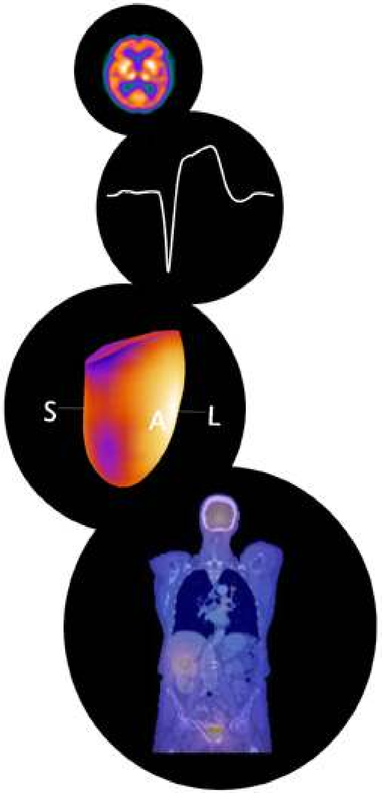


METU Informatics Institute

Min720

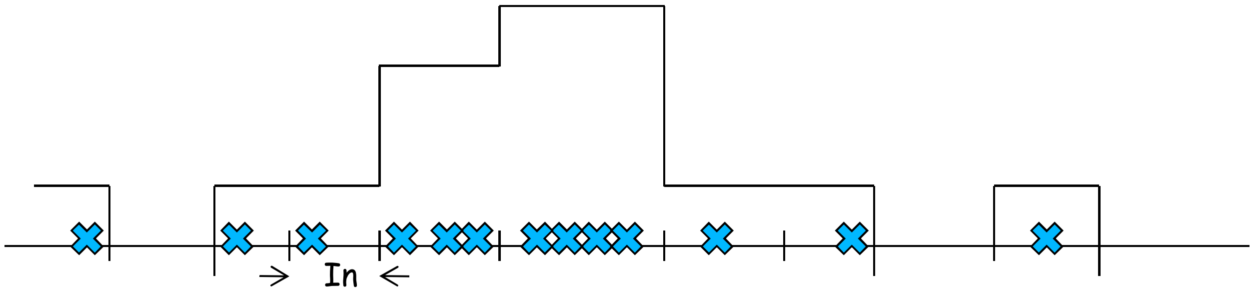
Pattern Classification with Bio-Medical Applications

Lecture Notes

Part 5: Density Estimation

Density Estimation

- We would like to know p(x) for a given x for Bayes Classifier. A histogram is crude estimation of a density function.



$$\hat{P}(x) = \frac{k_i}{n}\frac{1}{l_{In}}$$ for the interval In.

- n-total number of samples
- k_i - number of samples that fall in In.
- If x is multidimensional with dimension d, $\hat{P}_n(X) = \frac{k_i}{n}\frac{1}{V_i}$ where V_i is the volume of the d-dimensional hypercube.

- A smoother estimate would be obtained if a volume is extended around x to capture k_i samples.
  So k_i / n will be fixed. Measure V_i.

  In the limits,
- If $P_i(x)$ is to converge to p(x), then

$$\lim_{n\to\infty} V_i = 0$$
$$\lim_{n\to\infty} k_i = \infty$$
$$\lim_{n\to\infty} k_i / n = 0$$

that means, when there's enough number of samples in each interval, but still the number of samples in each interval is very small in the total number of samples, $\hat{p}_i(x) \to \hat{p}(x)$

To satisfy these conditions, either

- Specify and fix V_i as a function of n. $V_i = \frac{1}{\sqrt{n}}$ (parzen windows)

- Or, specify and fix k_i as a function of n.

  $k_i = \sqrt{n}$ (nearest neighbor estimation)

K-Nearest Neighbor Estimation of a Density Function

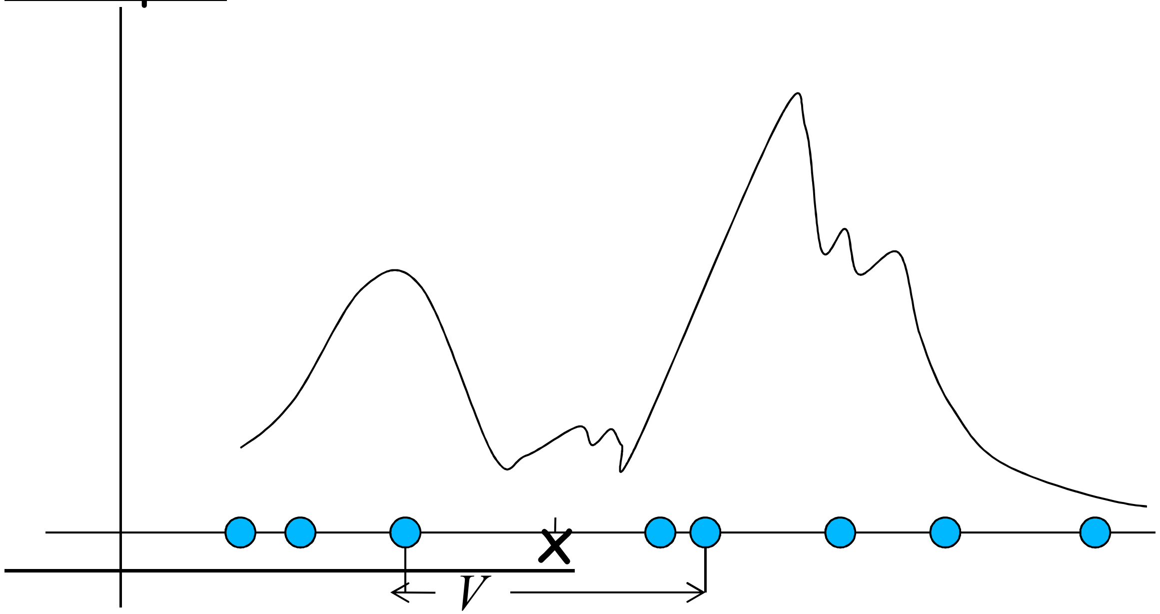
$$\hat{P}_i(X) = \frac{k/n}{V_i}$$

where we grow the region i to include k number of samples for region V_i

Curse of Dimensionality

- Each cube should have enough number of samples for a good approximation. The demand for a larger no. of samples grows exponentially with the dimensionality of a sample space.

Example:



Look at 3 nearest neighbors of each sample point and find volume V. Do it for all sample points. So, given x as above, what is the estimate of P(x)?

$$P(X) = (3/8)/V$$