# Computational Intelligence: Methods and Applications

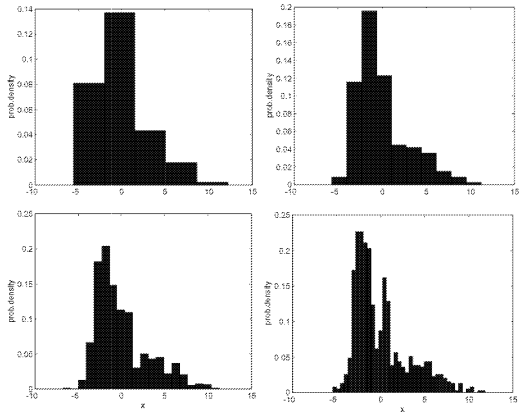
Lecture 28
Non-parametric density modeling

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

# Density from histograms

In 1-D or 2-D it is rather simple, histograms provide piecewise constant approximation – since we do not assume any particular functional dependence such estimation is called "nonparametric".

Histograms change, depending on the size of the bin $B_i$ that measures frequency $P(X \in B_i)$.




Smoothing histograms may be done by fitting some smooth functions, such as Gaussians.

How good is this approximation?

# Why histogram estimation works?

Probability that a data point comes from some region *R* (belongs to some category, etc) is:

$$P = \int_R P(\mathbf{X})\, d\mathbf{X}$$

We are given *n* data points, what is the chance Pr that *k* of these points are from region *R*? If *n*=*k*=1, this Pr=*P*, in general Pr is the number of combinations in which *k* points could be selected out of *n*, multiplied by probability of selecting *k* points from *R*, i.e. $P^k$, and selecting *n*-*k* points not from R, i.e. $(1-P)^{n-k}$, that is, the distribution is binomial:

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

Expected *k* value: $E(k) = nP$

Expected variance: $\sigma^2(k) = nP(1-P)$

Since $P(X)$ $V_R \cong P = k/n$, for a large number of samples *n* small variance of *k*/*n* is expected, therefore this is useful approximation to $P(X)$.

$$\sigma^2\left(\frac{k}{n}\right) = \frac{P}{n}(1-P)$$

# Parzen windows 1D

Density estimate using (for standardized data) a bin of size *h* (a window on the data) in each dimension.

For 1D cumulative density function CP($x$)=(# observation<x)/N

Density is given as a derivative of this function, estimated as:

$$P(x) = \frac{CP(x+h) - CP(x-h)}{2h}$$

For example, hyperrectangular windows with $H(u)=1$ for all $|u_j|<0.5$,

or hard sphere with 1 inside and 0 outside.

Number of points inside:

$$k = \sum_{i=1}^{n} H\left(\frac{\mathbf{X}^{(i)} - \mathbf{X}}{h}\right)$$

*h* is called a "smoothing" parameter

Density estimate:

$$P(\mathbf{X}) = \frac{k}{nV} = \frac{1}{nh^d} \sum_{i=1}^{n} H\left(\frac{\mathbf{X}^{(i)} - \mathbf{X}}{h}\right)$$

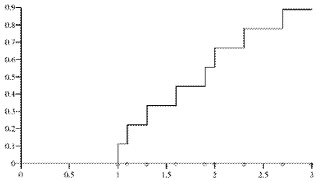
Kernel should be $H(u) \geq 0$ and should integrate to 1.

# Parzen windows 1D

Estimate density using (for standardized data) a bin of size *h* (a window on the data) in each dimension. For 1D cumulative density function is: $P(x<a)$=(# observation $x<a$)/$n$ $\leq 1$ (this is probability that $x<a$ ).

Cumulative contributions from all points should sum up to 1, contribution from each interval $[x-h/2, x+h/2]$ with a single observation $x_i$ inside is $1/n$.

For real data this is a stairway function.



Density is given as a derivative of this function, but for such staircase data it will be discontinuous, a series of spikes for $x=x_i$ values corresponding to real observations.

Numerical estimation of density at point $x$ is calculated as:

$$P(x=a)=\frac{P(x<a+h/2)-P(x\leq a-h/2)}{h}$$

# Parzen 1D kernels

We need continuous density estimation, not spikes.

Introduce a kernel function indicating if variable is in [−1,+1] interval:

$$H(u)=\begin{cases}0 & |u|>1\\ 1 & |u|\leq 1\end{cases}$$

Density may be now written as:

$$P(x)=\frac{1}{nh}\sum_{i=1}^{n}H\left(\frac{x_i-x}{h}\right)$$

Density in the window is constant=1, so integrating over each kernel:

$$\int_{-\infty}^{+\infty}H\left(\frac{x_i-x}{h}\right)dx=h$$

Integrating over all $x$ gives therefore total probability=1.

Smooth cumulative density for $x \leq a$ is then:

$$P(x\leq a)=\int_{-\infty}^{a}P(x)dx$$

This is equal to $1/n$ times the number of $x_i \leq a$ plus a fraction from the last interval $[x_i-h/2, a]$ if $a < x_i+h/2$

# Parzen windows dD

The window moves with *X* which is in the middle, therefore density is smoothed. 1D generalizes to dD situations easily:

Volume $V=h^d$ and the kernel (window) function:

$$H(\mathbf{u})=H\left(\frac{\mathbf{X}^{(i)}-\mathbf{X}}{h}\right)$$

Typically hyperrectangular windows with $H$(u)=1 for all $|u_j|<1$ are used, or hard sphere windows with 1 inside and 0 outside, or some other localized functions.

*h* is called a “smoothing“ parameter

Number of points inside:

$$k=\sum_{i=1}^{n}H\left(\frac{\mathbf{X}^{(i)}-\mathbf{X}}{h}\right)$$

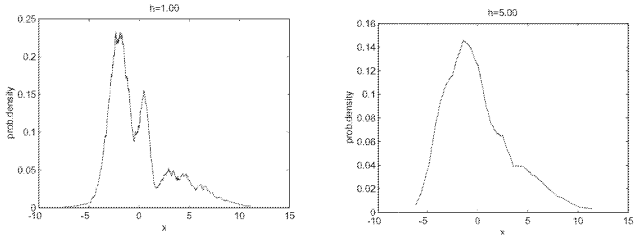
Density estimate:

$$P(\mathbf{X})=\frac{k}{nV}=\frac{1}{nh^d}\sum_{i=1}^{n}H\left(\frac{\mathbf{X}^{(i)}-\mathbf{X}}{h}\right)$$

Any function with $H$(u)≥0 integrating to 1 may be used as a kernel.

# Example with rectangles

With large *h* strong smoothing is achieved (imagine window covering all data ...)



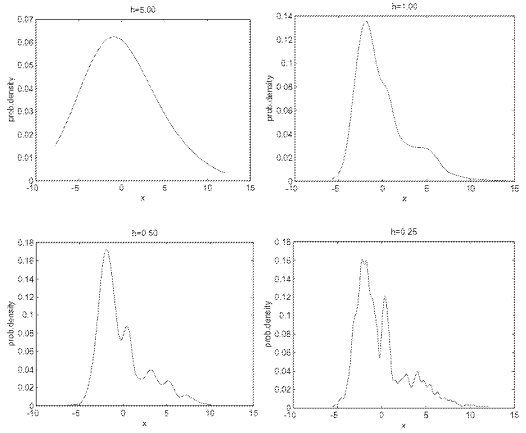

Details are picked up when *h* is small, general shape when it is large.

Use as $H$(u) a smooth function, such as Gaussian; if it is normalized than also the final density is normalized:

$$\int H(u)du=1 \Rightarrow \int P(x)dx=\frac{1}{nh}\sum_{i=1}^{n}\int H\left(\frac{x^{(i)}-x}{h}\right)dx=1$$

# Example with Gaussians

Dispersion *h* is also called here smoothing or regularization parameter.




A. Webb, Chapter 3.5 has a good explanation of Parzen windows.

# Idea

Assume that *P*(X) is a combination of some smooth functions Φ(X);

$$P(\mathbf{X}) = \sum_{i=1}^{m} W_i \Phi_i(\mathbf{X})$$

use an iterative algorithm that adapts the density to the incoming data.

Estimate density *P*(X|C) for each class separately.

Since calculation of parameters may be done on a network of independent processors, this leads to the basis set networks, such as the radial basis set networks.

This may be used for function approximation, classification and discovery of logical rules by covering algorithms.