Basis Functions

MLAI: Week 4

Neil D. Lawrence

Department of Computer Science
Sheffield University

20th October 2015

Review

- Last time: explored least squares for univariate and multivariate regression.
- Introduced matrices, linear algebra and derivatives.
- This time: introduce *basis functions* for non linear regression models.

Outline

Basis Functions

Basis Functions

Nonlinear Regression

- Problem with Linear Regression—$\mathbf{x}$ may not be linearly related to $\mathbf{y}$.
- Potential solution: create a feature space: define $\phi(\mathbf{x})$ where $\phi(\cdot)$ is a nonlinear function of $\mathbf{x}$.
- Model for target is a linear combination of these nonlinear functions

$$f(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}) \tag{1}$$

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

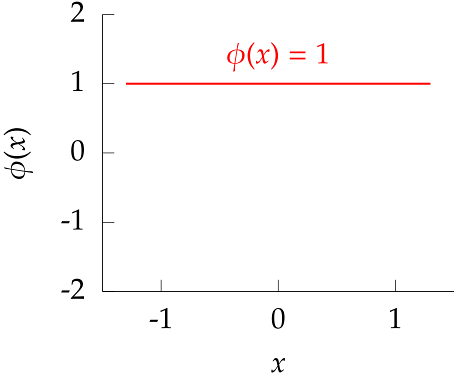


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

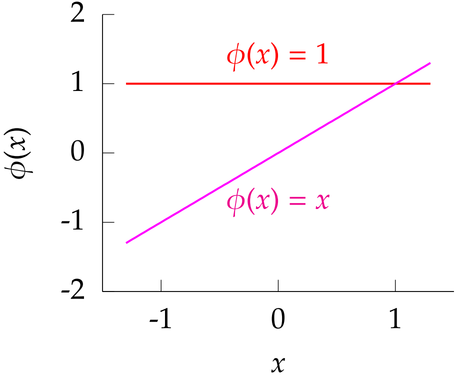


Figure: A quadratic basis.

Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

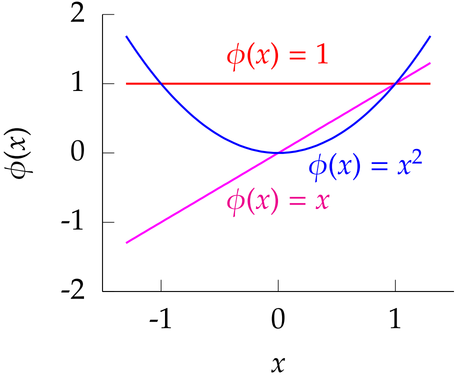


Figure: A quadratic basis.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

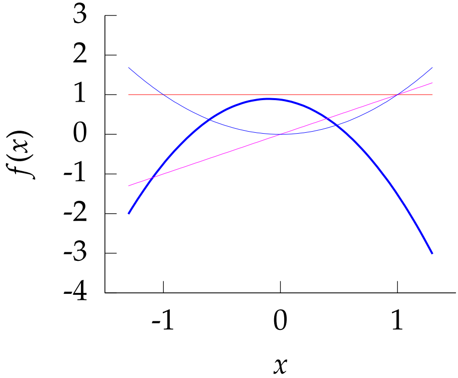


Figure: Function from quadratic basis with weights $w_1 = 0.87466$, $w_2 = -0.38835$, $w_3 = -2.0058$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

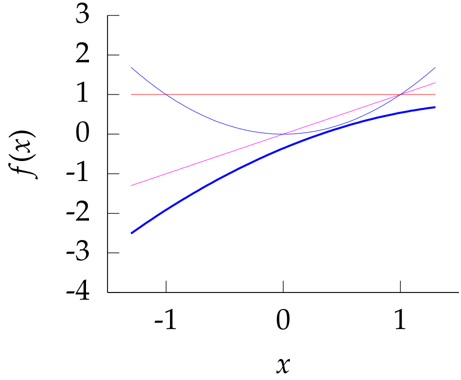


Figure: Function from quadratic basis with weights $w_1 = -0.35908$, $w_2 = 1.2274$, $w_3 = -0.32825$.

Functions Derived from Quadratic Basis

$$f(x) = w_1 + w_2 x + w_3 x^2$$

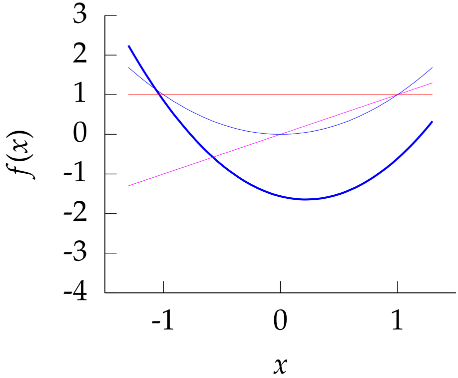


Figure: Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

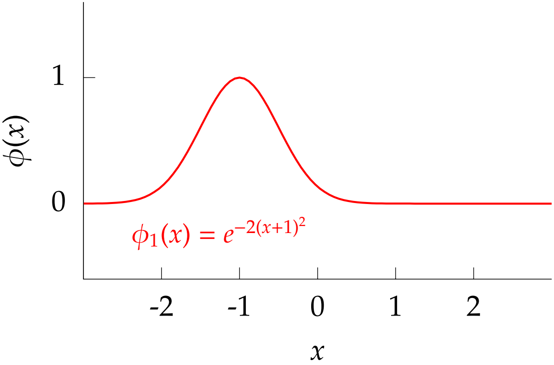


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

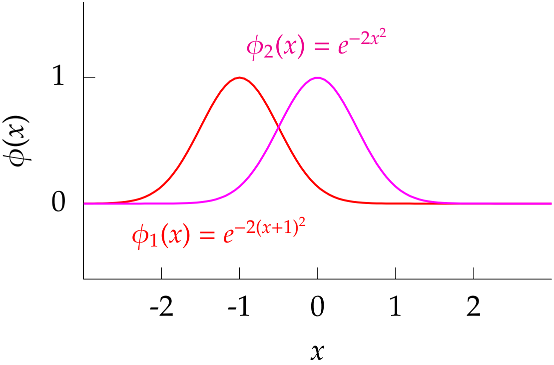


Figure: Radial basis functions.

Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

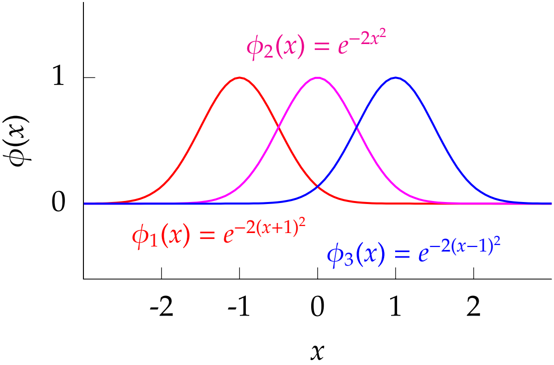


Figure: Radial basis functions.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

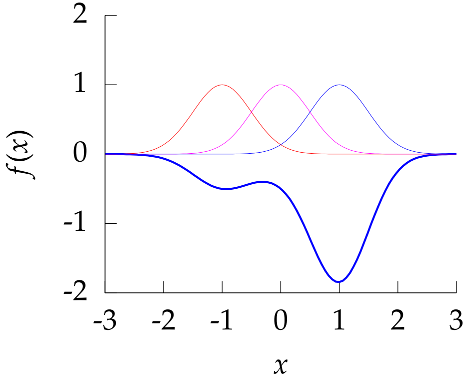


Figure: Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

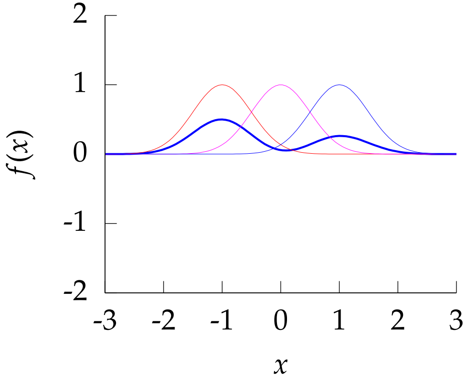


Figure: Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$.

Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

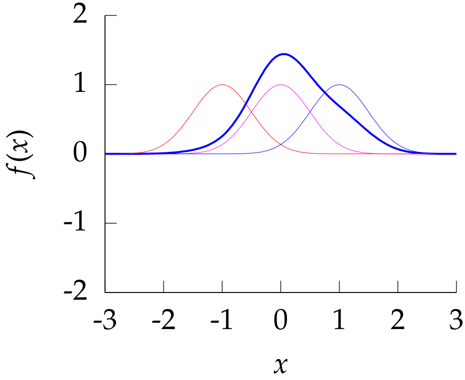


Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$.

Basis Function Models

- A Basis function mapping is now defined as

$$f(\mathbf{x}_i) = \sum_{j=1}^{m} w_j \phi_{i,j} + c$$

Vector Notation

- Write in vector notation,

$$f(\mathbf{x}_i) = \mathbf{w}^\top \boldsymbol{\phi}_i + c$$

Log Likelihood for Basis Function Model

- The likelihood of a single data point is

$$p\left(y_i|x_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)^2}{2\sigma^2}\right).$$

- Leading to a log likelihood for the data set of

$$L(\mathbf{w}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{n}{2} \log 2\pi - \frac{\sum_{i=1}^{n} \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)^2}{2\sigma^2}.$$

- And a corresponding error function of

$$E(\mathbf{w}, \sigma^2) = \frac{n}{2} \log \sigma^2 + \frac{\sum_{i=1}^{n} \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)^2}{2\sigma^2}.$$

Expand the Brackets

$$
\begin{aligned}
E(\mathbf{w}, \sigma^2) =& \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i \mathbf{w}^\top \boldsymbol{\phi}_i \\
&+ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \mathbf{w}^\top \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \mathbf{w} + \text{const.} \\
=& \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{1}{\sigma^2} \mathbf{w}^\top \sum_{i=1}^{n} \boldsymbol{\phi}_i y_i \\
&+ \frac{1}{2\sigma^2} \mathbf{w}^\top \left[\sum_{i=1}^{n} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \right] \mathbf{w} + \text{const.}
\end{aligned}
$$

Multivariate Derivatives Reminder

- We will need some multivariate calculus.

$$\frac{\mathrm{d}\mathbf{a}^\top\mathbf{w}}{\mathrm{d}\mathbf{w}} = \mathbf{a}$$

and

$$\frac{\mathrm{d}\mathbf{w}^\top\mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = \left(\mathbf{A} + \mathbf{A}^\top\right)\mathbf{w}$$

or if $\mathbf{A}$ is symmetric (*i.e.* $\mathbf{A} = \mathbf{A}^\top$)

$$\frac{\mathrm{d}\mathbf{w}^\top\mathbf{A}\mathbf{w}}{\mathrm{d}\mathbf{w}} = 2\mathbf{A}\mathbf{w}.$$

Differentiate

Differentiating with respect to the vector $\mathbf{w}$ we obtain

$$\frac{\partial L(\mathbf{w}, \beta)}{\partial \mathbf{w}} = \beta \sum_{i=1}^{n} \phi_i y_i - \beta \left[\sum_{i=1}^{n} \phi_i \phi_i^\top \right] \mathbf{w}$$

Leading to

$$\mathbf{w}^* = \left[\sum_{i=1}^{n} \phi_i \phi_i^\top \right]^{-1} \sum_{i=1}^{n} \phi_i y_i,$$

Rewrite in matrix notation:

$$\sum_{i=1}^{n} \phi_i \phi_i^\top = \mathbf{\Phi}^\top \mathbf{\Phi}$$

$$\sum_{i=1}^{n} \phi_i y_i = \mathbf{\Phi}^\top \mathbf{y}$$

Update Equations

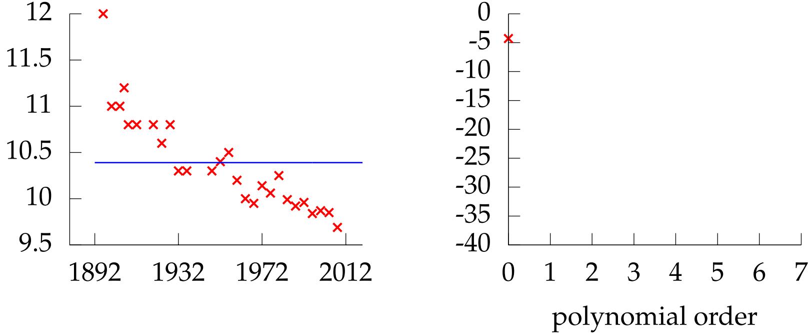
- Update for $\mathbf{w}^*$.

$$\mathbf{w}^* = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

- The equation for ${\sigma^2}^*$ may also be found

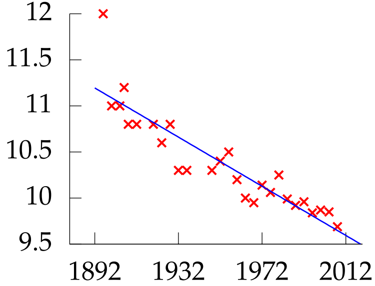
$${\sigma^2}^* = \frac{\sum_{i=1}^{n} \left(y_i - {\mathbf{w}^*}^\top \boldsymbol{\phi}_i\right)^2}{n}.$$

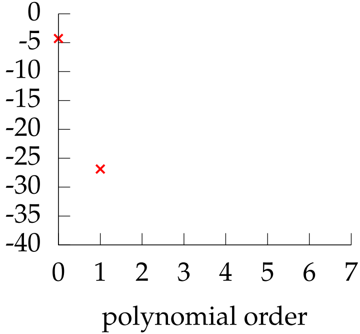
Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 0, model error -4.2717, $\sigma^2 = 0.268$, $\sigma = 0.518$.

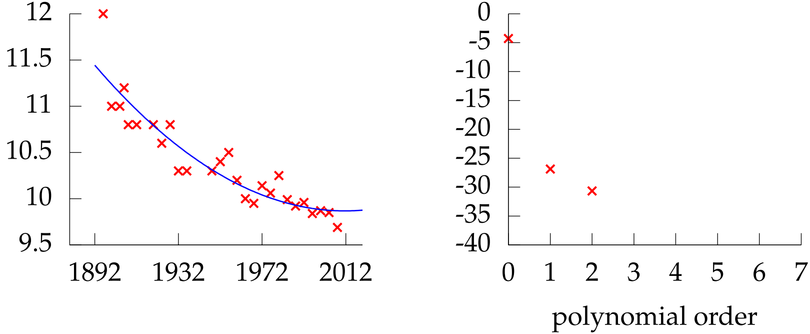
Polynomial Fits to Olympics Data





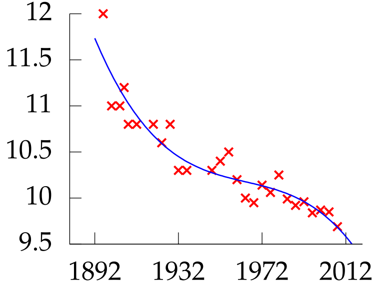
Left: fit to data, *Right*: model error. Polynomial order 1, model error -26.86, $\sigma^2 = 0.0503$, $\sigma = 0.224$.

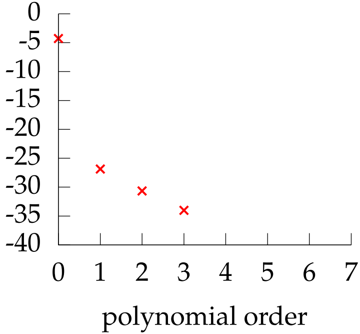
Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 2, model error -30.662, $\sigma^2 = 0.0380$, $\sigma = 0.195$.

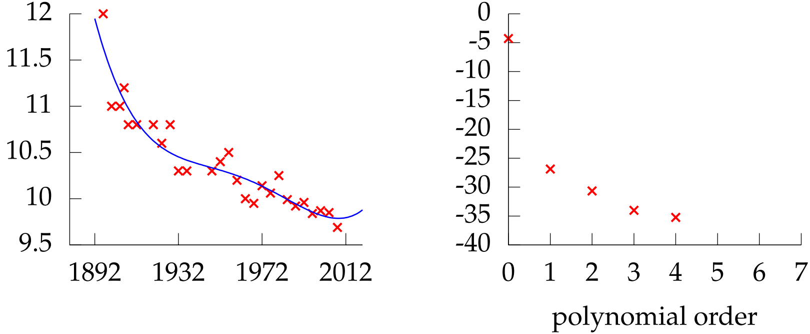
Polynomial Fits to Olympics Data





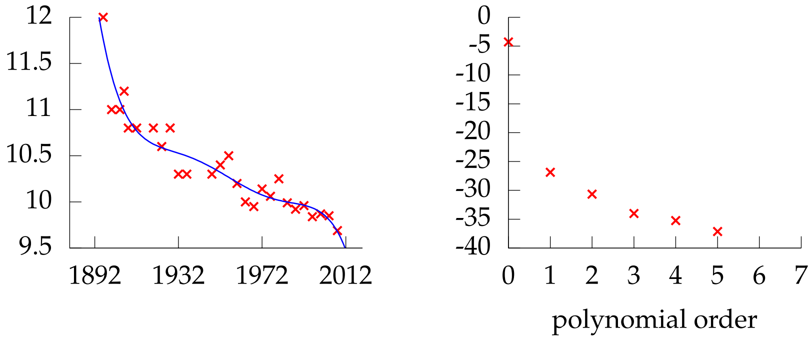
Left: fit to data, *Right*: model error. Polynomial order 3, model error -34.015, $\sigma^2 = 0.0296$, $\sigma = 0.172$.

Polynomial Fits to Olympics Data



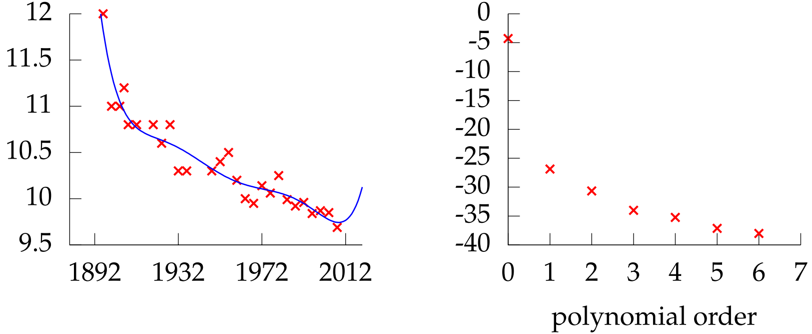
Left: fit to data, *Right*: model error. Polynomial order 4, model error -35.231, $\sigma^2 = 0.0271$, $\sigma = 0.165$.

Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 5, model error -37.138, $\sigma^2 = 0.0235$, $\sigma = 0.153$.

Polynomial Fits to Olympics Data



Left: fit to data, *Right*: model error. Polynomial order 6, model error -38.016, $\sigma^2 = 0.0220$, $\sigma = 0.148$.

Reading

- Section 1.4 of Rogers and Girolami.
- Chapter 1, pg 1-6 of Bishop.
- Chapter 3, Section 3.1 of Bishop up to pg 143.

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .