Lecture Pattern Analysis

Part 08: Mean Shift

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
May 13, 2021

Introduction

- Mean Shift is another widely known algorithm for clustering[1]
- It iteratively performs gradient ascent on the derivative of a kernel
- This iteration converges to a mode (local maximum) of the density **without** estimating the full density

[1] The associated paper by Comaniciu and Meer is available in studOn

Kernel Notation and Constraints

- The kernel has to be **radially symmetric**, i.e., the kernel function may only depend on the Euclidean distance of a sample
- The kernel functions must accept squared differences $\|\mathbf{x}_0 - \mathbf{x}\|_2^2$ as input,

$$K(\mathbf{x}_0, \mathbf{x}) = c \cdot k(\|\mathbf{x}_0 - \mathbf{x}\|_2^2) \tag{1}$$

  where c is an arbitrary constant and $k(x)$ is the so-called kernel profile
- This definition admits in particular the
  - Gaussian kernel with

$$k_{\text{Gauss}}(x) = \exp(-\frac{1}{2}x) \tag{2}$$

  - Epanechnikov kernel[2] with

$$k_{\text{Ep}}(x) = \begin{cases} 1 - x & \text{for} |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

- Kernel sizes: Assume that distances $\|\mathbf{x}_0 - \mathbf{x}\|_2^2$ are already size-normalized[3]

[2] The typical way to write the Epanechnikov kernel is shown in Hastie/Tibshirani/Friedman Eqn. (6.3) and Eqn. (6.4) for the 1-D case. Our notation differs because our input is already squared and pre-factors factors are absorbed in c in Eqn. 1

[3] See paper by Comaniciu/Meer Eqn. (1) and Eqn. (2)

Gradient Computation

- To find the maximum, calculate the gradient of a kernel density estimate

$$\nabla p(\mathbf{x}) = \nabla \frac{1}{N} \sum_{i=1}^{N} K_\lambda(\mathbf{x}_i, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \nabla K_\lambda(\mathbf{x}_i, \mathbf{x}) \tag{4}$$

- Insert $K_\lambda(\mathbf{x}_0, \mathbf{x}) = ck(\|\mathbf{x}_0 - \mathbf{x}\|_2^2)$, and substitute $s = \|\mathbf{x}_0 - \mathbf{x}\|_2^2$
- Then, the first derivative of $k(s)$ w.r.t. $\mathbf{x}$ consists of

$$\frac{\partial k(s)}{\partial s} = k'(s) \qquad \frac{\partial s}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}_i - \mathbf{x})^T(\mathbf{x}_i - \mathbf{x})}{\partial \mathbf{x}} = -2(\mathbf{x}_i - \mathbf{x}) \tag{5}$$

which leads to the gradient

$$\nabla p(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} c \cdot k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2)(-2(\mathbf{x}_i - \mathbf{x})) \stackrel{!}{=} 0 \tag{6}$$

From the Gradient to the Mean Shift Vector

- The gradient equation directly provides one gradient ascend step:

$$\nabla p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} c \cdot k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2)(-2(\mathbf{x}_i - \mathbf{x})) = 0 \tag{7}$$

$$\sum_{i=1}^{N} k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2)(-2(\mathbf{x}_i - \mathbf{x})) = 0 \tag{8}$$

$$\sum_{i=1}^{N} k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2) \cdot \mathbf{x}_i - \sum_{i=1}^{N} k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2) \cdot \mathbf{x} = 0 \tag{9}$$

$$\frac{\sum_{i=1}^{N} k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2) \cdot \mathbf{x}_i}{\sum_{i=1}^{N} k'(\|\mathbf{x}_i - \mathbf{x}\|_2^2)} - \mathbf{x} = 0 \tag{10}$$

- The last row is the normalized gradient, also called the **mean shift vector**

Mean Shift Algorithm

1. Calculate the mean shift vector $m^{(t)}(\mathbf{x})$ for iteration t:

$$m^{(t)}(\mathbf{x}) = \frac{\sum_{i=1}^{N} k'\left(\|\mathbf{x}_i - \mathbf{x}^{(t)}\|_2^2\right) \cdot \mathbf{x}_i}{\sum_{i=1}^{N} k'\left(\|\mathbf{x}_i - \mathbf{x}^{(t)}\|_2^2\right)} - \mathbf{x}^{(t)} \tag{11}$$

2. Update position $\mathbf{x}^{(t)}$:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + m^{(t)}(\mathbf{x}) = \frac{\sum_{i=1}^{N} k'\left(\|\mathbf{x}_i - \mathbf{x}^{(t)}\|_2^2\right) \cdot \mathbf{x}_i}{\sum_{i=1}^{N} k'\left(\|\mathbf{x}_i - \mathbf{x}^{(t)}\|_2^2\right)} \tag{12}$$

   (note that $\mathbf{x}^{(t)}$ cancels, since it also occurs in $m^{(t)}(\mathbf{x})$)

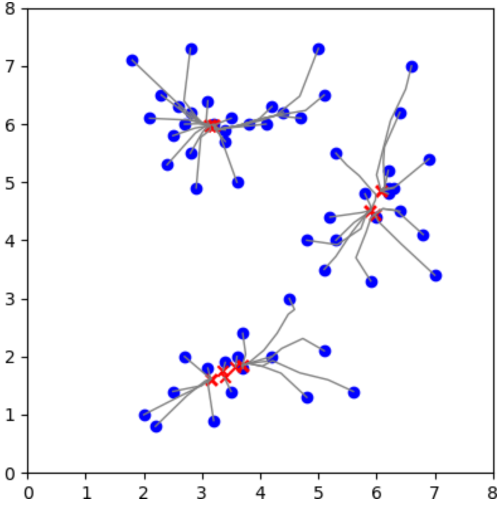
3. Goto 1) until convergence

Remarks

- With the Epanechnikov kernel, the update is just the mean of the samples within a D-dimensional sphere, hence the name “mean shift”
- Mean shift for clustering:
  - Run mean shift for each sample
  - Collect all samples in a cluster that converge to reasonably close locations
- Hence, the required parameters are
  - The kernel parameters (only window size for Epanechnikov and Gauss)
  - Cluster linking parameters for postprocessing (e.g., a distance threshold)
- Larger kernels lead to less clusters (less local maxima)
- Smaller kernels lead to more clusters (more local maxima)

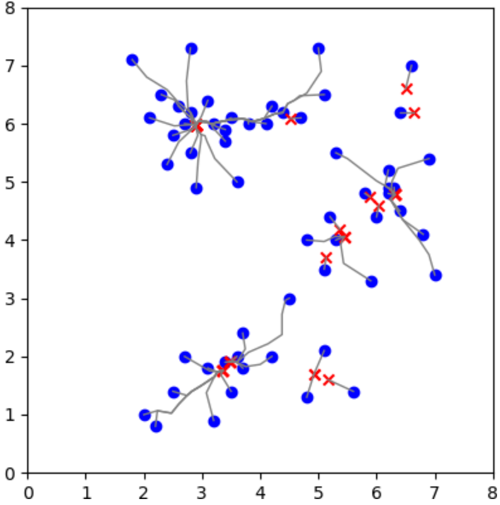
Example Run for Kernel Size 1.5

- Gray: path of a sample
- Red cross: Mode at the end of a path of a sample



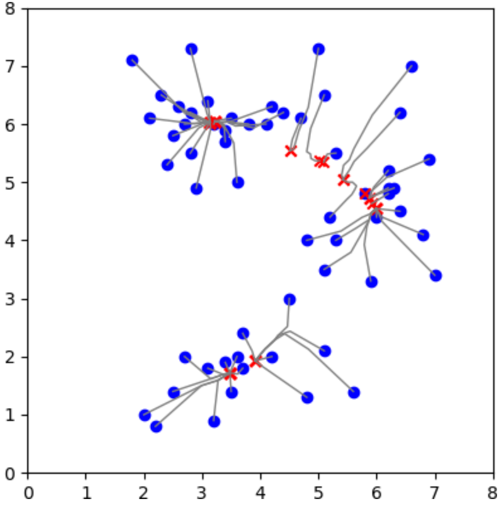
Example Run for Kernel Size 1

- Smaller kernel ($1.5 \rightarrow 1$): more clusters
- Gray: path of a sample
- Red cross: Mode at the end of a path of a sample



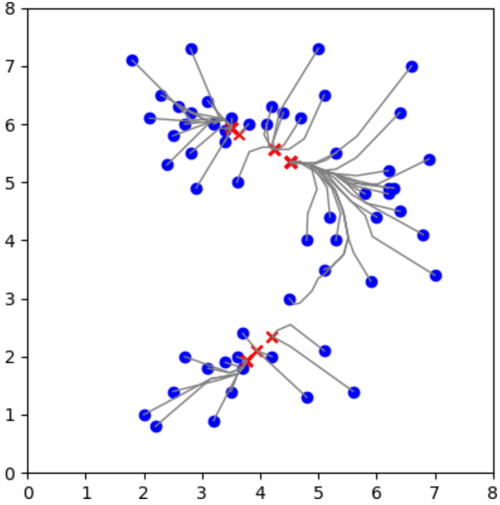
Example Run for Kernel Size 2

- Larger kernel ($1.5 \rightarrow 2$): clusters start to merge
- Note the characteristic mode ridge in the upper part
- Gray: path of a sample
- Red cross: Mode at the end of a path of a sample



Example Run for Kernel Size 2.5

- Larger kernel (1.5 $\rightarrow$ 2.5): upper 2 clusters almost merged
- Gray: path of a sample
- Red cross: Mode at the end of a path of a sample



Example Run for Kernel Size 3

- Larger kernel ($1.5 \to 3$): only a single cluster
- Gray: path of a sample
- Red cross: Mode at the end of a path of a sample