System Identification

Density Estimation, Basis Function Approximation

Olaf Wolkenhauer

Control Systems Centre

UMIST

o.wolkenhauer@umist.ac.uk

www.csc.umist.ac.uk/people/wolkenhauer.htm

Contents

Learning Objectives

- The identification of a model is an approximation of the function which relates independent (e.g input-) and dependent (e.g output-) variables.
- Linear parametric regression, employing the least squares principle, is an efficient tool to identify parameters from data - to learn linear functional relationships.
- In a probabilistic framework data are assumed to be distributed according to some unknown probability density function.
- Statistical learning can be seen as a generalisation of density estimation.
- Like the Fourier series, Kernel density estimation provides another example of the approximation of an unknown function by means of so called basis functions.

1. Regression Models

Let

$$\mathbf{x} \doteq [x_1, \ldots, x_r]$$

denote a vector of independent variables taking values in X_1,.., X_r where we write $X \doteq X_1 \times \cdots \times X_r$ for short. Then a system is specified by

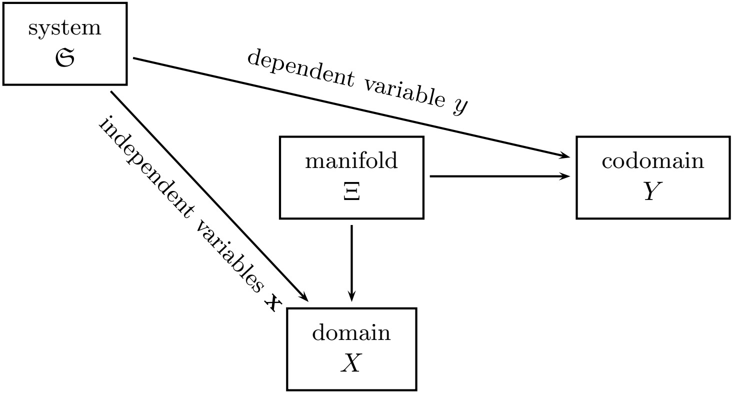
$$\begin{aligned} f : X &\rightarrow Y \\ \mathbf{x} &\mapsto y \ . \end{aligned}$$

The *identification* of a model $\mathfrak{M}$ is an approximation of $f\colon X \rightarrow Y$, based on a sampled set of training data, i.e measurements

$$\mathbf{m}_j = (\mathbf{x}_j, y_j) \qquad j = 1, 2, \ldots, d$$

The dependency between $\mathbf{x}$ and y is described by a *parameter vector* $\boldsymbol{\theta}$ such that

$$y \approx f(\mathbf{x}; \theta) \ .$$


system
$\mathfrak{S}$
dependent variable y
independent variables $\mathbf{x}$
manifold
Ξ
codomain
Y
domain
X

2. Linear Parametric Regression

The set of functions $f(\mathbf{x}, \theta)$ is specified as a polynomial of fixed degree,

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{r} \theta_i \cdot x_i$$

such that an appropriate set of θs can be found using least squares.

Examples: ARX, NARX Models.

Input-output "black-box" models, using an *auto-regressive* model structure :

$$\mathbf{x} \doteq [y(k), \ldots, y(k - n_y + 1), u(k), \ldots, u(k - n_u + 1)]^T \ .$$

ARX model structure :

$$y(k+1) = \sum_{i=1}^{n_y} \theta_i \cdot y(k-i+1) + \sum_{i=1}^{n_u} \theta_{n_y+i} \cdot u(k-i+1) \ .$$

This equation is called the *predictor* for model

$$y(k) = \sum_{i=1}^{n_y} a_i \cdot y(k-i) + \sum_{i=1}^{n_u} b_i \cdot u(k-i)$$

where

$$\boldsymbol{\theta} = [a_1, \cdots, a_{n_y}, b_1, \cdots, b_{n_u}]^T$$

and $r = n_y + n_u$.

NARX (Nonlinear AutoRegressive with eXogenous input) model :

$$\begin{aligned} y(k+1) &= f(\mathbf{x}, k) + \varepsilon(k) \\ &= f\big(y(k), \ldots, y(k-n_y+1), u(k), \ldots, u(k-n_u+1)\big) + \varepsilon(k) \end{aligned}$$

3. The Probabilistic Perspective

Random *input vectors*

$$\mathbf{x} \in \mathbb{R}^r \;\sim\; p(\mathbf{x}) \; .$$

Output values

$$y \sim p(y|\mathbf{x})$$

... unknown.

Training data

$$\mathbf{m}_j = (\mathbf{x}_j, y_j) \sim p(\mathbf{x}, y)$$

where

$$p(\mathbf{x}, y) = p(\mathbf{x}) \cdot p(y|\mathbf{x}) \; .$$

Find

$$f(\mathbf{x}) = \int y \cdot p(y|\mathbf{x}) \; \mathrm{d}y \tag{1}$$

such that

$$F = \big\{ \big(f(\mathbf{x}), \mathbf{x} \big) : f(\mathbf{x}) = (1) \big\} \; .$$

Using least squares, identify $f(\mathbf{x}; \theta)$ minimising the expected value of the *loss* :

$$\begin{aligned} E[L] &= \int L\big(y, f(\mathbf{x}; \boldsymbol{\theta})\big) \; p(\mathbf{x}, y) \; \mathrm{d}\mathbf{x}\mathrm{d}y \\ &\doteq R(\boldsymbol{\theta}) \end{aligned}$$

where $p(\mathbf{x}, y)$ is unknown.

... density estimation.

4. Kernel Density Estimation

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ be independent random variables identically distributed with cdf

$$\begin{aligned} F(x') &= Pr(\mathbf{x} \leq x') \\ &= \int_{-\infty}^{+\infty} p(x) \, \mathrm{d}x \,. \end{aligned} \tag{2}$$

Given training data $x_1, \ldots, x_d$, an empirical estimate of (2) is

$$\hat{F}(x') = \frac{1}{d} \sum_{j=1}^{d} \zeta(x_j \leq x') \tag{3}$$

where $\zeta(\cdot)$ is the indicator function. To estimate $p(x)$,

$$\hat{p}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2 \cdot h} \tag{4}$$

where h is a parameter.

Introducing the *kernel function* $K(\cdot)$, defined by

$$K(x') = \begin{cases} 0.5 & \text{if } |x'| \leq 1 \\ 0 & \text{if } |x'| > 1 \end{cases} ,$$

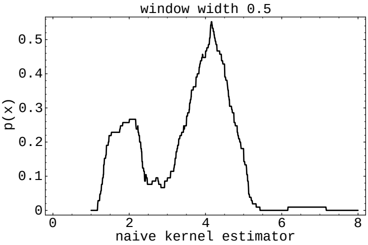
we can rewrite (4) as a weighted average over the sample distribution function :

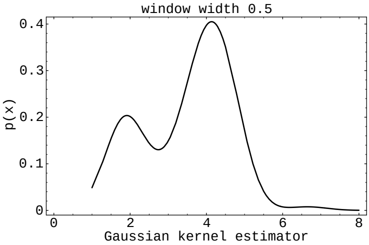
$$\begin{aligned} \hat{p}(x) &= \int_{-\infty}^{+\infty} \frac{1}{h} \, K\left(\frac{x - x'}{h}\right) \, \mathrm{d}\hat{F}(x') \\ &= \frac{1}{d \cdot h} \sum_{j=1}^{d} K\left(\frac{x - x_j}{h}\right) \; . \end{aligned} \tag{5}$$

Equation (5) is usually referred to as *kernel estimator*. A Gaussian kernel is frequently used :

$$K(x') = \frac{1}{\sqrt{2\pi}} \cdot e^{-0.5(x')^2} \; .$$

4.1. Example: Old Faithfull Data





5. Basis Function Approximation

The kernel density estimator (5)

$$\hat{p}(x) = \frac{1}{d \cdot h} \sum_{j=1}^{d} K\left(\frac{x - x_j}{h}\right)$$

suggests a general form for $f(\mathbf{x}; \theta)$, called *basis function approximation* :

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{c} \theta_i \cdot \phi_i(\mathbf{x})$$

5.1. Examples: Linear Regression, Fourier Series

Linear Regression

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{i=1}^{r} \theta_i \cdot x_i \ .$$

Fourier series :

$$\begin{aligned} f(t;\boldsymbol{\theta}) &= \sum_{i=1}^{r} \ \theta_i \cdot \phi_i(t) \ , \\ &\doteq \frac{a_0}{2} + \sum_{i=1}^{n_h} \left(a_i \cdot \cos(i \cdot \omega_0 \cdot t) + b_i \cdot \sin(i \cdot \omega_0 \cdot t)\right) \ . \end{aligned}$$

... and more to come.