ANLY-601
Advanced Pattern Recognition

Spring 2018

L16 --- Nonparametric Density Models (cont'd)

GEORGETOWN UNIVERSITY

K-NN Estimates

Parzen window (uniform kernel) -- kernel volume was fixed and we counted the number of samples falling inside the volume to estimate p(x).

K-nearest neighbor estimator, choose point x at which we estimate the density, and construct the smallest region L(x) that contains k points. Then estimate the density at x as

$$\hat{p}(x) = \frac{k-1}{N\,V(x)}$$

where N is the total number of points, V(x) the volume of the minimal n-dim region containing *k* points.

The numerator *k-1* gives the estimate lower bias than if it were *k*.

K-NN Estimates

Parzen window (uniform kernel) -- kernel volume was fixed and we counted the number of samples falling inside the volume to estimate p(x).

K-nearest neighbor estimator, choose point x at which we estimate the density, and construct the smallest region L(x) that contains k points. Then estimate the density at x as

$$\hat{p}(x) = \frac{k-1}{N\,V(x)}$$

If *r* is the distance from *x* to the kth nearest neighbor, then we can take *V(x)* to be the volume inside the n-sphere of radius *r*

$$V = \frac{\pi^{n/2}}{\Gamma\left(\frac{n+2}{2}\right)}\; r^n$$

where Γ is the Euler gamma function.

Bias and Variance of KNN

Bias and variance of $\hat{p}(x) = \dfrac{k-1}{N \; V(x)}$

$$E\left[\hat{p}(x)\right] = p(x)\left(1 + \tfrac{1}{2}\,\phi\left(\tfrac{k-1}{N}\right)^{2/n}\right)$$

where n is the dimension, and φ depends on the density and dimension.
The 2nd term is usually small, so $\hat{p}(x)$ is approx. unbiased.

$$\operatorname{var}\left[\hat{p}(x)\right] \approx \frac{p^2(x)}{k}$$

Notice that to decrease the variance, we increase the number of nearest neighbors *k* used. But this grows *V* and gives a coarser estimate of *p(x)* -- i.e. the bias increases. So we have a bias/variance trade-off to contend with.

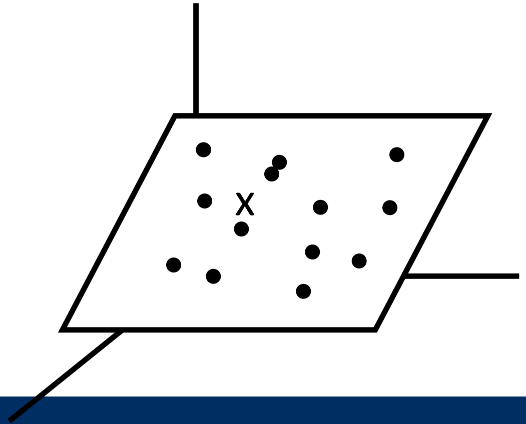
Source – Keinosuke Fukunaga, *Statistical Pattern Recognition,* 2nd Ed., Acad.Press

Conceptual Interlude: Intrinsic Dimensionality

Data handed to us in high-dimensional spaces may, in fact, actually lie near some lower dimensional sub-manifold. We can find the local dimension of this sub-manifold by using the k-NN density estimate.

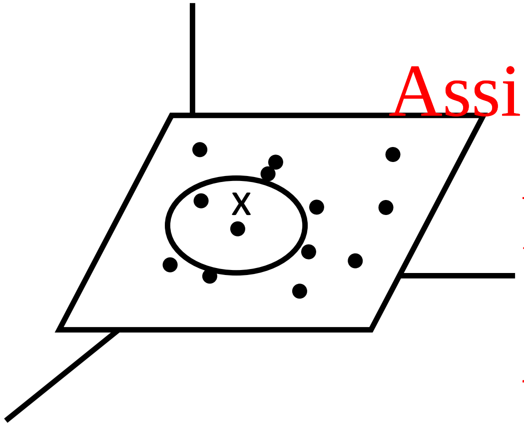
The idea is that the intrinsic dimension (of the manifold) dictates how the number of data points will increase as we increase the radius of the region L(x).

Consider data that lies on a plane in 3-space



Conceptual Interlude: Intrinsic Dimensionality

Consider data that lies on a plane in 3-space



If we erect a small sphere about the point x (open circle) and grow the radius of the sphere, the number of points inside will grow only as r^2 (the area in the plane intercepted by the sphere), not as r^3 as would be the case if the data filled 3-space rather than just the plane.

Intrinsic Dimensionality

We can derive the scaling of the number of points in the sphere with its radius on purely dimensional grounds. The volume of the n-sphere containing k points is

$$p \; V_k \cong \frac{k-1}{N} \quad \text{so} \quad \mathrm{V_k} \cong \frac{k-1}{N\;p}$$

The radius is related to the volume by $V_k = c \; r_k^{\,n_e}$ so $r_k = \left(\frac{V_k}{c}\right)^{1/n_e}$

So the radius of the sphere containing k points is

$$r_k = \left(\frac{V_k}{c}\right)^{1/n_e} = \left(\frac{k-1}{N\;p\;c}\right)^{1/n_e}$$

The radius containing k+1 points is

$$r_{k+1} = \left(\frac{k}{N\;p\;c}\right)^{1/n_e}$$

Consequently

$$\frac{r_{k+1}}{r_k} \approx \left(\frac{k}{k\;(1-1/k)}\right)^{1/n_e} \approx (1+1/k)^{1/n_e} \quad \text{(last equality for large } k\text{)}$$

and we can solve for the data dimensionality

Non-Parametric Methods ---
Expansion in Ortho-normal Basis Functions

Our last non-parametric technique is the use of orthogonal basis functions to represent the density. The model is

$$p(x) = \sum_{i=1}^{\infty} c_i \, \phi_i(x)$$

where the basis functions satisfy the orthogonality

$$\int_{-\infty}^{\infty} \phi_i(x) \, \phi_j(x) \, g(x) \, d^n x \;=\; \lambda_i \, \delta_{ij}$$

which provides solution for *c* $\quad c_i = \frac{1}{\lambda_i} \int_{-\infty}^{\infty} p(x) \, \phi_i(x) \, g(x) \, d^n x$

and completeness conditions $\quad \sum_{i=1}^{\infty} \frac{g(x') \, \phi_i(x') \, \phi_i(x)}{\lambda_i} \;=\; \delta(x - x')$

Completeness

Orthogonality is probably familiar to you, but completeness may not be. It is simply the statement that any function can be expanded in the basis – it's a complete set.

The analogs of

$$\int_{-\infty}^{\infty} \phi_i(x)\, \phi_j(x)\, g(x)\, d^n x \;=\; \lambda_i\, \delta_{ij}$$

$$\sum_{i=1}^{\infty} \frac{g(x')\, \phi_i(x')\, \phi_i(x)}{\lambda_i} \;=\; \delta(x - x')$$

for orthogonal, unit-norm, basis vectors e_i, $i=1..N$ in a finite-dimensional vector space are

$$e_i^T\, e_j \;=\; \delta_{ij}$$

$$\sum_{i=1}^{N} e_i\, e_i^T \;=\; 1 \equiv identity\ matrix$$

Error in Terminating the Series

We're obviously NOT going to use the whole infinite series, but rather will terminate it. The error incurred in terminating the series at *m* terms is

$$p(x) - p_m(x) = \sum_{i=1}^{\infty} c_i \ \phi_i(x) - \sum_{i=1}^{m} c_i \ \phi_i(x) = \sum_{i=m+1}^{\infty} c_i \ \phi_i(x)$$

a convenient integrated measure is

$$\int g(x) \left[p(x) - p_m(x)\right]^2 dx = \int g(x) \left[\sum_{i=m+1}^{\infty} c_i \ \phi_i(x)\right]\left[\sum_{j=m+1}^{\infty} c_j \ \phi_j(x)\right] dx$$

$$= \sum_{i=m+1}^{\infty} c_i^2 \ \lambda_i$$

so the best basis will have $c_i^2 \ \lambda_i$ drop off quickly with increasing i.

Example Basis Functions: Hermite Polynomials

A useful basis for densities close to Gaussian comes from Hermite polynomials $H_i(x)$ times Gaussians):

$$\varphi_i(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \; H_i(x)$$

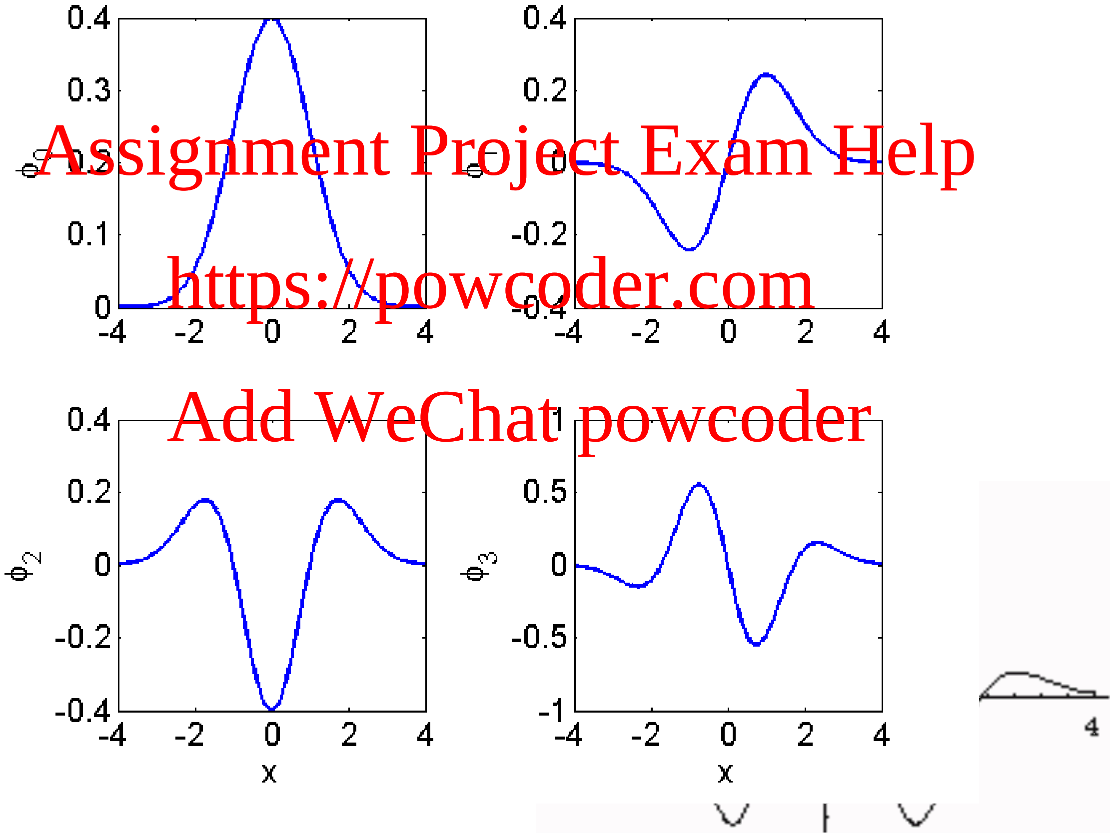
$$H_i(x) = (-\sigma)^i \; \exp\left(\frac{x^2}{2\sigma^2}\right) \; \frac{d^i}{dx^i} \left(\exp\left(-\frac{x^2}{2\sigma^2}\right)\right)$$

$$H_0(x) = 1 \qquad\qquad H_1(x) = x/\sigma$$

$$H_2(x) = -1 + (x/\sigma)^2 \qquad\qquad H_3(x) = (x/\sigma)^3 - 3\,x/\sigma$$

Hermite Polynomial Basis Functions

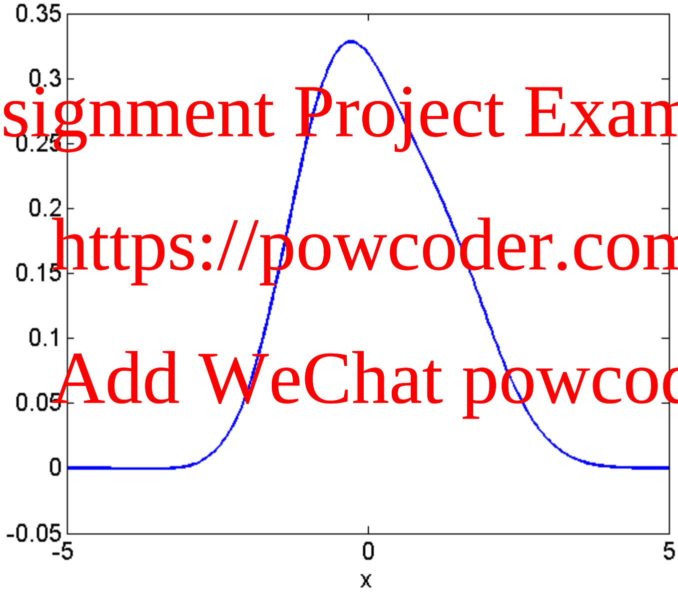
The first few basis functions ϕ_i look like this:



GEORGETOWN UNIVERSITY

Orthogonal Function Expansions

Example --- $p(x) = \varphi_0(x) + 0.15\varphi_1(x) + 0.2\varphi_2(x) + 0.15\varphi_3(x)$

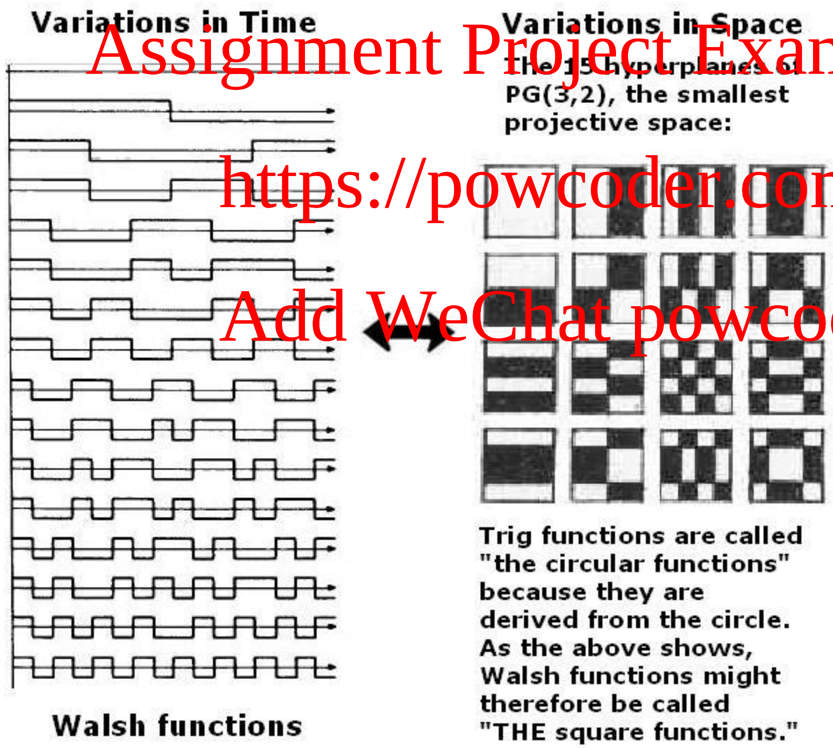


See Graham-Charlier and Edgeworth expansions of classical statistics – in e.g. Kendall & Stuart *The Advanced Theory of Statistics*

Binary Input Variables

For binary n-vectors --- You only need 2^n basis functions to represent the density without error.

One such basis are the *Walsh* functions that appear in digital image processing.



GEORGETOWN UNIVERSITY