Outlines

- Density estimation
- Nonparametric kernel density estimation
- Mean shift
- Mean shift clustering
- Mean shift filtering and segmentation
Statistical Background

Density Estimation: Given a sample $S = \{x_i\}_{i=1..N}$ from a distribution obtain an estimate of the density function $\hat{f}(\cdot)$ at any point.

**Parametric**: Assume a parametric density family $f(.) | \theta$, (ex. $N(\mu, \sigma^2)$) and obtain the best estimator $\hat{\theta}$ of $\theta$

Advantages:
- Efficient
- Robust to noise: robust estimators can be used

Problem with parametric methods
- An incorrectly specified parametric model has a bias that cannot be removed even by large number of samples.

**Nonparametric**: directly obtain a good estimate $\hat{f}(\cdot)$ of the entire density $f(.)$ from the sample.

Most famous example: Histogram

Kernel Density Estimation

- 1950s + (Fix & Hodges 51, Rosenblatt 56, Parzen 62, Cencov 62)
- Given a set of samples $S = \{x_i\}_{i=1..N}$ we can obtain an estimate for the density at $x$ as:

$$
\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) = \frac{1}{N} \sum_{i=1}^{N} K_h(x - x_i)
$$
where \( K_h(t) = K(t/h)/h \) called kernel function (window function)

\( h \): scale or bandwidth

\( K \) satisfies certain conditions, e.g.:

\[
\int K_h(x) \, dx = 1
\]

\( K_h(x) \geq 0 \)

Kernel Estimation

- A variety of kernel shapes with different properties.
- Gaussian kernel is typically used for its continuity and differentiability.

- Multivariate case: Kernel Product
  Use same kernel function with different bandwidth \( h \) for each dimension.
- General form: avoid to store all the samples

\[
\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{d} K_{h_j}(x^j - x_i^j)
\]

\[
\hat{f}(x) = \sum_{i=1}^{N} \alpha_i K_h(x - x_i)
\]
Kernel Density Estimation

Advantages:
- Converge to any density shape with sufficient samples. asymptotically the estimate converges to any density.
- No need for model specification.
- Unlike histograms, density estimates are smooth, continuous and differentiable.
- Easily generalize to higher dimensions.
- All other parametric/nonparametric density estimation methods, e.g., histograms, are asymptotically kernel methods.
- In computer vision, the densities are multivariate and multimodal with irregular cluster shapes.

Example: color clusters
- Cluster shapes are irregular
- Cluster boundaries are not well defined.

From D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis,"
Conversion - KDE

Estimation using Gaussian Kernel

Estimation using Uniform Kernel
Scale selection

- Important problem. Large literature.
- Small $h$ results in ragged densities.
- Large $h$ results in over smoothing.
- Best choice for $h$ depends on the number of samples:
  - small $n$, wide kernels
  - large $n$, Narrow kernels
  - $\lim_{n \to \infty} h(n) = 0$

Optimal scale

- Optimal kernel and optimal scale can be achieved by minimizing the mean integrated square error – if we know the density!
- Normal reference rule:
  $$h^{opt} = \left(\frac{4}{3}\right)^{1/5} \sigma \cdot n^{-1/5} = 1.06\hat{\sigma} \cdot n^{-1/5}$$
Scale selection

Mean Shift

- Given a sample $S = \{s_i : s_i \in \mathbb{R}^n\}$ and a kernel $K$, the sample mean using $K$ at point $x$:

$$m(x) = \frac{\sum_i s_i K(s_i - x)}{\sum_i K(s_i - x)}$$

- Iteration of the form $x \leftarrow m(x)$ will lead to the density local mode.
- Let $x$ is the center of the window.
  - Iterate until convergence.
  - Compute the sample mean $m(x)$ from the samples inside the window.
  - Replace $x$ with $m(x)$

Fukunaga and Hostler 1975 introduced the mean shift as the difference $m(x) - x$ using a flat kernel.

Cheng 1995 generalized the definition using general kernels and weighted data:

$$m(x) = \frac{\sum_i s_i K(s_i - x)w(s_i)}{\sum_i K(s_i - x)w(s_i)}$$

Recently popularized by D. Comaniciu and P. Meer 99+

Applications: Clustering [Cheng,Fu 85], image filtering, segmentation [Meer 99] and tracking [Meer 00].
Mean Shift

- Iterations of the form $x \leftarrow m(x)$ are called mean shift algorithm.
- If $K$ is a Gaussian (e.g.) and the density estimate using $K$ is
  \[ \hat{P}(x) = C \sum_i K(x - s_i)w(s_i) \]
- Using Gaussian Kernel $K_\sigma(x)$, the derivative is $K'_\sigma(x) = \frac{x}{\sigma^2}K_\sigma(x)$
  we can show that:
  \[ \nabla \frac{\hat{P}(x)}{\hat{P}(x)} = m(x) - x \]
- the mean shift is in the gradient direction of the density estimate.

Mean Shift

- The mean shift is in the gradient direction of the density estimate.
- Successive iterations would converge to a local maxima of the density, i.e., a stationary point: $m(x) = x$.
- Mean shift is a steepest-ascent like procedure with variable size steps that leads to fast convergence “well-adjusted steepest ascent”.
Mean shift and Image Filtering

Discontinuity preserving smoothing

- Recall, average or Gaussian filters blur images and do not preserve region boundaries.

Mean shift application:

- Represent each pixel \( x \) as spatial location \( x^s \) and range \( x^r \) (color, intensity)
- Look for modes in the joint spatial-range space
- Use a product of two kernels: a spatial kernel with bandwidth \( h_s \) and a range kernel with bandwidth \( h_r \)

\[
K_{h_s,h_r} = k_{h_s}(x^s)k_{h_r}(x^r)
\]

- Algorithm:
  - For each pixel \( x^s(x^r, x^r') \)
  - apply mean shift until conversion. Let the conversion point be \( (x^s', x^r') \)
  - Assign \( z_i = (x^s', x^r') \) as filter output
- Results: see the paper.
Mean Shift and Segmentation

- Similar to filtering but group clusters from the filtered image: group together all \( z_i \) which are closer than \( h_s \) in the spatial domain and closer than \( h_r \) in the range domain.

Let \( x_i \) and \( z_i, i = 1, \ldots, n \) be the \( d \)-dimensional input and filtered image pixels in the joint spatial-range domain and \( L_i \) the label of the \( i \)th pixel in the segmented image.

1. Run the mean shift filtering procedure for the image and store all the information about the \( d \)-dimensional convergence point in \( z_i \), i.e., \( z_i = y_{ic} \).
2. Delineate in the joint domain the clusters \( \{ C_{ip} \}_{p=1}^{m} \) by grouping together all \( z_i \) which are closer than \( h_s \) in the spatial domain and \( h_r \) in the range domain, i.e., concatenate the basins of attraction of the corresponding convergence points.
3. For each \( i = 1, \ldots, n \), assign \( L_i = \{ p \mid z_i \in C_{i_p} \} \).
4. Optional: Eliminate spatial regions containing less than \( M \) pixels.
Meanshift tracking

Appearance-Based Tracking

- current frame + previous location
- likelihood over object location
- current location
- Mode-Seeking
  - (e.g. mean-shift, Lucas-Kanade; particle filtering)
  - appearance model
    - (e.g. image template, or
      - color, intensity, edge histograms)

Sources

- Slides by D. Comaniciu