Pattern Analysis Thursday Meeting
Juliane Hoffmann

Worksheet 3: Thursday, May 22, 2025

Task 1: Mean Shift Clustering

(a) What is the idea behind the mean shift clustering?

> The idea behind mean shift clustering is to iteratively shift the center of a kernel function towards the direction of the highest density of data points.

(b) What are the kernel constraints?

> - Radial symmetry
> - Kernel function must accept squared differences

(c) What are the steps of the mean shift algorithm?

> 1. Mean Shift Iteration: For each data point, calculate the mean shift vector, which represents the direction to move the point towards a higher density region
> 2. Update data points: Shift each point in the direction of the mean shift vector
> 3. Repeat 1. and 2. until convergence
> 4. Group samples that converge to nearby locations into same clusters

Task 2: Kernel Selection for Mean Shift Clustering

(a) What role does the bandwidth of the kernel play in Mean Shift Clustering?

> The bandwith determines the size of the region around each data point where the algorithm searches for the mode.

(b) How does increasing the bandwidth affect the clustering result?

> A larger bandwidth will:
>
> - consider a larger neighborhood
> - lead to fewer, larger clusters

- produce smoother density estimates
- lead to a more generalized solution, possibly missing smaller structures in the data

(c) What do the clustering results look like when using a smaller bandwidth?

A smaller bandwidth will:

- consider a smaller/local neighborhood
- lead to more, smaller clusters
- produce more detailed/granular density estimates
- lead to a more detailed and localized solution, possibly overfitting to noise

(d) What happens to the clustering process if the kernel is not radially symmetric (going against the defined constraints of the kernel function)?

When using a non-radially symmetric kernel:

- the influence of a data point not only depends on distance but direction
- a directional bias in the kernel is introduced, which leads to a distorted density estimate

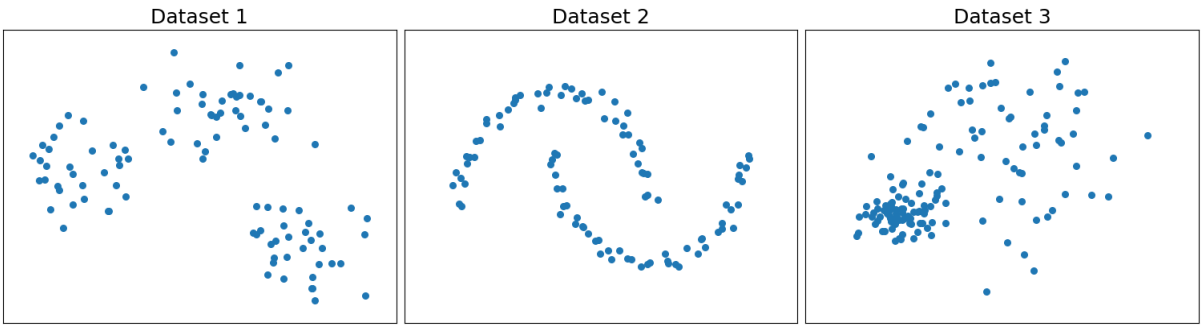
Task 3: K-means Model Selection

What does the optimality criterion given in Equation 1 aim to achieve?

$$k^* = \underset{k}{\operatorname{argmin}}\{k \mid G(k) \geq G(k+1) - s'_{k+1}\} \tag{1}$$

The optimality criterion is used for selecting the optimal number of clusters k. The aim is to determine the smallest number of clusters k where the gap statistic $G(k)$ indicates sufficient separation between clusters. It helps in identifying the point beyond which increasing the number of clusters k does not provide a statistically significant improvement in the clustering structure.

Task 4: Algorithm Selection



For each of the three datasets shown above:

(a) Describe the shape and distribution of the data

(b) Decide which clustering algorithm among the following are most suitable and explain why:

- K-means
- Gaussian Mixture Model (GMM)
- Mean Shift

(c) Justify why the other methods might fail

Dataset 1

- Shape and Distribution:
  - three well-seperated, roughly spherical clusters
- Suitable Clustering Method:
  K-means is most suitable, because the algorithm assumes spherical, equally size clusters, which fits the dataset well.
- Less suitable:
  - GMM: would work, but is unnecessarily complex since clusters are well-seperated and isotropic
  - Mean Shift: would work but is computationally more expensive

Dataset 2

- Shape and Distribution:
  - two interleaved, moon shaped clusters, non-convex and not linearly seperable
- Suitable Clustering Method:
  Mean Shift is most suitable, because it can handle arbitrary cluster shapes and identifies clusters based on the mode of the density distribution
- Less suitable:
  - K-means: assumes convex clusters and will cut across the moons
  - GMM: assumes elliptical Gaussian distributions and will also struggle with non-convex shape

Dataset 3

- Shape and Distribution:
  - two overlapping clusters, one dense and compact, the other sparse and more dispersed
- Suitable Clustering Method:
  GMM is most suitable because the approach can model clusters of different shapes, sizes and densities by adjusting the covariance matrices
- Less suitable:
  - K-means: assumes equal sized clusters with straight boundaries and therefore cannot handle the overlapping clusters
  - Mean Shift: might produce too many clusters or struggle to define dense vs. sparse areas depending on bandwidth