonging to a distribution
  - E.g., assign objects to Gaussian most likely to belong to
- Other types: mean-shift (KDE), spectral (eigenvalue → pairwise similarity)

In all cases attempting to assign "similar" points to the same cluster

K-means Clustering – Setting

- Assume **fixed** K no. of clusters with cluster centroids $\boldsymbol{\mu}_k$ for $k \in \{1, \ldots K\}$
- Given $D = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ with $\mathbf{x} \in \mathbb{R}^n$ but **no labels** $y^{(i)}$ available
- **Similarity?** Computed in terms of a **distance measure**, most commonly Euclidean
- Two steps:
  1. **Cluster assignment**
  2. Move centroid (**mean update**)

K-means Clustering – Algorithm

Randomly initialize K distinct cluster centroids $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^n$

Can be at sample locations OR random points in data space

Repeat until convergence (values no longer changing) {

1. **Cluster assignment:** for every i

$$c^{(i)} = \arg\min_k ||\mathbf{x}^{(i)} - \boldsymbol{\mu}_k||_2^2$$

2. **Mean update:** for every k

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:c^{(i)}=k} \mathbf{x}^{(i)}$$

}

K-means Clustering – Illustration

K-means Clustering – Objective

- The two update steps in the K-means algorithm are in fact finding the **local minimum** for the following **distortion** objective function:

$$\mathcal{J}(c^{(1)}, \ldots, c^{(N)}, \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K) = \mathcal{J}(c, \boldsymbol{\mu}) = \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}||_2^2,$$

- Measures the **sum of squared distances** between each example $\mathbf{x}^{(i)}$ and centroid $\boldsymbol{\mu}_{c^{(i)}}$ to which it has been assigned

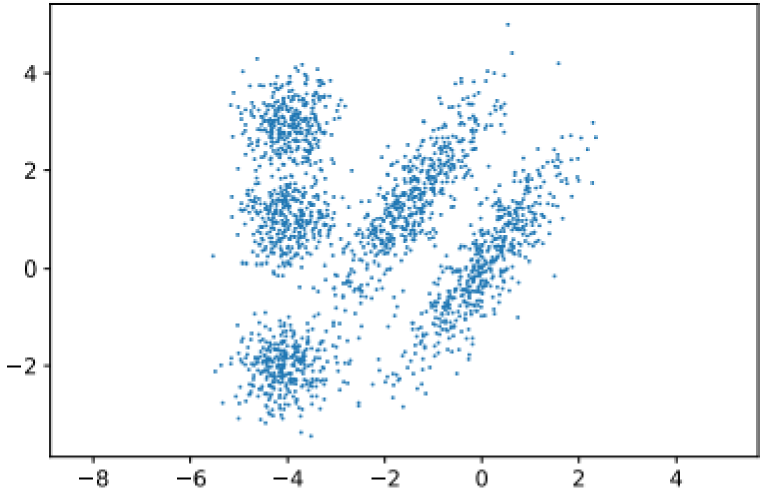
- **Non-convex** objective so K-means may suffer from bad local minima

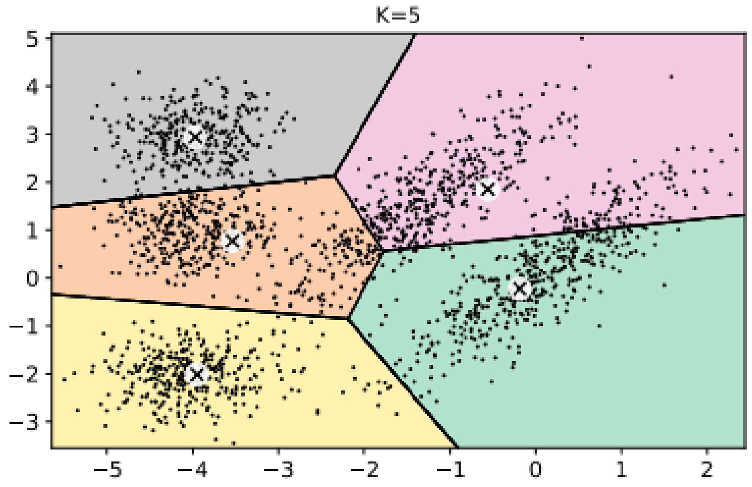
- Discussion of other properties/limitations provided in Notebook…

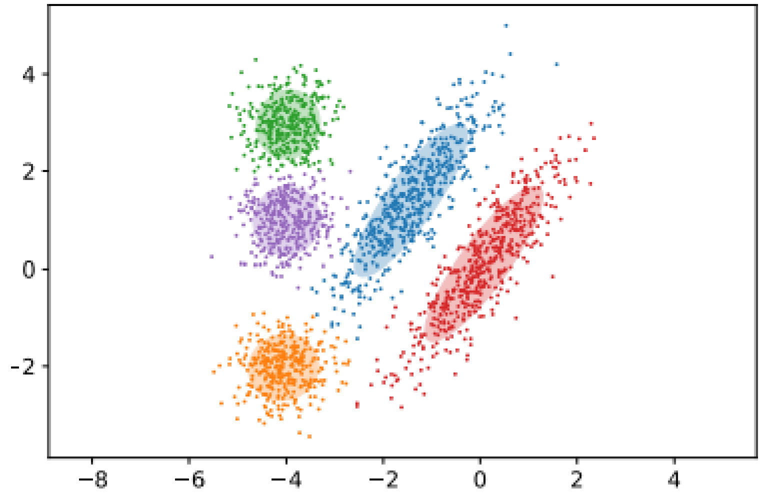
Coding Break

GMM Clustering – Setting

- Similarity now in terms of **likelihood**
- Two steps:
  1. Fit the model, e.g. by computing the MLE of the parameters
  2. Then associate each sample $\mathbf{x}^{(i)}$ with a discrete variable $z^{(i)}$ (**cluster label**)







GMM Clustering – Algorithm

Given data $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$

1. Fit a GMM $p(\mathcal{D} \mid \boldsymbol{\theta})$, *e.g.* using **Expectation Maximization (EM)**

$$\begin{aligned}
\arg\max \log p(\mathcal{D} \mid \boldsymbol{\theta}) &= \arg\max \log \left(\sum_{k=1}^{K} a_k p_k(\mathbf{x}) \right) \\
&= \arg\max \log \left(\sum_{k=1}^{K} p(z = k \mid \boldsymbol{\theta}) p(\mathbf{x} \mid z = k, \boldsymbol{\theta}) \right)
\end{aligned}$$

2. Can then perform MAP clustering for each $\mathbf{x}^{(i)}$ as:

$$\arg\max_{k} \left[\log p(\mathbf{x}^{(i)} \mid z^{(i)} = k, \boldsymbol{\theta}) + \log p(z^{(i)} = k \mid \boldsymbol{\theta}) \right].$$

GMM Clustering – Illustration

K-means & GMM Clustering Overlaps

- K-means clustering is a special case of the EM algorithm for GMMs

- Two simplifications:
  1. All covariance matrices are assumed **fixed** $\mathbf{\Sigma}_k = \sigma^2\mathbf{I}$
  2. The spherical Gaussian clusters all have **equal prior probability** $a_k = \frac{1}{K}$

- Both assume K is provided beforehand… **Kernel Density Estimation (KDE)** based clustering approaches circumvent this requirement

Concluding Remarks

- **Clustering** is a powerful unsupervised learning technique with a diverse array of applications, understandably so given the benefits of algorithms that operate without the requirement for labeled data

- Code:

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/unsupervised_learning/gmm_fitting_clustering.ipynb

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/unsupervised_learning/k_means_clustering.ipynb

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/unsupervised_learning/k_means_image_segmentation.ipynb