Statistical Machine Learning
Assignment 3

Inez Wijnands (s4149696) & Guido Zuidhof (s4160703)
Radboud University Nijmegen
03/12/2015

The entire code listing is included in the zip-file. The listings shown here are merely code snippets.

1 Bayesian linear regression

1.

$$\{x_1, t_1\} = (0.4, 0.05) \tag{1.1}$$
$$\{x_2, t_2\} = (0.6, -0.35) \tag{1.2}$$
$$\alpha = 2 \tag{1.3}$$
$$\beta = 10 \tag{1.4}$$
$$\tag{1.5}$$

We need to compute the predictive distribution $p(t|x, \boldsymbol{t}, \boldsymbol{x}, \alpha, \beta)$ after the two data points (where x is the input and t is the target) are observed.

Step 1: Identify the vector of basis functions $\phi(\boldsymbol{x})$, and write out $\boldsymbol{\Phi}^T \boldsymbol{t}$ and $\boldsymbol{\Phi}^T \boldsymbol{\Phi}$ in terms of the data $\{x_n, t_n\}$.

$$\boldsymbol{\Phi}^T \boldsymbol{t} = \sum_n \boldsymbol{\phi}(x_n) t_n = N \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \tag{1.6}$$

$$\boldsymbol{\Phi}^T \boldsymbol{\Phi} = \sum_n \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^T = N \begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix} \tag{1.7}$$

Where:

$$\bar{\mu}_t = \frac{1}{N} \sum_n t_n \qquad \bar{\mu}_{xt} = \frac{1}{N} \sum_n x_n t_n$$
$$\bar{\mu}_x = \frac{1}{N} \sum_n x_n \qquad \bar{\mu}_{xx} = \frac{1}{N} \sum_n x_n^2$$

Combined:

$$\boldsymbol{\Phi}^T \boldsymbol{t} = \sum_n \boldsymbol{\phi}(x_n) t_n = N \begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \tag{1.8}$$
$$= N \begin{pmatrix} \frac{1}{N} \sum_n t_n \\ \frac{1}{N} \sum_n x_n t_n \end{pmatrix} \tag{1.9}$$
$$= 2 \begin{pmatrix} \frac{1}{2}(0.05 + -0.35) \\ \frac{1}{2}(0.4 \cdot 0.05 + 0.6 \cdot -0.35) \end{pmatrix} \tag{1.10}$$
$$= \begin{pmatrix} -0.3 \\ -0.19 \end{pmatrix} \tag{1.11}$$

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \sum_n \phi(x_n)\, t_n = N\begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix} \tag{1.12}$$

$$= N\begin{pmatrix} 1 & \frac{1}{N}\sum_n x_n \\ \frac{1}{N}\sum_n x_n & \frac{1}{N}\sum_n x_n^2 \end{pmatrix} \tag{1.13}$$

$$= 2\begin{pmatrix} 1 & \frac{1}{2}(0.4+0.6) \\ \frac{1}{2}(0.4+0.6) & \frac{1}{2}(0.4^2+0.6^2) \end{pmatrix} \tag{1.14}$$

$$= \begin{pmatrix} 2 & 1 \\ 1 & 0.52 \end{pmatrix} \tag{1.15}$$

Step 2: Compute the posterior $p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{x},\alpha,\beta)$.

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0,\alpha^{-1}\boldsymbol{I}) \tag{1.16}$$

$$p(\boldsymbol{t}|\boldsymbol{w},\boldsymbol{x}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{\Phi}\boldsymbol{w},\beta^{-1}\boldsymbol{I}) \tag{1.17}$$

$$\rightarrow \tag{1.18}$$

$$p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{x}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N) \tag{1.19}$$

$$\boldsymbol{S}_N^{-1} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} + N\beta\begin{pmatrix} 1 & \bar{\mu}_x \\ \bar{\mu}_x & \bar{\mu}_{xx} \end{pmatrix} \tag{1.20}$$

$$= \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + 10\begin{pmatrix} 2 & 1 \\ 1 & 0.52 \end{pmatrix} \tag{1.21}$$

$$= \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 20 & 10 \\ 10 & 5.2 \end{pmatrix} \tag{1.22}$$

$$= \begin{pmatrix} 22 & 10 \\ 10 & 7.2 \end{pmatrix} \tag{1.23}$$

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t} = N\beta\boldsymbol{S}_N\begin{pmatrix} \bar{\mu}_t \\ \bar{\mu}_{xt} \end{pmatrix} \tag{1.24}$$

$$= 10\begin{pmatrix} 22 & 10 \\ 10 & 7.2 \end{pmatrix}^{-1}\begin{pmatrix} -0.3 \\ -0.19 \end{pmatrix} \tag{1.25}$$

$$= \begin{pmatrix} -0.0445 \\ -0.2021 \end{pmatrix} \tag{1.26}$$

Step 3: Compute the predictive distribution $p(t|x,\boldsymbol{t},\boldsymbol{x},\alpha,\beta) = \mathcal{N}(t|m(x),s^2(x))$ in terms of known or computable quantities, we do this the same way as we obtained the posterior distribution in step 2.

$$p(\boldsymbol{w}|\boldsymbol{t},\boldsymbol{x}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N) \tag{1.27}$$

$$p(t|\boldsymbol{w},\boldsymbol{t},\boldsymbol{x}) = \mathcal{N}(t|\boldsymbol{\phi}(x)^T\boldsymbol{w},\beta^{-1}) \tag{1.28}$$

$$\rightarrow \tag{1.29}$$

$$p(t|x,\boldsymbol{t},\boldsymbol{x}) = \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(x),\sigma_N^2(x)) \tag{1.30}$$

With $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ defined as before, and

$$\sigma_N^2(x) = \frac{1}{\beta} + \boldsymbol{\phi}(x)^T\boldsymbol{S}_N\boldsymbol{\phi}(x) \tag{1.31}$$

$$p(t|x, \boldsymbol{t}, \boldsymbol{x}, \alpha, \beta) = \mathcal{N}(t|m(x), s^2(x)) \tag{1.32}$$

$$m(x) = \boldsymbol{m}_N^T \boldsymbol{\phi}(x) \tag{1.33}$$

$$= \begin{pmatrix} -0.0445 \\ -0.2021 \end{pmatrix}^T \begin{pmatrix} 1 \\ x \end{pmatrix} \tag{1.34}$$

$$s(x) = \sigma_N^2 \tag{1.35}$$

$$= \frac{1}{10} + \begin{pmatrix} 1 \\ x \end{pmatrix}^T \begin{pmatrix} 22 & 10 \\ 10 & 7.2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ x \end{pmatrix} \tag{1.36}$$

2. See Figure 1.1 where the mean of the predictive Gaussian distribution and one standard deviation on both sides are plotted as a function over x over the interval $[0, 1]$, with the two observed data points. When we compare this plot with Figure 3.8b (Bishop, p.157), is our linear mean. This can be explained by the nature of our basis function ϕ_j, which is two-dimensional in our case and Bishop uses a 9-dimensional function.

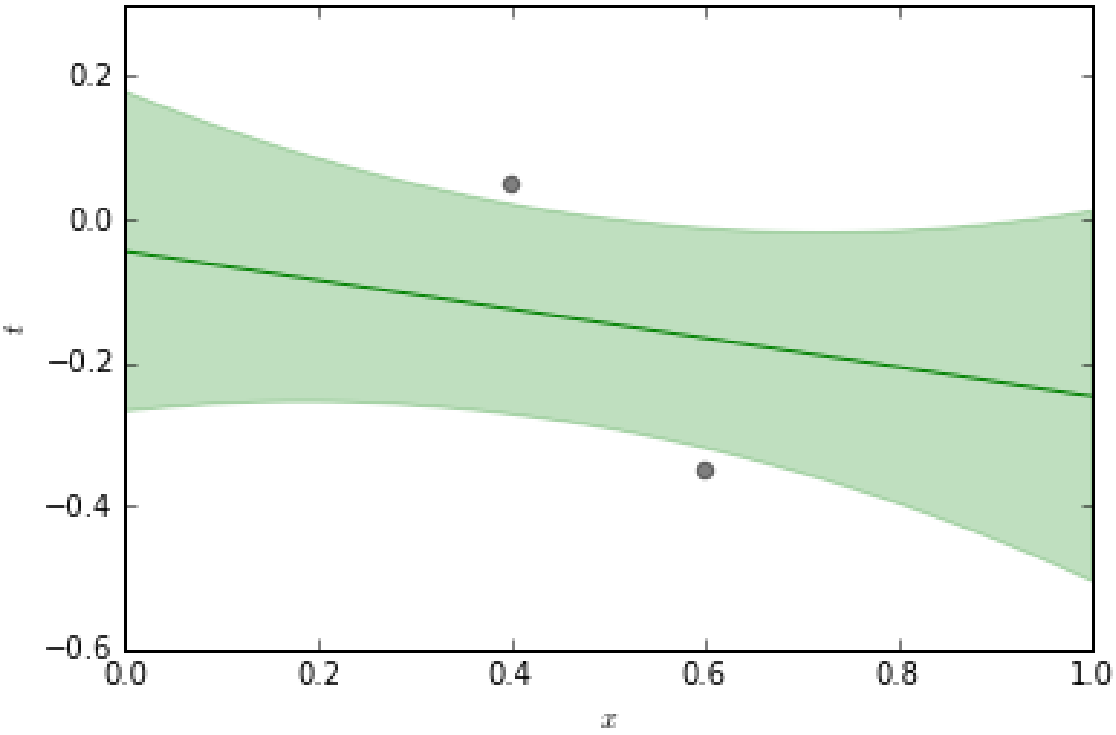

Figure 1.1: The mean of the predictive Gaussian distribution is represented by the green line, plotted against x. The two grey dots are the data points $\{x_1, t_1\}$ and $\{x_2, t_2\}$. The light green area indicates the standard deviation and is bound by the mean plus the standard deviationd and the mean minus the standard deviation.

3. We sampled five functions $y(x, \boldsymbol{w})$ from the posterior distribution over $\boldsymbol{w}$ (see Exercise 1.1 step 2) for this data set and plotted them with the predictive distribution (see Figure 1.2).

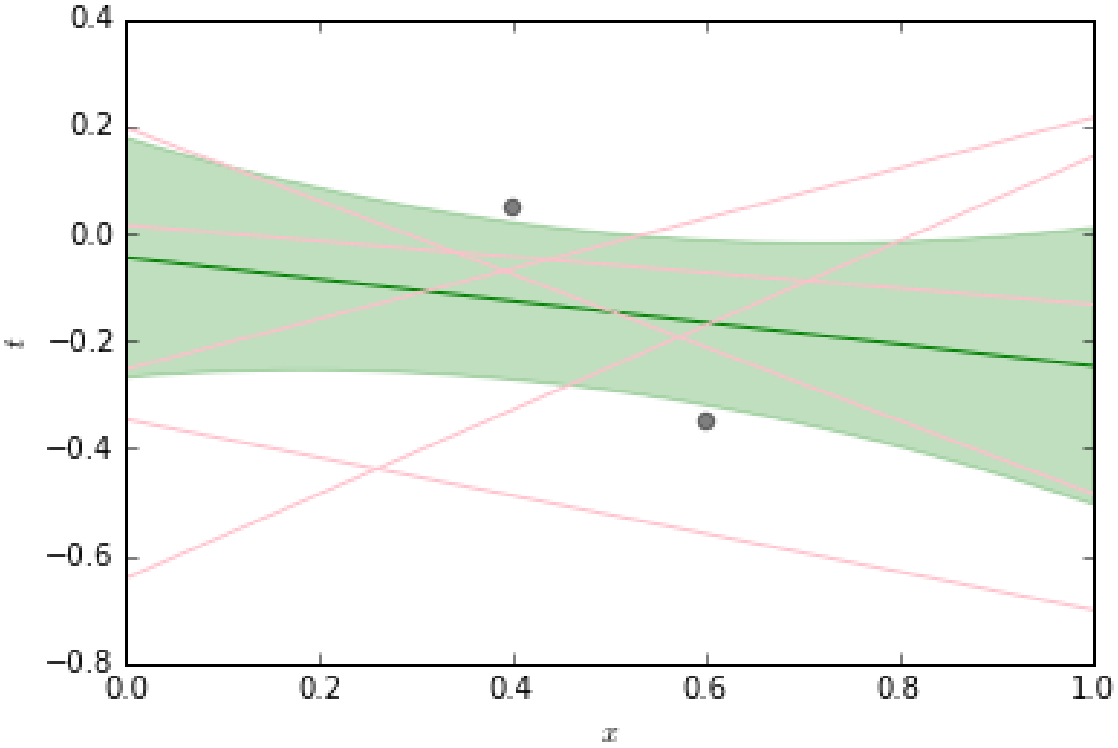

Figure 1.2: The pink lines represent $y(x, \boldsymbol{w})$ where the weights $\boldsymbol{w}$ are sampled from the posterior distribution with mean m_N and variance S_N.

2 Logistic regression

2.1 The IRLS algorithm

1.

2.

3.

2.2 Two-class classification using logistic regression

1.