Statistical and Learning Techniques in Computer Vision
Lecture 3: Non-Parametric Density Estimation

Jens Rittscher and Chuck Stewart

1 Overview

- Motivation for non-parametric methods
- Review of point and histogram techniques
- Kernel density methods
- Nearest neighbor methods
- Application: multimodal image registration based on mutual information

2 Motivation

- Parametric models may not capture the data effectively: multiple peaks and heavy tails.
- Histograms are non-smooth and require many samples for accuracy

3 What We Are Given and What We Want

- A set of data points sampled from a single, unknown distribution:

$$\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \tag{1}$$

  (We switch to the lower case notation here to avoid confusion with exponentiation.)

- Our goal is to estimate the density function $p(\mathbf{x})$.
- We will denote an estimate of the function as

$$\hat{p}(\mathbf{x}) \tag{2}$$

4 Sample-Based Representation

- The choice of representation depends on what we need to compute using the density.
- For example, if we just need the mean and variance of the density (or even higher-order moments), we can just use the samples.

- Problems with this:
  - Multimodal distributions
  - Higher order moments require a large number of samples to obtain good estimates.

5 Histogram Representations

- For points in 1-d, choose an interval $[x_{\text{min}}, x_{\text{max}})$, select a number of desired bins, M, and compute
$$\Delta = \frac{x_{\text{max}} - x_{\text{min}}}{M} \tag{3}$$
- Number the bins from 0 to $M-1$ and let k_j be the number of sample points that fall within bin j, i.e.
$$k_j = |\{i : \lfloor (x_i - x_{\text{min}})/\Delta \rfloor = j\}| \tag{4}$$
- Then
$$\hat{p}(x) = k_{\lfloor (x - x_{\text{min}})/\Delta \rfloor}/N \tag{5}$$
- Problems:
  - Choice of M
  - Non-differentiable
  - Number of bins is exponential in the number of dimensions

6 Probabilities, Points and Regions

As a prelude to the rest of the discussion we consider the relationship between $p(\mathbf{x})$, the samples, and regions.

- Given a location $\mathbf{x}$, a region $\mathcal{R}$ of volume V centered at $\mathbf{x}$, and a value P for the probability that a point drawn from $p(x)$ falls into $\mathcal{R}$, the expected number of points that fall into this region is easily shown to be
$$k = NP \tag{6}$$
- If the region is relatively small, then an approximation to P is obtained as
$$P = \int_{\mathcal{R}} p(\mathbf{u}) d\mathbf{u} \approx V p(\mathbf{x}) \tag{7}$$
- Combining and rearranging gives the approximation
$$p(\mathbf{x}) \approx \frac{k/N}{V} \tag{8}$$
- As the number of points increases, we can decrease the volume V and therefore obtain more and more accurate approximations to $p(\mathbf{x})$.

7 Kernel Density Estimation with Hypercube Regions

The following technique is also called "Parzen Windows"

- Define the rectangle function

$$\psi(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq 1/2, j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

  (Here we are using the subscripts j to denote components of the point vector $\mathbf{u}$.) Note that this integrates to 1 over the d-dimensional domain.

- Define the region $\mathcal{R}$ from above as a d-dimensional hypercube of width h_N on each side. This implies

$$V = h_N^d \tag{10}$$

- Now

$$k = k_N = \sum_{i=1}^{N} \psi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_N}\right) \tag{11}$$

- Substituting these into (8) yields

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_N^d} \psi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_N}\right) \tag{12}$$

- It is easy to show that this is a density.

- A primary difference with histograms is that the functions ψ are centered on the data points $\mathbf{x}_i$.

8 Other Kernels

The rectangle function as defined in (9) is usually replaced with smoother functions:

- One common version is the normal distribution:

$$\psi(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}} e^{-\mathbf{u}^\top \mathbf{u}/2} \tag{13}$$

- The Epanechnikov kernel:

$$\psi(\mathbf{u}) = \begin{cases} \frac{d+2}{2c_d}(1 - \mathbf{u}^\top \mathbf{u}) & \mathbf{u}^\top \mathbf{u} \leq 1 \\ 0 & \mathbf{u}^\top \mathbf{u} > 1 \end{cases} \tag{14}$$

  where c_d is the volume of the d-dimensional unit hypersphere, e.g. $c_1 = 2$, $c_2 = \pi$, $c_3 = 4\pi/3$

- These all result in $p(\mathbf{x})$ being a density.

- Other kernels have been derived to satisfy asymptotic statistical properties, although the normal and the Epanechnikov kernels are widely used in practice.

9 Issues With Parzen Windows

- All data points are used:
  - $O(N)$ computation in the worst case
  - Spatial data structures, such as k-d trees, may be used to make this substantially more efficient — approximately $O(\log N)$ in practice.
  - When N is large, we can sample the points and only use a subset.
- Under mild conditions on the relationship between the kernel width and N, the estimated density $\hat{p}(\mathbf{x})$ (see 12) converges to $p(\mathbf{x})$ as $N \to \infty$.
- Kernel width
  - If the kernel width is too large the density is over-smoothed, when the kernel width is too small the density is too noisy
  - Ideal kernel widths may be derived if the form of $p(\mathbf{x})$ is known. For example for distributions that are approximately normal, if we can robustly estimate the standard deviation σ, then

$$h_n = \sigma \left(\frac{4n}{3} \right)^{-1/5} \approx 1.06 \sigma n^{-1/5} \tag{15}$$

    For densities suspected to have more than one mode, the leading multiplier 1.06 should be reduced.
  - Adapative widths may be used
- Samples on the tail of the distribution can cause trouble. This is one place where adaptive widths can be important.

10 K-Nearest Neighbor Methods

- Returning to (8),

$$p(\mathbf{x}) \approx \frac{k/N}{V}, \tag{16}$$

  we can think of the kernel density approach as adapting k for fixed V (for fixed N). Now we consider adapting V.
- For a given $\mathbf{x}$, order the N samples by their distance to $\mathbf{x}$ and find the k^{th}. Call this sample $\mathbf{x}_{k:N}$ (we leave the dependence on the point $\mathbf{x}$ were we are evaluating the density implicit). Then the k-nearest neighbor approximation to p is

$$\hat{p}_k(\mathbf{x}) = \frac{k}{N} \frac{1}{(2\|\mathbf{x}_{k:N} - \mathbf{x}\|)^d}. \tag{17}$$

- The choice of k is the major parameter and should typically increase with increasing N, but

$$\lim_{N \to \infty} \frac{k}{N} = 0 \tag{18}$$

in order for $\hat{p}(\mathbf{x})$ to converge to $p(\mathbf{x})$.

- Weaknesses
  - Continuous, but not differentiable
  - Not a density
  - $O(\log N)$ in the number of stored samples, using a k-d tree, but more samples are required than kernel-based methods.

11 Application: Mutual Information Registration

- Almost simultaneously in the mid-1990's two independent papers, [MCV$^+$97] and [WVA$^+$96], proposed registration algorithms that use Shannon's "mutual information" as an objective function to measure the alignment of two images.
- The following is a summary of this approach, including
  - Definition of entropy
  - Definition of mutual information
  - Maes's algorithm [MCV$^+$97], which is based on histograms
  - Wells and Viola [WVA$^+$96], which is based on Parzen windows.

  We may not have time for all of the following detail in class.

12 Images, Intensities and Probabilities

- Consider the two images $J_A(\mathbf{x})$ and $J_B(\mathbf{x})$.
- Abusing notation, we will use A and B to denote both the set of all possible intensities in the two images and to indicate the images themselves.
- We will think of intensities as samples from a random variable, which means each image forms a distribution (of intensities).
- We will denote the two image distributions as

$$p_A(a) \qquad \text{and} \qquad p_B(b) \tag{19}$$

where the domain of both a and b is the set of possible intensity values in each image. These may be different!

13 Entropy, Joint Entropy and Conditional Entropy

- The entropy of a distribution is the negative expected value of the log of the density:
$$H(A) = -\sum_{a \in A} p_A(a) \ln p_A(a). \tag{20}$$
- Entropy is always non-negative (because $-p \ln p$ is non-negative on the interval $[0..1]$).
- Entropy is maximized when p_A is uniform, and minimized when p_A is an impulse function. When p_A is a (discretized) Gaussian distribution, then $H(A)$ increases with increasing variance of the distribution.
- The joint entropy between distributions is
$$H(A, B) = -\sum_{a \in A, b \in B} p_{A,B}(a, b) \ln p_{A,B}(a, b). \tag{21}$$
Note that when p_A and p_B are independent, $H(A, B) = H(A) + H(B)$, whereas when p_A and p_B are perfectly correlated $H(A, B) = H(A) = H(B)$.
- The conditional entropy is
$$H(A|B) = -\sum_{a \in A, b \in B} p_{A,B}(a, b) \ln p_{A|B}(a|b). \tag{22}$$
At first this is somewhat counter-intuitive, but the following point should make it clearer:
  - The sum is the expected value of $\ln p_{A|B}(a|b)$, just as in the other definitions of entropy. In fact, if we put $p_{A,B}(a, b)$ in each and sum over a and b, we'd get the same definitions.

  Intuitively, the conditional entropy is low when A is well-explained by B.
- Finally, note that
$$H(A, B) = H(A|B) + H(B). \tag{23}$$

14 Mutual Information

- Defined in terms of entropy:
$$\begin{aligned} I(A, B) &= H(A) + H(B) - H(A, B) && (24)\\ &= \sum_{a,b} p_{A,B}(a, b) \ln \frac{p_{A,B}(a, b)}{p_A(a) p_B(b)} \\ &= H(A) - H(A|B) \\ &= H(B) - H(B|A) && (25)\\ & && (26) \end{aligned}$$

- Some properties:
  - $I(A, B) \geq 0$
  - If p_A and p_B are independent (bad, in the case of registration) then $I(A, B) = 0$.
  - If p_A and p_B are perfectly correlated (good, in the case of registration), then $I(A, B) = H(A) = H(B)$.
  - The second expression (the summation) is the Kullback-Leibler measure between two densities. In this case the densities are the the joint density and what the joint density would be if the two were distributions were independent.
- Intuitively, $I(A, B)$ is high when A is well-explained by B (B is well-explained by A).
- Finally, maximizing $I(A, B)$ is better than minimizing $H(A, B)$. In minimizing $H(A, B)$, all that is sought is a region of overlap between the images where there is low entropy. This could (and often is) the background region. Including $H(A)$ and $H(B)$, which increase with increasing complexity and variability in the image regions, forces the alignment into areas of both significant content as well as low joint entropy.

15 Mutual information as an alignment evaluation function

- Let A be the fixed image and B be the moving image.
- Let $T(B; \boldsymbol{\alpha})$ be the transformation function described by parameters $\boldsymbol{\alpha}$.
- Our goal is to find the parameters $\boldsymbol{\alpha}$ maximizing

$$I(A, T(B; \boldsymbol{\alpha})) = H(A) + H(T(B; \boldsymbol{\alpha})) - H(A, T(B; \boldsymbol{\alpha})). \tag{27}$$

- In order to evaluate this objective function, we must transform image B based on the parameters, re-compute the resulting densities and entropies, and then re-evaluate the objective function.
- One subtlety is that $H(A)$ should be re-evaluated as the transformation changes because the region of overlap between the images will change.

16 Algorithm 1: Non-derivative search [MCV+97]

- Powell's method, starting with searches in the directions of the individual rigid transformation parameters. Within plane parameters are manipulated first.

- Recompute marginal densities at all steps, including only the region of overlap between images, as above.
- Do NOT do trilinear histogramming. Instead, do partial-volume interpolation in the histogram. (This is justified both intuitively and empirically.)
- Expensive computation, slow convergence.

17 Algorithm 2: Density modeling through Parzen windows

- Parzen windows density:

$$p_A(a) = \frac{1}{N} \sum_i G(a - a_i; \Sigma) \tag{28}$$

  - a_i is the set of intensities of a randomly-chosen set of N points
  - G is the multivariate Gaussian density, with covariance $\Sigma = \sigma \mathbf{I}$
  - Width parameter σ is estimated from the data by minimizing the entropy.
- A similar form holds for the joint density.
- Empirical expected value of entropy is evaluated using a second set of M randomly chosen points:

$$H(A) \approx \frac{-1}{M} \sum_j \ln \sum_i G(a_j - a_i; \Sigma) \tag{29}$$

- Surprisingly small values of M and N are typically used — often as low as 50.
- For fixed sets $\{a_i\}$ and $\{a_j\}$ this is now a differentiable function.
- We can form the MI objective function

$$I(A, T(B; \boldsymbol{\alpha})) = H(A) + H(T(B; \boldsymbol{\alpha})) - H(A, T(B; \boldsymbol{\alpha})) \tag{30}$$

  using the sampling techniques described (sampling from A and B to compute the joint density), compute the derivative with respect to the parameters in $\boldsymbol{\alpha}$, and apply gradient descent.

18 Further Reading

- Bishop [Bis95, Section 2.5] provides a brief, but clear introduction.
- Duda, Hart and Stroke [DHS01, Chapter 4] provides a more detailed introduction, especially to Parzen windows.
- B.W. Silverman [Sil86] presents one of the most complete introductions to the topic. In addition he also illustrated the limitations of the techniques.

References

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.

[MCV+97] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[Sil86] B.W. Silverman. *Density Estimation for Stastistics and Data Analysis*. Chapman and Hall, 1986.

[WVA+96] William M. Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.