In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Classification: Non-Parametric Modeling

Hamid R. Rabiee
Jafar Muhammadi

Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Template designed by Jafar Muhammadi

Agenda

- **Parametric Modeling**
- **Non-Parametric Modeling**
  - **Density Estimation**
  - **Parzen Window**
    - Parzen Window - Illustration
    - Parzen Window and Classification
  - **K_n-Nearest Neighbor (K-NN)**
    - K-NN - Illustration
    - K-NN and a-posteriori probabilities
    - K-NN and Classification
  - **Pros and cons**

Template designed by Jafar Muhammadi

Parametric Modeling

- **Data availability in a Bayesian framework**
  - **We could design an optimal classifier if we knew $P(w_i)$ and $P(x|w_i)$**
  - **Unfortunately, we rarely have this complete information!**
- **Assumptions**
  - **A priori information about the problem**
  - **The form of underlying density**
    - Example: Normality of $P(x|w_i)$: Characterized by 2 parameters
- **Estimation techniques (studied in stochastic process course)**
  - **Maximum-Likelihood (ML) and the Bayesian estimations (MAP: Maximum A Posteriori)**
    - Results are nearly identical, but the approaches are different!
- **Other techniques (will be discussed later)**
  - **Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM)**

Template designed by Jafar Muhammadi

Non-Parametric Modeling

- **Non-parametric modeling tries to model arbitrary distribution without assuming certain parametric form.**
- **Non-parametric models can be used with arbitrary distributions and without the assumption that the forms of the underlying densities are known.**
- **Moreover, They can be used with multimodal distributions which are much more common in practice than unimodal distributions.**
- **There are two types of non-parametric methods:**
  - **Estimating $P(x|w_j)$**
    - Parzen window
  - **Bypass probability and go directly to a-posteriori probability estimation (Estimating $P(w_j|x)$)**
    - K_n-Nearest Neighbor

Template designed by Jafar Muhammadi

Density Estimation

- **Basic idea:**
  - **Probability that a vector x will fall in region R is:** $P = \int_R P(x')dx'$
    - P is a smoothed (or averaged) version of the density function P(x).
  - **If we have a sample of size n; therefore, the probability that k points fall in R is then:**

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}$$

    - The expected value for k is E(k) = nP
  - **ML estimation of *P* is reached for** $\hat{P}_{ML} = \hat{\theta} = \frac{k}{n}$
    - Therefore, the ratio k/n is a good estimate for the density function p.
  - **Assuming P(x) is continuous and that the region R is so small that P does not vary significantly within it, we can write (V is the volume of R):**

$$P = \int_R P(x')dx' \cong P(x)V$$

  - **Combining above equations, the density estimate becomes:**

$$P(x) \cong \frac{k/n}{V}$$

Template designed by Jafar Muhammadi

Density Estimation

- **The volume V needs to approach zero if we want to use this estimation**
  - Practically, V cannot be allowed to become small (since the number of samples is always limited).
  - Theoretically, if an unlimited number of samples is available, we can circumvent this difficulty
- **To estimate the density of x regarding above limitations, we do following steps:**
  - In n^{th} step, consider a total of n data samples with the centrality of x
  - Form *a region* R_n containing x
  - Let V_n be the volume of R_n, k_n the number of samples falling in R_n and $P_n(x)$ be the n^{th} estimate for P(x), then:

$$P_n(x) = (k_n/n)/V_n$$

  - Three necessary conditions for converging $P_n(x)$ to P(x) are:

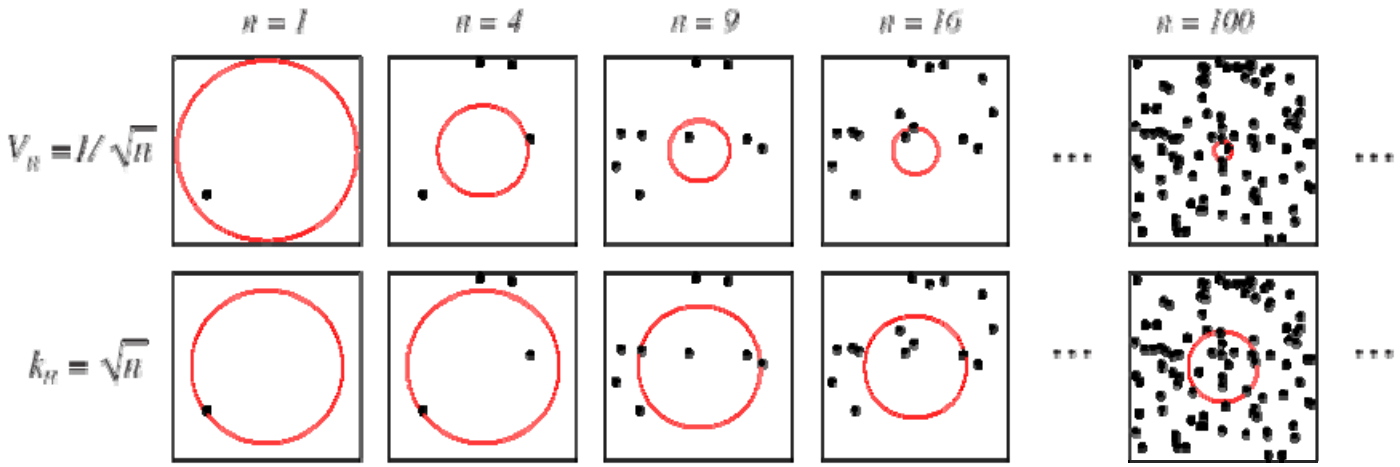
$$\lim_{n\to\infty} V_n \to 0 \qquad \lim_{n\to\infty} 1/k_n \to 0 \qquad \lim_{n\to\infty} k_n / n \to 0$$

  - There are two different ways of obtaining sequences of regions that satisfy these conditions:
    - Parzen-window estimation method: Shrink an initial region where $V_n = 1/\sqrt{n}$ and show that $P_n(x) \underset{n\to\infty}{\to} P(x)$
    - k_n-nearest neighbor estimation method: Specify k_n as some function of n, such as $k_n = \sqrt{n}$; the volume V_n is grown until it encloses k_n neighbors of x.

Template designed by Jafar Muhammadi

Density Estimation

- **Parzen window vs. k-nearest neighbor**



Template designed by Jafar Muhammadi

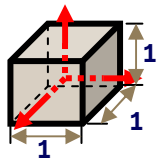
Parzen Window

- **Parzen-window approach to estimate densities**
  - **assume that the region R_n is a d-dimensional hypercube**

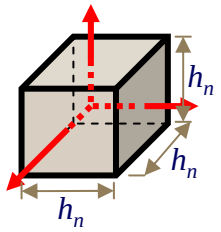
$$V_n = h_n^d \quad (h_n : \text{length of the edge of } R_n)$$

$$Let\ \varphi(u) \text{ be the following window function:}$$

$$\varphi(u) = \begin{cases} 1 & |u_j| \le \frac{1}{2} \quad j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$



  - **$\varphi((x-x_i)/h_n)$ is equal to unity if x_i falls within the hypercube of volume V_n centered at x and equal to zero otherwise.**
  - **The number of samples in this hypercube is:** $k_n = \sum_{i=1}^{i=n} \varphi\left(\frac{x - x_i}{h_n}\right)$



  - **Then, we obtain the following estimate:** $P_n(x) = \frac{1}{n} \sum_{i=1}^{i=n} \frac{1}{V_n} \varphi\left(\frac{x - x_i}{h_n}\right)$
    - $P_n(x)$ estimates $p(x)$ as an average of functions of x and the samples (x_i) (i = 1,… ,n). These functions φ can be general density function!

Template designed by Jafar Muhammadi

Parzen Window

- **Example:**
  - **The behavior of the Parzen-window method for the case where both P(x) & φ(u)~*N(0,1)***
  - **Let**

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}},\ h_n=\frac{h_1}{\sqrt{n}};\ (n>1, h_1: known\ parameter)$$

  - **Thus:**

$$P_n(x)=\frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h_n}\varphi\left(\frac{x-x_i}{h_n}\right)$$

P_n is an average of normal densities centered at the samples x_i.

  - **Numerical results for *n=1* and $h_1=1$**

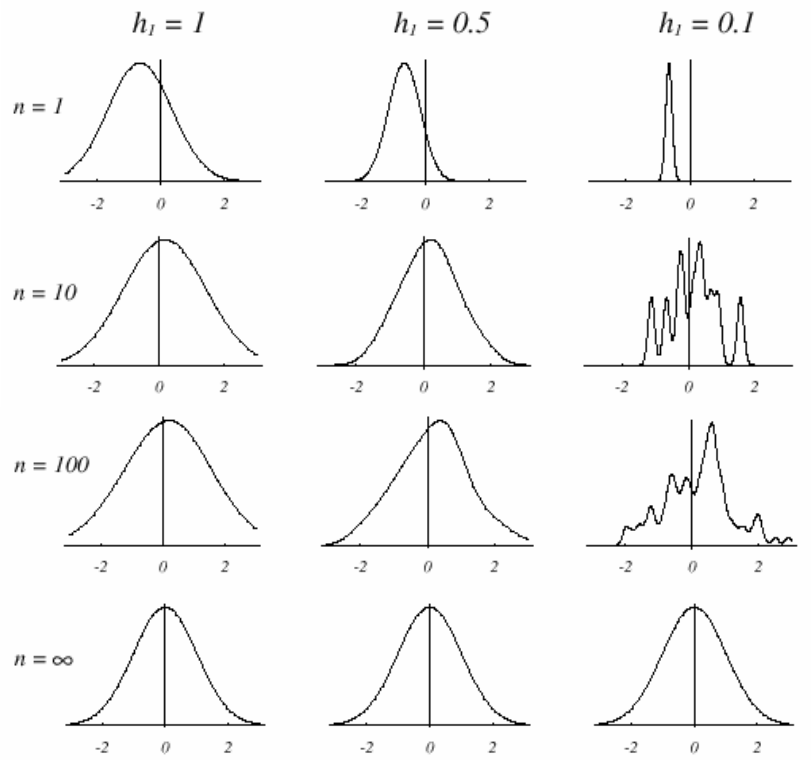
$$\mathrm{P}_1(x)=\varphi(x-x_1)=\frac{1}{\sqrt{2\pi}}e^{-1/2}(x-x_1)^2\rightarrow N(x_1,1)$$

  - For *n=10* and *h=0.1*, the contributions of the individual samples are clearly observable!

Template designed by Jafar Muhammadi

Parzen Window - Illustration

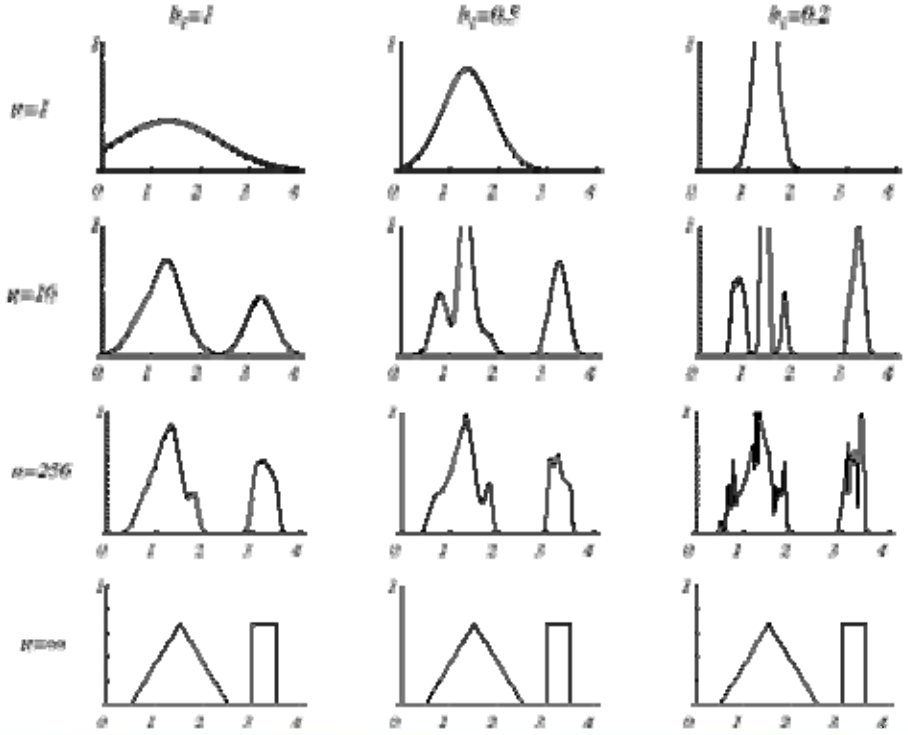
- **Example illustration**
  - Note that the n=∞ estimates are the same and match the true density function regardless of window width.



Template designed by Jafar Muhammadi

Parzen Window - Illustration

- **Example 2**
  - **Case where $P(x) = \lambda_1 U(a,b) + \lambda_2 T(c,d)$ (unknown density) – mixture of a uniform and a triangle density**
    - The P_n as the same as previous example



Template designed by Jafar Muhammadi

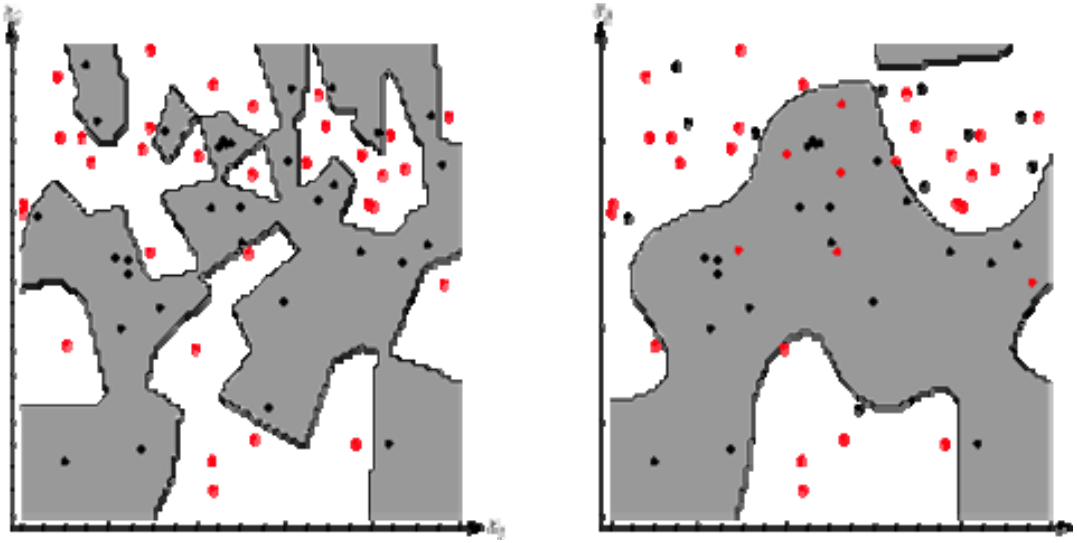
Parzen Window and Classification

- **In classifiers based on Parzen-window estimation:**
  - **We estimate the densities for each category and classify a test point by the label corresponding to the maximum posterior**
    - Using the points of only category w_i, $P(x|w_i)$ can be estimated
    - Knowing $P(w_i)$, posterior probabilities can be found
  - **The decision region for a Parzen-window classifier depends upon the choice of window function as illustrated in the following figure. (See next slide)**

Template designed by Jafar Muhammadi

Parzen Window and Classification

- **The left one: a small h (complicated boundaries) - The right one: a larger h (simple boundaries)**
  - compare the upper and lower regions of two cases
  - small h is appropriate for the upper region, large h for the lower region
  - No single window width is ideal overall



Template designed by Jafar Muhammadi

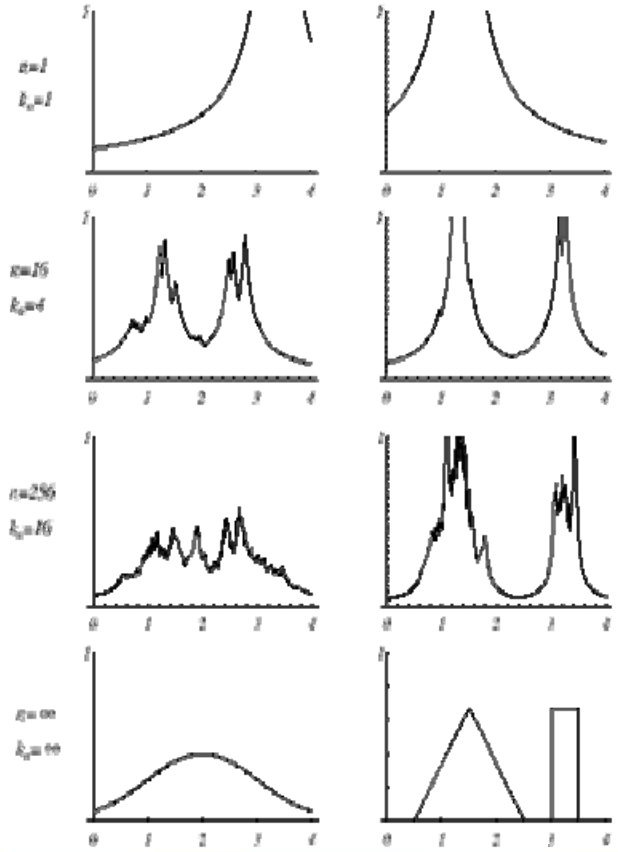
K_n-Nearest Neighbor

- **Goal: a solution for the problem of the unknown "best" window function**
  - **Let the cell volume be a function of the training data**
  - **Center a cell about x and let it grows until it captures k_n samples ($k_n = f(n)$)**
    - k_n samples are called the k_n nearest-neighbors of x
- **Two possibilities can occur:**
  - **Density is high near x; therefore the cell will be small which provides a good resolution**
  - **Density is low; therefore the cell will grow large and stop until higher density regions are reached**
- **We can obtain a family of estimates by setting $k_n=k_1/\sqrt{n}$ and choosing different values for k_1**

Template designed by Jafar Muhammadi

K-NN - Illustration

- For $k_n=\sqrt{n}$ and for n=1 the estimate becomes $P_n(x)=k_n/n$, $V_n= 1/V_1=1/2|x-x_1|$



Template designed by Jafar Muhammadi

K-NN and a-posteriori probabilities

- **Goal: estimate $P(w_i|x)$ from a set of n labeled samples**
  - **Let's place a cell of volume V around x and capture k samples**
  - **k_i samples amongst k turned out to be labeled w_i then**

$$P_n(X, w_i) = P_n(X|\, w_i) * P_n(w_i) = \frac{k_i}{n_i} \times \frac{n_i}{n} = \frac{k_i}{n}$$

  - **An estimate for $p_n(w_i|\, x)$ is:**

$$P_n(w_i \mid x) = \frac{P_n(x, w_i)}{\sum_{j=1}^{j=c} P_n(x, w_j)} = \frac{k_i}{k}$$

  - ***k_i/k* is the fraction of the samples within the cell that are labeled *w_i***
  - **For minimum error rate, the most frequently represented category within the cell is selected**
  - **If k is large and the cell sufficiently small, the performance will approach the best possible**

Template designed by Jafar Muhammadi

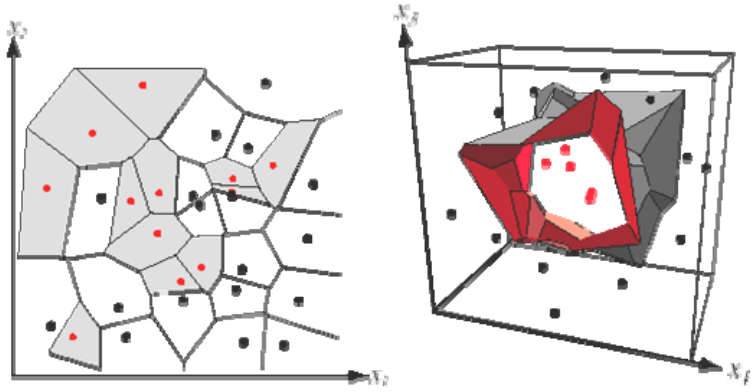
K-NN and Classification

- **The nearest neighbor Rule (K=1)**
  - **Let $D_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n labeled prototypes**
  - **Let $x' \in D_n$ be the closest prototype to a test point x then the nearest-neighbor rule for classifying x is to assign it the label associated with x'**
  - **The nearest-neighbor rule leads to an error rate greater than the minimum possible: the Bayes rate**
  - **If the number of prototype is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Bayes rate (it can be demonstrated!)**
    - Think more about it. It means that 50% of the information needed to optimally classify point x is aggregated within its nearest labeled neighbor.
  - **If $n \rightarrow \infty$, it is always possible to find x' sufficiently close so that $P(w_i \mid x') \cong P(w_i \mid x)$**
  - **If $P(w_m|x) \cong 1$, then the nearest neighbor selection is almost always the same as the Bayes selection**

Template designed by Jafar Muhammadi

K-NN and Classification

- **The nearest neighbor rule**
  - **In 2D the nearest neighbor leads to a partitioning of the input space into Voronoi cells**
  - **In 3D the cells are 3D and the decision boundary resembles the surface of a crystal**



Template designed by Jafar Muhammadi

Pros and Cons

- ✧ **No assumptions are needed about the distributions ahead of time (generality).**
- ✧ **With enough samples, convergence to an arbitrarily complicated target density can be obtained.**
- ✧ **The number of samples needed may be very large (number grows exponentially with the dimensionality of the feature space).**
- ✧ **These methods are very sensitive to the choice of window size (if too small , most of the volume will be empty, if too large, important variations may be lost).**
- ✧ **There may be severe requirements for computation time and storage.**

Template designed by Jafar Muhammadi

Any Question

End of Lecture 8

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/

Template designed by Jafar Muhammadi