Regression

Reading: Chapter 8.1-8.4 Prince Book

Outline

- Linear regression
- Bayesian regression
- Non-linear regression
- Kernel “trick”
- Gaussian process regression

Linear Regression

We have one equation for each x,w training pair:

$$Pr(w_i|\mathbf{x}_i,\boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i,\sigma^2\right]$$

Joint Likelihood over whole training dataset

$$Pr(\mathbf{w}|\mathbf{X}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi},\sigma^2\mathbf{I}]$$

where

$$\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2\ldots\mathbf{x}_N] \quad \mathbf{w} = [w_1,w_2,\ldots,w_N]^T$$

Linear Regression

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta})\right] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\log Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta})\right]$$

Substituting in

$$\hat{\phi},\hat{\sigma}^2 = \underset{\phi,\sigma^2}{\operatorname{argmax}}\left[-\frac{N\log[2\pi]}{2} - \frac{N\log[\sigma^2]}{2} - \frac{(\mathbf{w}-\mathbf{X}^T\phi)^T(\mathbf{w}-\mathbf{X}^T\phi)}{2\sigma^2}\right]$$

Take derivative, set result to zero and re-arrange:

$$\begin{aligned}\hat{\phi} &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w}\\ \hat{\sigma}^2 &= \frac{(\mathbf{w}-\mathbf{X}^T\phi)^T(\mathbf{w}-\mathbf{X}^T\phi)}{N}\end{aligned}$$

Bayesian Regression

Likelihood

$$Pr(\mathbf{w}|\mathbf{X}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}]$$

Prior

$$Pr(\phi) = \text{Norm}_{\phi}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

encourages values of ϕ to be small

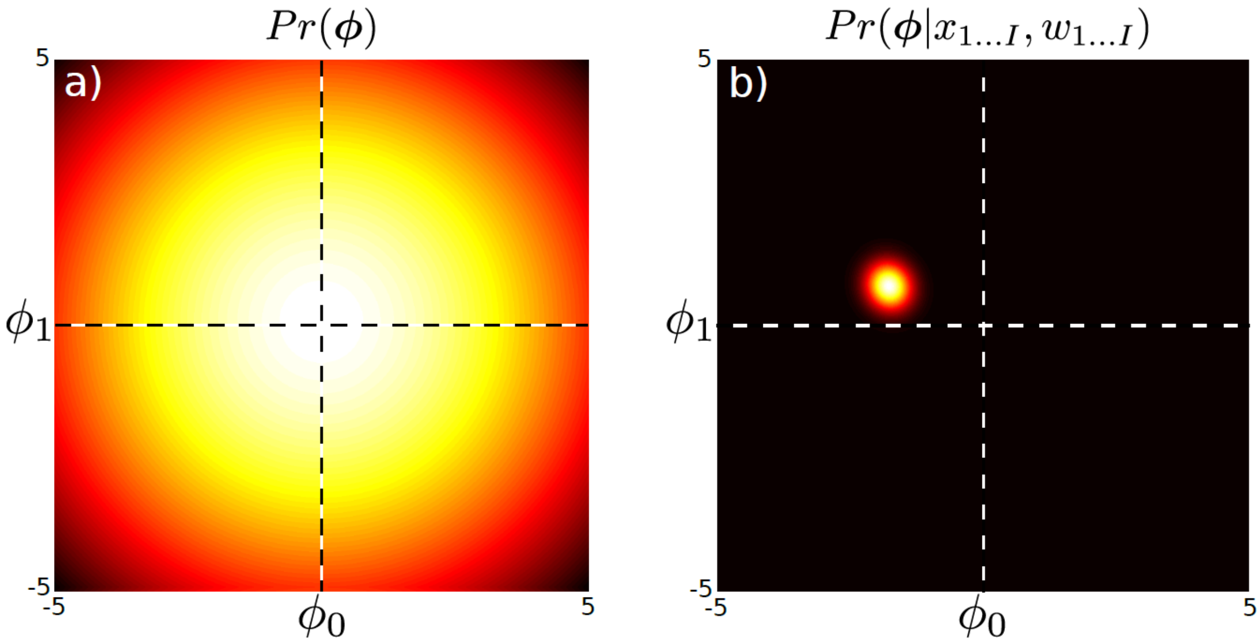
Bayes rule'

$$Pr(\phi|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \phi)Pr(\phi)}{Pr(\mathbf{w}|\mathbf{X})}$$

Bayesian Regression

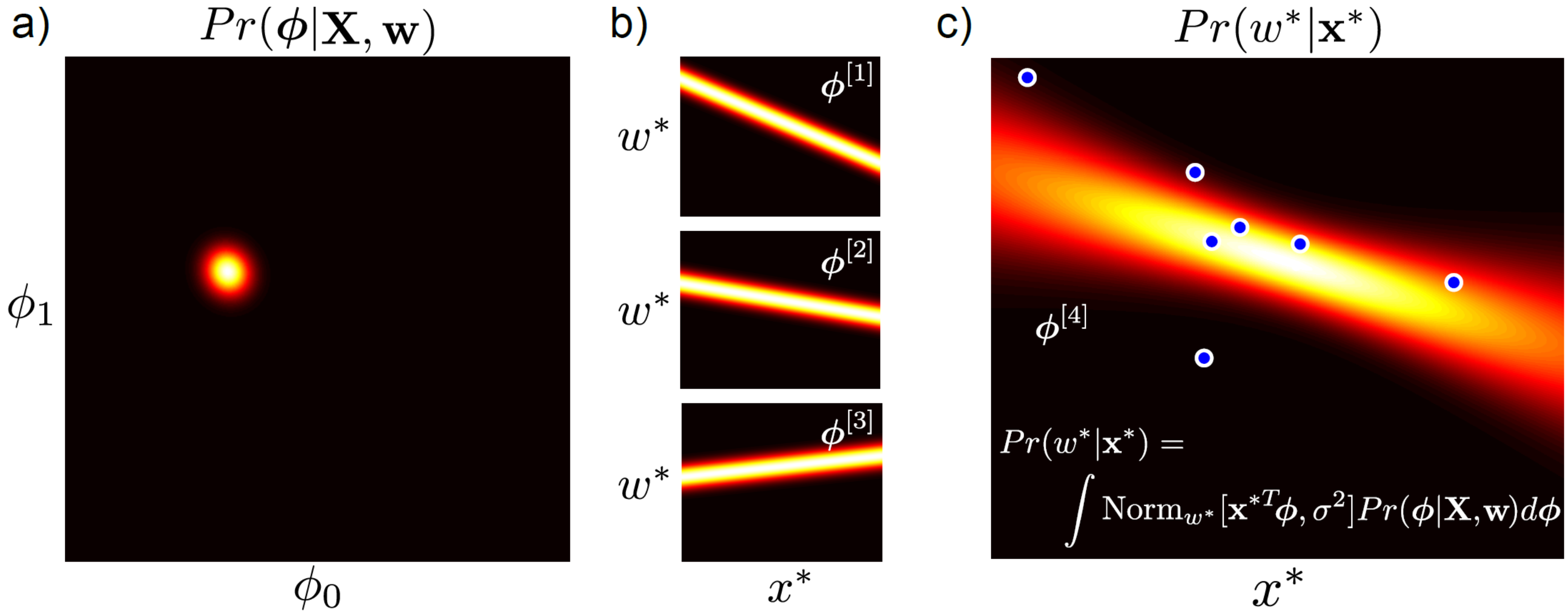
$$Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w}) = \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right]$$

where $$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}$$



Bayesian Regression

$$
\begin{aligned}
Pr(w^*|\mathbf{x}^*,\mathbf{X},\mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*,\boldsymbol{\phi})Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w})d\boldsymbol{\phi} \\
&= \int \mathrm{Norm}_{w^*}[\boldsymbol{\phi}^T\mathbf{x}^*,\sigma^2]\mathrm{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w},\mathbf{A}^{-1}\right]d\boldsymbol{\phi} \\
&= \mathrm{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w},\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^*+\sigma^2\right].
\end{aligned}
$$



Non-Linear Regression

GOAL:

Keep the math of linear regression, but extend to more general functions

KEY IDEA:

You can make a non-linear function from a linear weighted sum of non-linear basis functions

Non-linear regression

Linear regression:

$$Pr(w_i|\mathbf{x}_i,\boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i,\sigma^2\right]$$

Non-Linear regression:

$$Pr(w_i|\mathbf{x}_i,\boldsymbol{\theta}) = \text{Norm}_{w_i}[\boldsymbol{\phi}^T\mathbf{z}_i,\sigma^2]$$

where $\mathbf{z}_i = \mathbf{f}[\mathbf{x}_i]$

In other words, create z by evaluating x against basis functions, then linearly regress against z.

Example: polynomial regression

$$Pr(w_i|x_i) = \text{Norm}_{w_i}[\phi_0 + \phi_1 x_i + \phi_2 x_i^2 + \phi_3 x_i^3, \sigma^2].$$

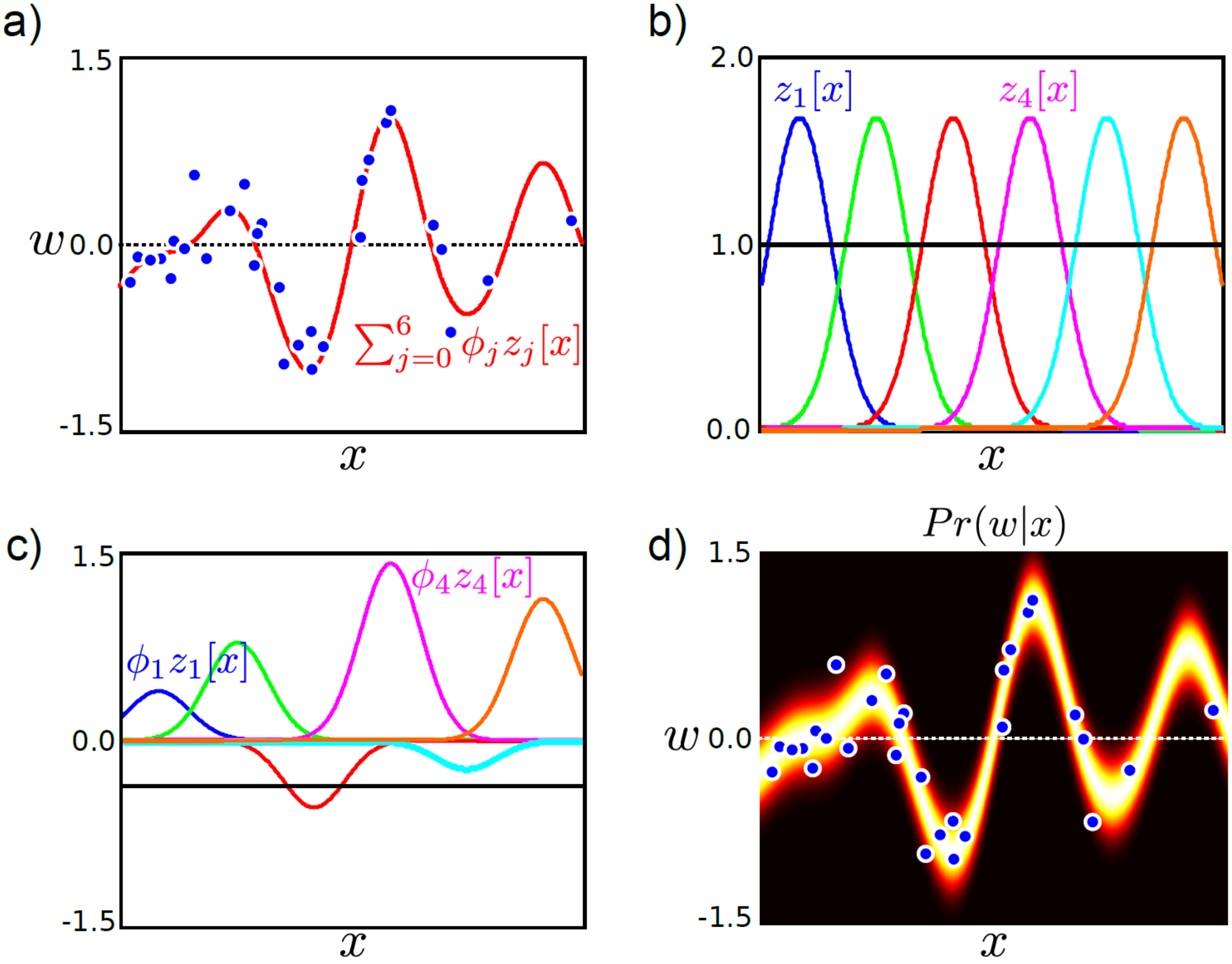
A special case of

$$Pr(w_i|\mathbf{x}_i) = \text{Norm}_{w_i}[\boldsymbol{\phi}^T \mathbf{z}_i, \sigma^2]$$

Where

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ x_i^3 \end{bmatrix}$$

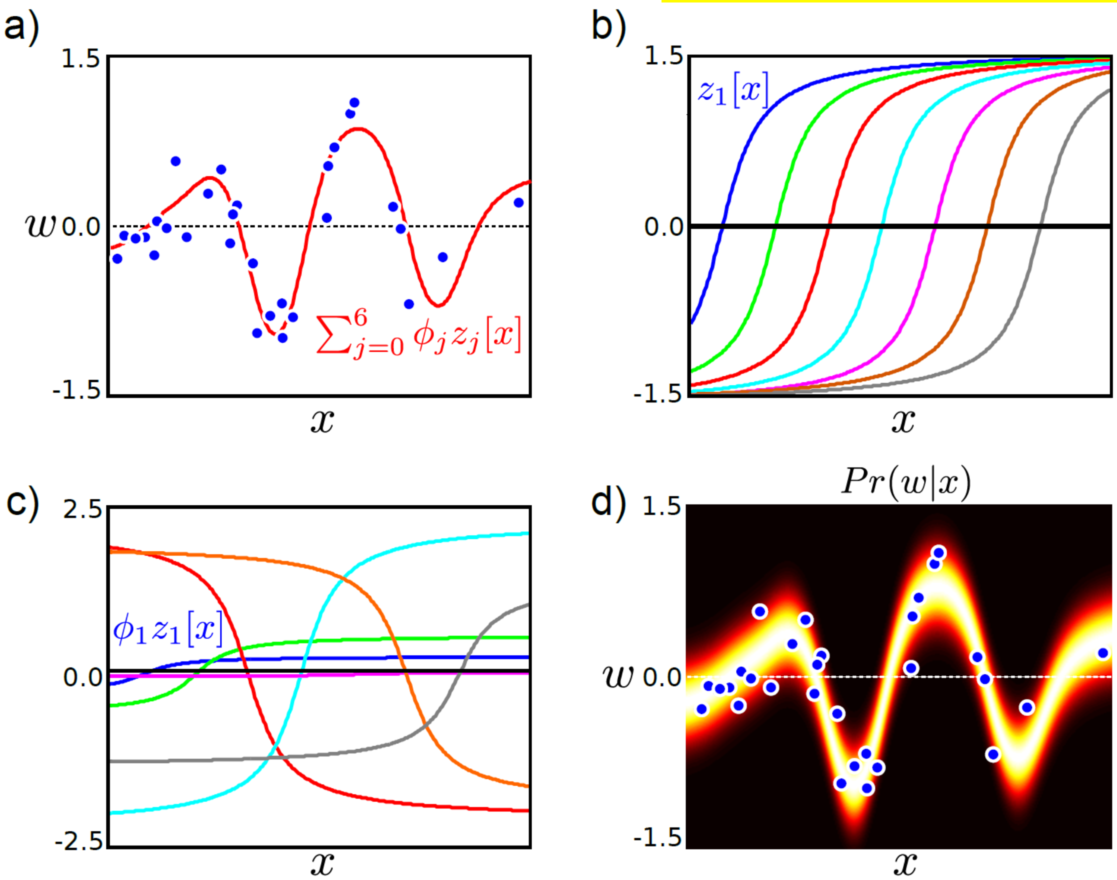
Radial basis functions



$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \exp\left[-(x_i-\alpha_1)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_2)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_3)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_4)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_5)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_6)^2/\lambda\right] \end{bmatrix}$$

Arc Tan Functions

note:sigmoid-like functions



$$
\mathbf{z}_i = \begin{bmatrix} \arctan[\lambda x_i - \alpha_1] \\ \arctan[\lambda x_i - \alpha_2] \\ \arctan[\lambda x_i - \alpha_3] \\ \arctan[\lambda x_i - \alpha_4] \\ \arctan[\lambda x_i - \alpha_5] \\ \arctan[\lambda x_i - \alpha_6] \\ \arctan[\lambda x_i - \alpha_7] \end{bmatrix}
$$

Nonlinear Regression

Maximum likelihood estimation for nonlinear regression is the same as linear regression, but substitute in **Z** for **X**:

$$
\begin{aligned}
\hat{\phi} &= (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{w} \\
\hat{\sigma}^2 &= \frac{(\mathbf{w}-\mathbf{Z}^T\phi)^T(\mathbf{w}-\mathbf{Z}^T\phi)}{N}
\end{aligned}
$$

Bayesian Nonlinear Regression

Bayesian nonlinear regression likewise proceeds similarly to Bayesian linear regression, using features Z instead of original data X

Through suitable rewriting of the matrix inverse equations, we can write the predictive distribution as

$$
\begin{aligned}
Pr(w^*|\mathbf{z}^*, \mathbf{X}, \mathbf{w}) = \\
&\text{Norm}_w\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w},\right. \\
&\left.\sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2\right]
\end{aligned}
$$

The Kernel Trick

Notice that the final equation doesn’t need the data itself, but just dot products between data items of the form $\mathbf{z}_i^T\mathbf{z}_j$

$$
\begin{aligned}
&Pr(w^*|\mathbf{z}^*, \mathbf{X}, \mathbf{w}) = \\
&\quad \mathrm{Norm}_w\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w},\right. \\
&\qquad\qquad \left.\sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2\right]
\end{aligned}
$$

So, we take data $\mathbf{x}_i$ and $\mathbf{x}_j$ pass through non-linear function to create $\mathbf{z}_i$ and $\mathbf{z}_j$ and then take dot products of different $\mathbf{z}_i^T\mathbf{z}_j$

The Kernel Trick

So, we take data $\mathbf{x}_i$ and $\mathbf{x}_j$ pass through non-linear function to create $\mathbf{z}_i$ and $\mathbf{z}_j$ and then take dot products of different $\mathbf{z}_i^T\mathbf{z}_j$

Key idea:

Define a “kernel” function that does all of this together.

- Takes data $\mathbf{x}_i$ and $\mathbf{x}_j$
- Returns a value for dot product $\mathbf{z}_i^T\mathbf{z}_j$

If we choose this function carefully, then it will correspond to some underlying $\mathbf{z}=\mathbf{f}[\mathbf{x}]$.

Never compute $\mathbf{z}$ explicitly - can be very high or infinite dimension

Kernelized Regression

Before

$$
\begin{aligned}
Pr(w^*|\mathbf{z}^*, \mathbf{X}, \mathbf{w}) &= \\
&\text{Norm}_w\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w},\right. \\
&\left.\sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2\right]
\end{aligned}
$$

After

$$
\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \\
&\text{Norm}_{w^*}\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\left(\mathbf{K}[\mathbf{X}, \mathbf{X}] + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X}, \mathbf{X}]\mathbf{w},\right. \\
&\left.\sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{x}^*] - \sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\left(\mathbf{K}[\mathbf{X}, \mathbf{X}] + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X}, \mathbf{x}^*] + \sigma^2\right]
\end{aligned}
$$

where the notation $\mathbf{K}[\mathbf{X}, \mathbf{X}]$ represents a matrix of dot products where element (i, j) is given by $\text{k}[\mathbf{x}_i, \mathbf{x}_j]$.

Example Kernels

- linear $\mathrm{k}[\mathbf{x}_i, \mathbf{x}_j] = \mathbf{x}_i^T \mathbf{x}_j,$

- degree p polynomial $\mathrm{k}[\mathbf{x}_i, \mathbf{x}_j] = (\mathbf{x}_i^T \mathbf{x}_j + 1)^p,$

- radial basis function (RBF) or Gaussian

$$\mathrm{k}[\mathbf{x}_i, \mathbf{x}_j] = \exp\left[-0.5\left(\frac{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}{\lambda^2}\right)\right].$$

(Equivalent to having an infinite number of radial basis functions at every position in space. You would not be able to explicitly generate z_i=f(x_i) for this, must use kernel function.)

Gaussian Process Regression

- Bayesian nonlinear regression using kernels!

Gaussian Process Regression

Bayesian nonlinear regression

$$Pr(w^*|\mathbf{z}^*,\mathbf{X},\mathbf{w}) =$$

$$\text{Norm}_w\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w},\right.$$

$$\left.\sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2\right]$$

Gaussian Process regression

$$Pr(w^*|\mathbf{x}^*,\mathbf{X},\mathbf{w}) =$$

$$\text{Norm}_{w^*}\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*,\mathbf{X}]\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*,\mathbf{X}]\left(\mathbf{K}[\mathbf{X},\mathbf{X}]+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X},\mathbf{X}]\mathbf{w},\right.$$

$$\left.\sigma_p^2\mathbf{K}[\mathbf{x}^*,\mathbf{x}^*] - \sigma_p^2\mathbf{K}[\mathbf{x}^*,\mathbf{X}]\left(\mathbf{K}[\mathbf{X},\mathbf{X}]+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X},\mathbf{x}^*] + \sigma^2\right]$$

where the notation $\mathbf{K}[\mathbf{X},\mathbf{X}]$ represents a matrix of dot products where element (i,j) is given by $\text{k}[\mathbf{x}_i,\mathbf{x}_j]$.

RBF Kernel Fits

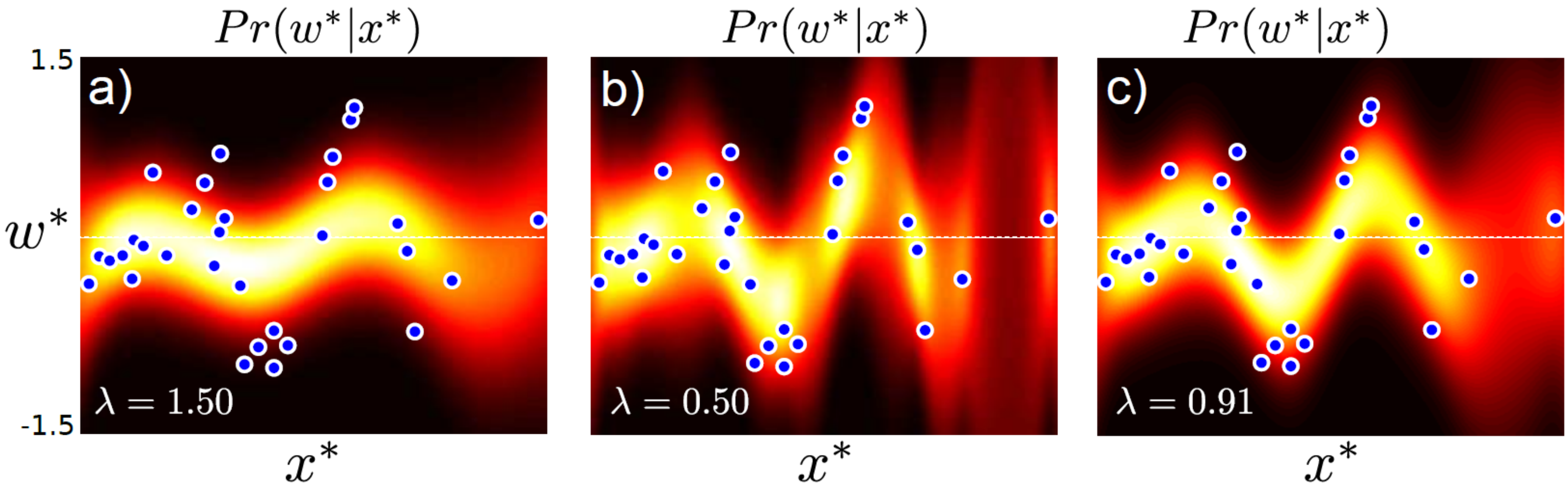


Figure 8.9 Gaussian process regression using an RBF kernel a) When the length scale parameter λ is large, the function is too smooth. b) For small values of the length parameter the model does not successfully interpolate between the examples. c) The regression using the maximum likelihood length scale parameter is neither too smooth nor disjointed.

$$k[\mathbf{x}_i, \mathbf{x}_j] = \exp\left[-0.5\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\lambda}\right)^2\right]$$