THE UNIVERSITY OF TENNESSEE KNOXVILLE

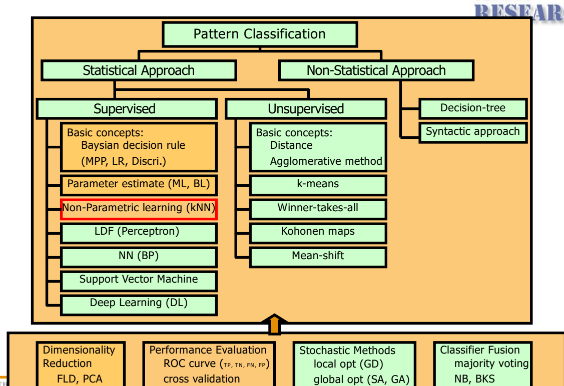
AICIP RESEARCH

ECE471-571 – Pattern Recognition

Lecture 9 – Nonparametric Density Estimation – Parzen Windows

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

AICIP RESEARCH



AICIP RESEARCH

Review - Bayes Decision Rule

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

Maximum Posterior Probability: For a given x, if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then x belongs to class 1, otherwise, 2.

Likelihood Ratio: If $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$, then x belongs to class 1, otherwise, 2.

Discriminant Function: The classifier will assign a feature vector x to class ω_i if $g_i(x) > g_j(x)$

Case 1: Minimum Euclidean Distance (Linear Machine), $\Sigma_i=\sigma^2 I$
Case 2: Minimum Mahalanobis Distance (Linear Machine), $\Sigma_i = \Sigma$
Case 3: Quadratic classifier, Σ_i = arbitrary

Estimate Gaussian, Two-modal Gaussian

Dimensionality reduction

Performance evaluation and ROC curve

3

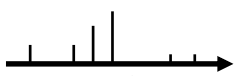
Motivation

- Estimate the density functions without the assumption that the pdf has a particular form

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

*Start from Histogram

- In order to generate a reasonable representation for the density, we'd like to first "smooth" the data over cells



- The probability that a vector x will fall into a region R is

$$P = \int_R p(x')dx'$$

- If $p(x)$ does not vary significantly within R, then $P = p(x)V$
  - V is the volume enclosed by R
- For a training set of n samples, k of them fall into the hypervolume V, we can then estimate $p(x)$ by

$$p(x) \approx p_n(x) = \frac{k_n / n}{V_n}$$

*Parzen Windows

$$p_n(x) = \frac{k_n / n}{V_n}$$

- The density estimation at x is calculated by counting the number of samples fall within a hypercube of volume V_n centered at x
- Let R be a d-dimensional hypercube, whose edges are h_n units long. Its volume is then $V_n = h_n^d$
- The window function

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le 0.5, \quad j = 1, \cdots, d \\ 0 & \text{otherwise} \end{cases} \qquad k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

- Therefore

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)}{V_n}$$

*Problem

AICIP RESEARCH

- Hypercube – why should a point just inside the hypercube contribute the same as a point very near to **x**, while a point just outside the hypercube contributes nothing?
- Use a continuous window function

THE UNIVERSITY OF TENNESSEE 7

*Continuous Window Function

AICIP RESEARCH

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)}{V_n}$$

- Univariate $\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$
- Multi-variate

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^d}\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)^T\Sigma^{-1}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right]$$

- Making **Σ** an identity matrix

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1 h_2\cdots h_d}\frac{1}{(2\pi)^{d/2}}\prod_{j=1}^{d}\exp\left[-\frac{1}{2}\left(\frac{x_j - x_{ij}}{h_j}\right)^2\right]$$

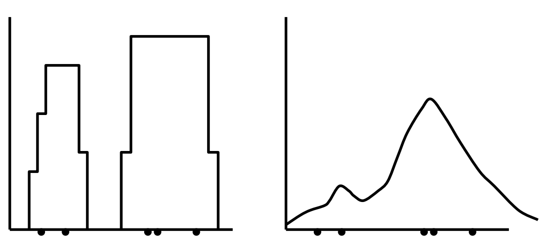
- h_j reflects the variance (spread) of the smoothing kernel (window function) in the jth coordinate direction. If we assume the spread is equal in all directions

$$p(x) = \frac{1}{nh^d(2\pi)^{d/2}}\sum_{i=1}^{n}\prod_{j=1}^{d}\exp\left[-\frac{1}{2}\left(\frac{x_j - x_{ij}}{h}\right)^2\right]$$

THE UNIVERSITY OF TENNESSEE 8

*Comparison

AICIP RESEARCH



THE UNIVERSITY OF TENNESSEE 9

*Another Problem

- How to choose *h*?
- A large *h* will result in a great deal of smoothing and loss of resolution
- A very small *h* will tend to degenerate the estimator into a collection of *n* sharp peaks, each centered at a sampling point
- Solution: *h* should depend on the number of samples. If only a few samples are available, we require a large *h* and considerable smoothing, whereas if many points are available, we can use a smaller *h* without the danger of degenerating into separate peaks.

10

*The Choice of h

- We make *h* a function of *n*

$$h = \frac{1}{\sqrt{n}}$$

11

*Problem with Parzen Windows

- Discontinuous window function -> Gaussian
- The choice of h
- Still another one: fixed volume

12