SYDE 372
Introduction to Pattern Recognition

Estimation and Learning: Part II

Alexander Wong

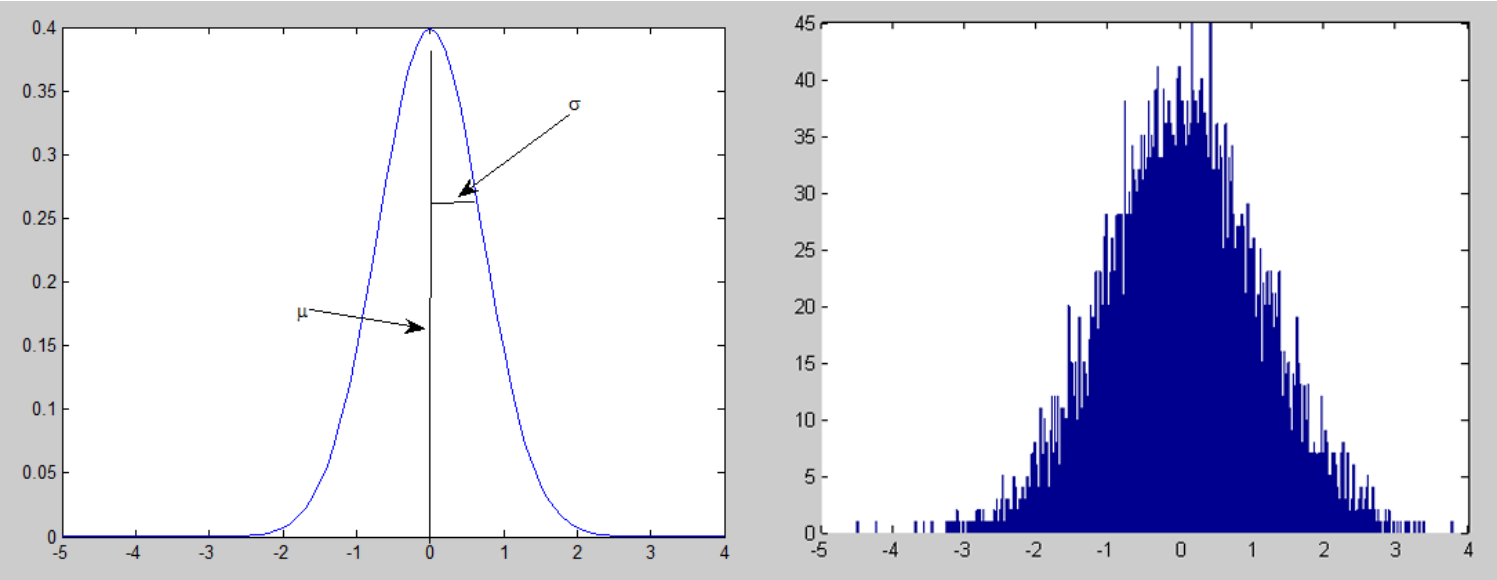
Department of Systems Design Engineering
University of Waterloo

Outline

Motivation

- So far, all of the model learning methods have been based on parametric estimation, where functional form of PDF is assumed to be known and the necessary parameters of the PDF are estimated.
- However, in many pattern recognition problems:
  - The functional form of the PDF is not known, or
  - The statistical distribution cannot be well modeled using parametric PDF models (i.e., $p(\underline{x}|\theta)$).
- Solution: we need to **directly** estimate the class distribution $p(\underline{x})$ from the samples $\{\underline{x}_i\}$!
- Such methods are called non-parametric estimation methods.

Motivation



(left) Parametric estimation (right) Non-parametric estimation

Non-parametric Learning

- Here, we assume that we do not know ANYTHING about the class conditional probability function.
- All we are given are N samples $\underline{x}_i$, which are labeled so we know that they belong to a single class.
- In this scenario, what we want to do is estimate the distribution directly based on a set of labeled samples for the class!
- Here, we will focus on learning a single class from labeled samples. In practice, we need to learn multiple distributions based on samples with different class labels.
- All we need to do is just run the non-parametric learning process multiple times, once for each class.

Types of Non-parametric Estimation

- There are three main categories of non-parametric estimation approaches:
  - **Histogram estimation:** Group given labeled samples into discrete regions to approximate $p(\underline{x})$.
  - **Parzen window estimation:** Approximate $p(\underline{x})$ in a continuous manner based on the local contribution of each sample, thus controlling resolution along x-axis explicitly, with resolution along PDF axis data dependent.
  - **kNN Estimation:** Approximate $p(\underline{x})$ in a continuous manner based on contribution of nearest neighbor samples, thus controlling resolution along PDF axis explicitly, with resolution along x-axis data dependent.

Histogram Estimation

- The simplest approach to non-parametric estimation of $p(\underline{x})$ is just constructing the normalized histogram!
- Why is this true?
  - Consider some interval $R = [a, b]$
  - If $p(x)$ is constant over this region, then

$$Pr(x \in R) = p_R = \int_a^b p(x)dx = p(a) \cdot (b - a) \qquad (1)$$

Histogram Estimation

- Supposed that we have N samples $x_1, \ldots, x_N$ taken from PDF $p(x)$
- The number of samples M that fall within region R must obey a binomial distribution:

$$p(M) = \left[\begin{array}{c} N \\ M \end{array} \right] p_R^M (1 - p_R)^{N-M} \tag{2}$$

- Taking the log gives us:

$$\log p(M) = \log \left[\begin{array}{c} N \\ M \end{array} \right] + \log[p_R^M] + \log[(1 - p_R)^{N-M}] \tag{3}$$

$$\log p(M) = \log \left[\begin{array}{c} N \\ M \end{array} \right] + M \log[p_R] + (N - M) \log[1 - p_R] \tag{4}$$

Histogram Estimation

- Taking the derivative and setting to zero:

$$\frac{\partial \log p(M)}{\partial p_R} = \frac{M}{p_R} - \frac{N-M}{1-p_R} = 0 \tag{5}$$

$$(1-p_R)M = p_R(N-M) \tag{6}$$

- This gives us the ML estimate of p_R as:

$$\hat{p}_R = \frac{M}{N} \tag{7}$$

Histogram Estimation

- Recall that:

$$p_R = p(a) \cdot (b - a) \tag{8}$$

- Plugging our ML estimate for p_R gives us:

$$\hat{p}_R = \hat{p}(x) \cdot (b - a) \tag{9}$$

$$\hat{p}(x) = \frac{\hat{p}_R}{(b - a)} \tag{10}$$

$$\hat{p}(x) = \frac{M/N}{(b - a)} \tag{11}$$

$$\hat{p}(x) = \frac{M}{N|R|} \tag{12}$$

where $|R|$ is the size of region R

Histogram Estimation: Steps

- Given a set of bins R_i:
  1. Count the number of samples M_i that falls into each bin i
  2. Count the total number of samples N
  3. For a particular pattern x, $p(x)$ can be computed as:

$$\hat{p}(x) = \frac{M_i}{N|R_i|} \quad \text{for } x \in R_i \tag{13}$$

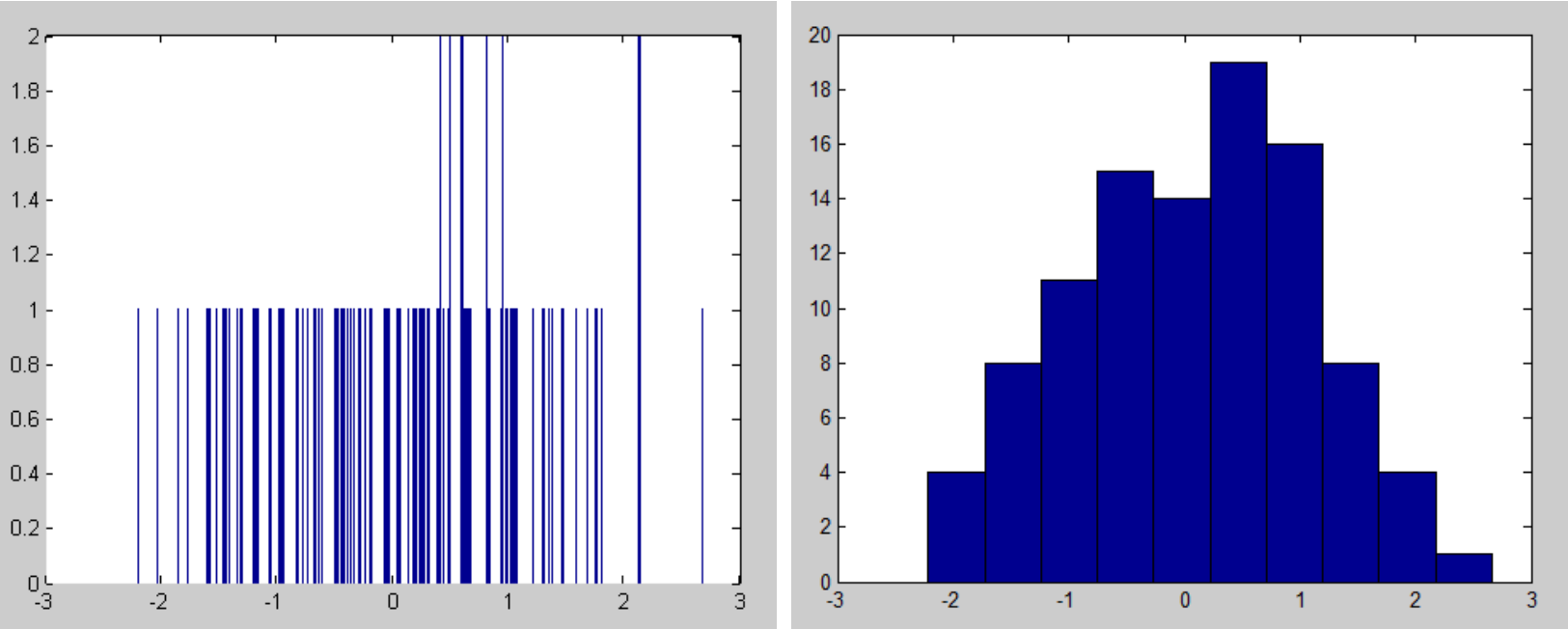
- What this means: For a particular pattern x, find which bin i it belongs to (R_i), and compute the ratio between the number of samples in that bin (M_i) and the total number of samples N times the region size $|R|$.

Histogram Estimation

- There is a fundamental tradeoff between resolution along x and resolution along $p(x)$:
  - For good resolution along x, we want to have small-sized regions. However this leads to small M_i, thus poor PDF resolution.
  - For good resolution along $p(x)$, we want to have large M_i and thus large-sized regions. However this leads to poor x resolution.
- Other major disadvantages:
  - Estimated PDF is ALWAYS discontinuous.
  - Shifting origin changes the shape of the estimated PDF.

Histogram Estimation: Example

Example: 100 samples from a Gaussian distribution



(left) Small region sizes (right) Large region sizes

Histogram Estimation: Example

- Suppose you are given the following set of samples $x = \{1, 1.1, 1.23, 2, 3, 3.1, 3.5, 4.1, 5, 5.8\}$.
- Based on the histogram estimation method, with a region size of $|R| = 2$ and a range between 0 and 6, what is $p(x = 3.2)$?
  - Step 1: Count the number of samples M_i that falls into each bin i ($M_1 = 3$, $M_2 = 4$, $M_3 = 3$)
  - Step 2: Count the total number of samples N ($N = 10$)
  - Step 3: Find the bin the pattern belongs to ($x = 3.2$ belongs to bin R_2)
  - Step 4: Compute $p(x = 3.2)$ based on identified bin

$$\hat{p}(x = 3.2) = \frac{4}{10(2)} = 0.2 \tag{14}$$

Parzen Window Estimation

- Aside from the problems mentioned before, the most obvious flaw in the histogram estimation approach is that we assume that the PDF $p(x)$ is constant over each region.
- Solution: the Parzen Window estimation approach does away with predefined regions and bin counting!
- Advantages:
  - No need to predefine region sizes
  - Estimated PDF is ALWAYS continuous.
  - Method is origin-independent.

Parzen Window Estimation

- Fundamental idea:
  - Every sample x_i locally influences the estimated PDF in the vicinity of x_i
  - In other words, since we observed x_i, the PDF there can't be too small; and if we see lots of samples in an area, then we expected the PDF to be correspondingly larger.
  - Given this mentality, the estimated PDF is then the sum of the contributions from each sample:

$$\hat{p}(x) \propto \sum_i \phi(x - x_i) \tag{15}$$

where ϕ is a window function which controls how each observed sample influences the PDF.

Parzen Window Estimation

- Properties of window function ϕ:
  - Window function must be normalized:

$$\int_{-\infty}^{\infty} \phi(x)dx = 1 \tag{16}$$

  - To change the locality of influence of a sample, we may wish to stretch or compress the window function:

$$\phi\left(\frac{x - x_i}{h}\right) \tag{17}$$

    where h is a scaling factor
  - To keep things normalized:

$$\int_{-\infty}^{\infty} \frac{1}{h}\phi\left(\frac{x - x_i}{h}\right) dx = 1 \tag{18}$$

Parzen Window Estimation

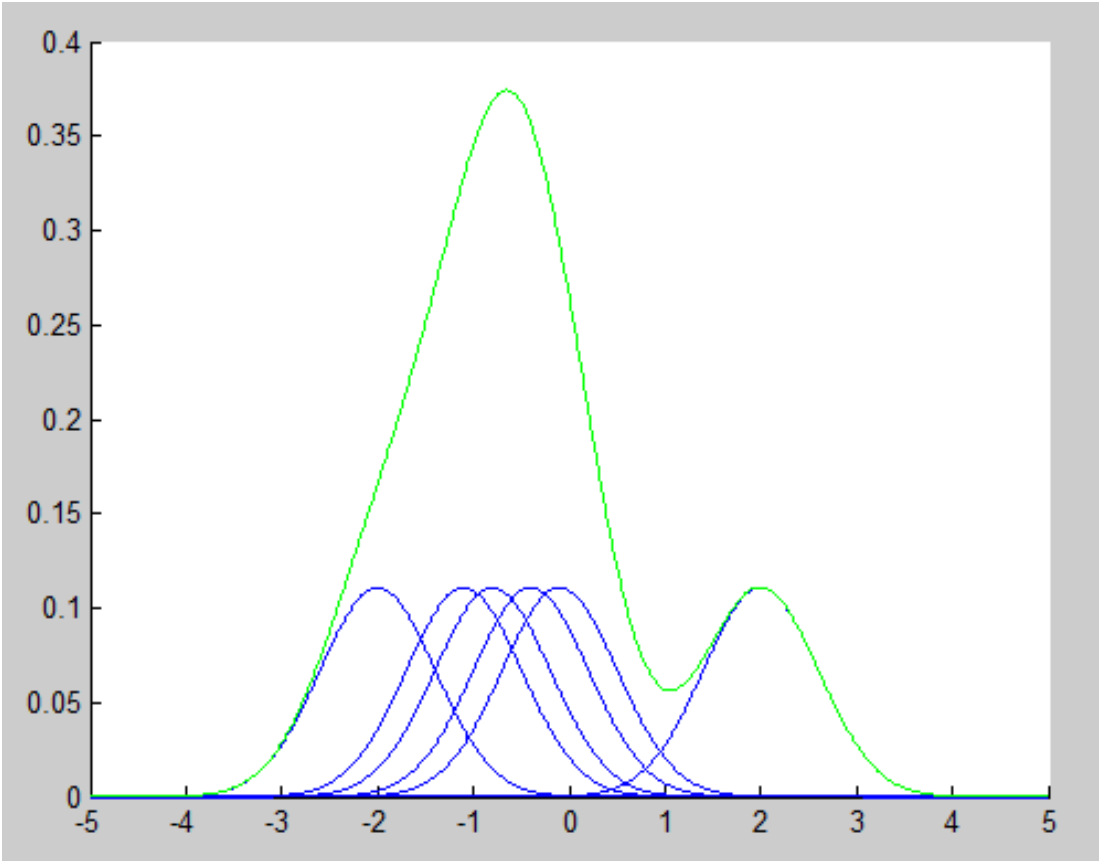
- Given N samples $\{x_i, \ldots, x_N\}$, the Parzen Window estimation of $p(x)$ is:

$$\hat{p}(x) = \frac{1}{N} \sum_i \frac{1}{h} \phi \left(\frac{x - x_i}{h} \right) \tag{19}$$

- Common window functions include: rectangular, triangular, Gaussian, and exponential.

Parzen Window Estimation: Example

Example: 6 samples, Gaussian window function



Parzen Window Estimation: Example

- Suppose you are given the following set of samples $x = \{1, 1.1, 1.23, 2, 5.8\}$.
- Based on the Parzen Window estimation method, with a Gaussian window function with $h = 1$, what is $p(x = 3.2)$?

$$\hat{p}(x) = \frac{1}{N} \sum_i \frac{1}{h} \phi \left(\frac{x - x_i}{h} \right) \tag{20}$$

$$\hat{p}(x = 3.2) = \frac{1}{5} \sum_{i=1}^{5} \phi \left(3.2 - x_i \right) \tag{21}$$

Parzen Window Estimation: Example

- Therefore,

$$\begin{aligned}\hat{p}(x=3.2) = & \frac{1}{5}\left[\frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}(3.2-1)^2\right)\right)\right] \\ & +\frac{1}{5}\left[\frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}(3.2-1.1)^2\right)\right)\right] \\ & +\frac{1}{5}\left[\frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}(3.2-1.23)^2\right)\right)\right] \\ & +\frac{1}{5}\left[\frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}(3.2-2)^2\right)\right)\right] \\ & +\frac{1}{5}\left[\frac{1}{\sqrt{2\pi}}\left(\exp\left(-\frac{1}{2}(3.2-5.8)^2\right)\right)\right]\end{aligned} \tag{22}$$

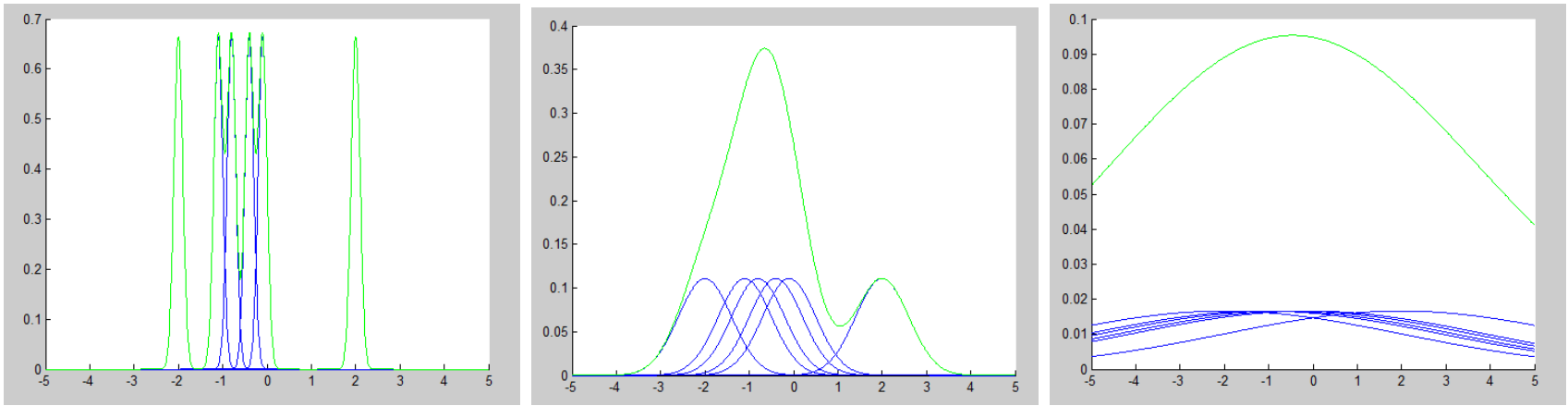
$$\hat{p}(x=3.2) = 0.0689 \tag{23}$$

Parzen Window Estimation: Scaling factor

- One drawback of Parzen Window estimation is that we still need to choose a window function and scale factor h
- As a general rule, one option is to try $h = K/\sqrt{N}$ for some constant K.
- For small N, the constant K is important:
  - If it is too small, the estimate is noisy with sharp peaks at the samples
  - If it is too large, the estimate is smeared with low resolution

Parzen Window Estimation: Scaling factor

Example: 6 samples, Gaussian window function, different K



(left) small K, (middle) medium K, (right) large K

k-Nearest-Neighbor Estimation

- In histogram and Parzen Window estimation methods, we fix the region size/window function width
- This explicitly controls the resolution along the x-axis, and the resolution along the PDF axis is data dependent.
- In the kNN method, we instead fix the number of samples M, and determine the size of region required at each point to enclose this many samples.
- Therefore, this explicitly controls the resolution along the PDF axis, and the x-axis resolution becomes data dependent.

k-Nearest-Neighbor Estimation

- To compute the kNN estimate of $p(x)$:
  - Create an interval $[x - \alpha, x + \alpha]$ centered around x
  - Increase α until it contains a suitable number of observations M
  - Compute estimate of $p(x)$ as:

$$\hat{p}(x) = \frac{M}{N|R(x)|} = \frac{M}{N \cdot 2\alpha} \tag{24}$$

where $R(x)$ is the smallest region, centered around x, which encloses M sample points.

k-Nearest-Neighbor Estimation

- Frequently we set $M = \sqrt{N}$, in which case the kNN method has no free parameters.
- If sample density is high, $|R(x)|$ will be small, and the estimate will have high resolution where it is needed.
- If sample density is low, $|R(x)|$ will be large, and resolution will be low which is probably acceptable in sparsely populated regions.
- **Main advantage of kNN estimation**: avoids setting $p(x)$ identically to zero in regions which happen not to have any samples, but instead results in a more realistic non-zero probability.
- **Disadvantage**: estimated PDF is highly "peaked" and non-normalized.

k-Nearest-Neighbor Estimation: Example

- Suppose you are given the following set of $N = 5$ samples $x = \{1, 1.1, 1.23, 2, 5.8\}$.
- Based on the k-Nearest-Neighbor estimation method, with $M = 3$, what is $p(x = 3.2)$?
- Given $x = 3.2$, the interval around x is $[3.2 - 2.1, 3.2 + 2.1] = [1.1, 5.3]$, as it encloses $M = 3$ samples $\{1.1, 1.23, 2\}$
- Compute estimate of $p(x = 3.2)$ as:

$$\hat{p}(x = 3.2) = \frac{M}{N \cdot 2\alpha} = \frac{3}{5 \cdot 2(2.1)} = 0.1429 \qquad (25)$$