Linear regression

Dr. E.S.Gopi

Series editor, Signals and Communication, Springer publications,
Co-ordinator and Head, Pattern recognition and Computational intellligence laboratory
Associate professor, Department of ECE
National Institute of Technology Tiruchirappalli, Tamil Nadu, India

January 21, 2022

Motivation

Estimate the outcome of the random vector $\mathbf{x}$ based on the noisy observations on the outcome of the random variable $t = f(\mathbf{x}) + \epsilon$.

Linear Regression

1. Linear Regression-Parametric approach
2. Maximum Likelihood approach
3. Least square estimation
4. Regularization technique
5. Error in Regression= $Bias^2 + var + Noise$
6. Bayes technique
7. Kernel smoothing

Linear regression-Parametric approach

1. Consider $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$
2. $\mathbf{t}$ is the observation and ϵ is Guassian distributed random variable with mean zero and variance $\frac{1}{\beta}$

Linear regression-Parametric approach

1. Consider $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$
2. $\mathbf{t}$ is the observation and ϵ is Guassian distributed random variable with mean zero and variance $\frac{1}{\beta}$
3. Training data: $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_N}$ and the corresponding noisy observations $t_1, t_2 \cdots t_N$

Linear regression-Parametric approach

1. Consider $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$
2. **t** is the observation and ϵ is Guassian distributed random variable with mean zero and variance $\frac{1}{\beta}$
3. Training data: $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_N}$ and the corresponding noisy observations $t_1, t_2 \cdots t_N$
4. Let us assume that the basis functions $\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}) \cdots \phi_{M-1}(\mathbf{x})$ are known.

Linear regression-Parametric approach

1. Consider $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$
2. **t** is the observation and ϵ is Guassian distributed random variable with mean zero and variance $\frac{1}{\beta}$
3. Training data: $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_N}$ and the corresponding noisy observations $t_1, t_2 \cdots t_N$
4. Let us assume that the basis functions $\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \phi_2(\mathbf{x}) \cdots \phi_{M-1}(\mathbf{x})$ are known.
5. The requirement is to estimate obtain the optimal value of $\mathbf{w}$

Linear regression-Parametric approach

1. How do we estimate $\mathbf{w}$
2. What is needed to estimate $\mathbf{w}$

Linear regression-Parametric approach

1. How do we estimate $\mathbf{w}$
2. What is needed to estimate $\mathbf{w}$
3. Let the prior density function of the unknown random vector $\mathbf{w}$ is $f(\mathbf{w})$.

Linear regression-Parametric approach

1. How do we estimate $\mathbf{w}$
2. What is needed to estimate $\mathbf{w}$?
3. Let the prior density function of the unknown random vector $\mathbf{w}$ is $f(\mathbf{w})$.
4. The posterior density function of the random vector w given the observations $t_1, t_2, \cdots t_N$ $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_N}$, is represented as $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$

Linear regression-Parametric approach

1. How do we estimate $\mathbf{w}$
2. What is needed to estimate $\mathbf{w}$?
3. Let the prior density function of the unknown random vector $\mathbf{w}$ is $f(\mathbf{w})$.
4. The posterior density function of the random vector w given the observations $t_1, t_2, \cdots t_N$ $\mathbf{x_1}, \mathbf{x_2} \cdots \mathbf{x_2}$, is represented as $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$
5. $\mathbf{w}$ is estimated as the conditional mean of the posterior density function $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$

Maximum Likelihood Estimation

1. MAP Estimate: $\mathbf{w}$ is estimated that maximizes the posterior density function $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$
2. The likelihood function is obtained as follows $f(\mathbf{w}/\mathbf{t}, \mathbf{x}) = \frac{f(\mathbf{t}/\mathbf{w}, \mathbf{x}) f(\mathbf{w}/\mathbf{x})}{f(\mathbf{t})}$

Maximum Likelihood Estimation

1. MAP Estimate: $\mathbf{w}$ is estimated that maximizes the posterior density function $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$
2. The likelihood function is obtained as follows $f(\mathbf{w}/\mathbf{t}, \mathbf{x}) = \frac{f(\mathbf{t}/\mathbf{w}, \mathbf{x}) f(\mathbf{w}/\mathbf{x})}{f(\mathbf{t})}$
3. In this case, $f(\mathbf{w}/\mathbf{x})$ is assumed to be uniform distributed (Constant).
4. Also the denominator $(\mathbf{t})$ is not playing any role in optimizing $\mathbf{w}$ that maximizes $f(\mathbf{w}/\mathbf{x})$

Maximum Likelihood Estimation

1. MAP Estimate: $\mathbf{w}$ is estimated that maximizes the posterior density function $f(\mathbf{w}/\mathbf{t}, \mathbf{x})$
2. The likelihood function is obtained as follows $f(\mathbf{w}/\mathbf{t}, \mathbf{x}) = \frac{f(\mathbf{t}/\mathbf{w}, \mathbf{x})f(\mathbf{w}/\mathbf{x})}{f(\mathbf{t})}$
3. In this case, $f(\mathbf{w}/\mathbf{x})$ is assumed to be uniform distributed (Constant).
4. Also the denominator $(\mathbf{t})$ is not playing any role in optimizing $\mathbf{w}$ that maximizes $f(\mathbf{w}/\mathbf{x})$
5. The optimal value of $\mathbf{w}$ is estimated that maximizes the likelihood function $f(\mathbf{t}/\mathbf{w}, \mathbf{x})$

Maximum Likelihood Estimation

1. Thus optimal value of $\mathbf{w}$ is estimated that maximizes the likelihood function $f(\mathbf{t}/\mathbf{w}, \mathbf{x})$
2. THIS IS KNOWN AS MAXIMUM LIKELIHOOD ESTIMATION

Likelihood function with N observations

$$f(t_1 t_2 \cdots t_N / \mathbf{w}, \mathbf{x}) = K \prod_{i=1}^{i=N} e^{-\frac{(t_i - w^T \phi(x_i))^2}{2\sigma^2}} \tag{1}$$

1. As logarithm is the increasing function, Maximizing Likelihood function is equivalent to maximizing the logarithm of the Likelihood function
2. Taking logarithm of (1), we get the following.

$$\ln(f(t_1 t_2 \cdots t_N / \mathbf{w}, \mathbf{x})) = -\sum_{i=1}^{i=N} \frac{(t_i - w^T \phi(x_i))^2}{2\sigma^2} + lnK \tag{2}$$

Maximum Likelihood versus Least square solution

$$\ln(f(t_1 t_2 \cdots t_N / \mathbf{w}, \mathbf{x})) = -\sum_{i=1}^{i=N} \frac{(t_i - w^T \phi(x_i))^2}{2\sigma^2} + lnK \tag{3}$$

1. Maximizing (2) is equivalent to minimizing the following.

$$\frac{(t_i - w^T \phi(x_i))^2}{2\sigma^2} \tag{4}$$

2. This is the Least square solution

Matrix Representation

This can be written in the matrix form as follows.

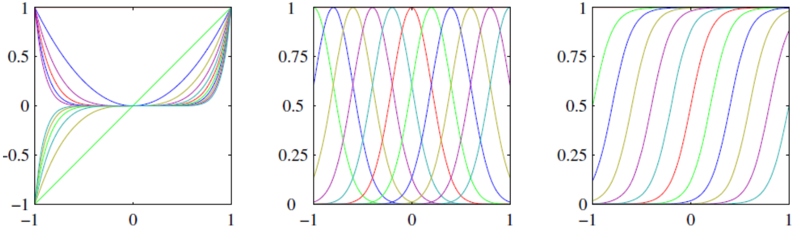
$$\begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \cdots & \phi_{M-1}(x_2) \\ \cdots & \cdots & \cdots \\ \phi_0(x_N) & \cdots & \phi_{M-1}(x_N) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_N \end{bmatrix} \tag{5}$$

This is represented as follows. $\mathbf{\Phi w} = \mathbf{t}$The solution is obtained using pseudo inverse as $\hat{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$

Observation

1. We understand that Maximum Likelihood estimation is identical as that of the Least square estimation (?) $\cdots$

Basis function



Examples of basis functions, showing polynomials on the left, Gaussians of the form in the centre, and sigmoidal of the form on the right.

$$\phi_j(x) = x^j \qquad \phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\} \qquad \phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

$$\sigma(a) = \frac{1}{1+\exp(-a)} \qquad \tanh(a) = 2\sigma(a) - 1$$

Prediction distribution

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w}$$

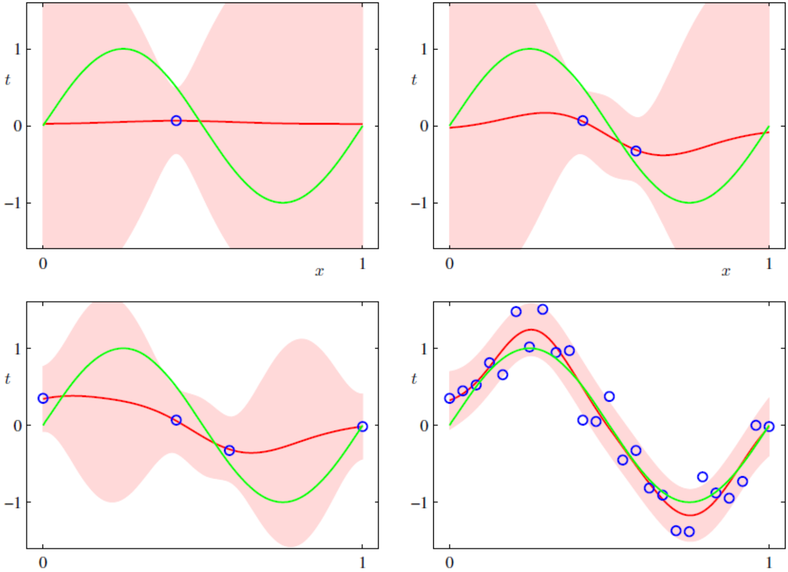
$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

where the variance $\sigma_N^2(\mathbf{x})$ of the predictive distribution is given by

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

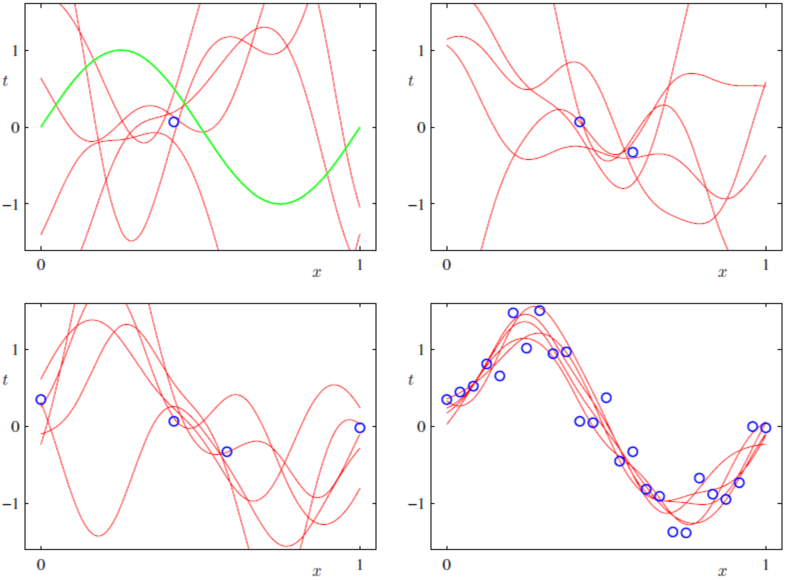
$$\sigma_{N+1}^2(\mathbf{x}) \leqslant \sigma_N^2(\mathbf{x}).$$

Prediction distribution



Examples of the predictive distribution (3.58) for a model consisting of 9 Gaussian basis functions

Prediction distribution



Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w}$ corresponding to the plots in Figure

Regularization techniques.

1. The observation $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$ is the parametric model.
2. In this, t is the scalar observation corresponding to the input vector $\mathbf{x}$.
3. ϵ is Gaussian distributed with mean zero and variance $\frac{1}{\beta}$
4. Given the training data, establishing the relationship $y(\mathbf{x}) = w^T \phi(\mathbf{x})$ needs estimating the value of $\mathbf{w}$.

Regularization techniques.

1. Given the training data, establishing the relationship $y(\mathbf{x}) = w^T\phi(\mathbf{x})$ needs estimating the value of $\mathbf{w}$.
2. $f(\mathbf{w})$ is the prior density function
3. $f(\mathbf{w}/t_1 t_2 \cdots t_N)$ is the posterior density function.
4. $f(t_1 t_2 \cdots t_N/\mathbf{w})$ is the likelihood function.
5. They are related using Bayes as follows.

$$f(\mathbf{t}/\mathbf{w}) = \frac{f(\mathbf{t}/\mathbf{w})f(\mathbf{w})}{f(\mathbf{t})} \tag{6}$$

Regularization techniques

1. It is observed that $f(\mathbf{t}/\mathbf{w})$ is modelled as Gaussian distributed with mean $\mathbf{w}^T\phi(x)$ and variance $\frac{1}{\beta}$
2. In Likelihood estimation, $f(\mathbf{w})$ is uniform distributed and hence maximizing $f(\mathbf{w}/\mathbf{t})$ (MAP) is identical as that of maximizing $f(\mathbf{t}/\mathbf{w})$
3. This is known as Maximum Likelihood estimation
4. As log is the increasing function Maximizing $f(\mathbf{t}/\mathbf{w})$ is same as that of maximizing the logarithm of the likelihood function.
5. This ends up solving the matrix $\Phi\mathbf{w} = \mathbf{t}$

Regularization Techniques

1. $$\begin{bmatrix} \phi_0(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \cdots & \phi_{M-1}(x_2) \\ \cdots & \cdots & \cdots \\ \phi_0(x_N) & \cdots & \phi_{M-1}(x_N) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ \vdots \\ t_N \end{bmatrix} \tag{7}$$
2. Using pseudo inverse computation **w** is estimated as the following.
$$\mathbf{w} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t} \tag{8}$$
3. From the above, it is understood that the estimated vector **w** is data dependent
4. Ends up with Overfitting.

Regularization

1. To circumvent this, the Least square problem is formulated with the constraints $\sum_{n=1}^{n=M} |w_n|^2 \leq \eta$ as given below.

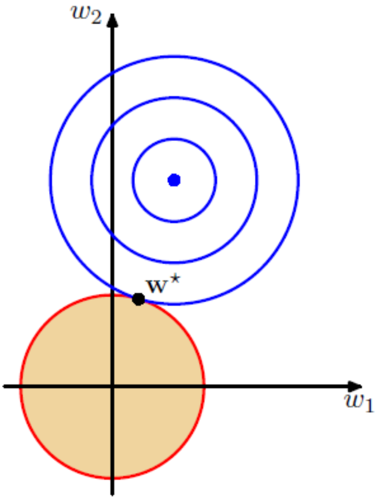
$$J = \frac{1}{2}\sum_{n=1}^{n=N}(t_n - \mathbf{w}^T\phi(\mathbf{x}))^2 + \frac{\lambda}{2}\sum_{n=1}^{n=M}|w_n|^2$$

2. The estimate is given as the following.

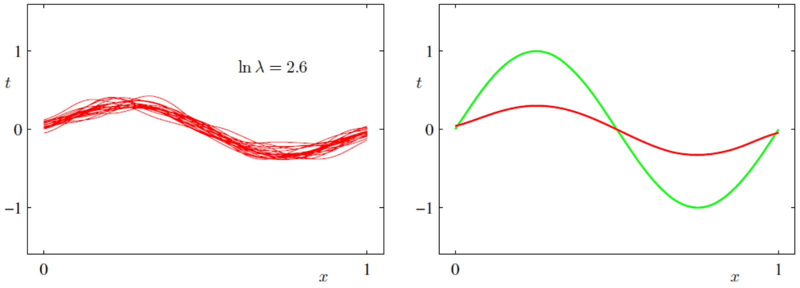
$$\mathbf{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{t} \tag{9}$$

3. λ is known as Regularization constant.

Regularization

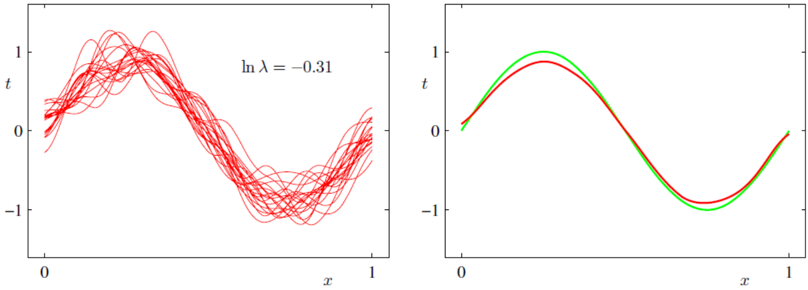


Regularization



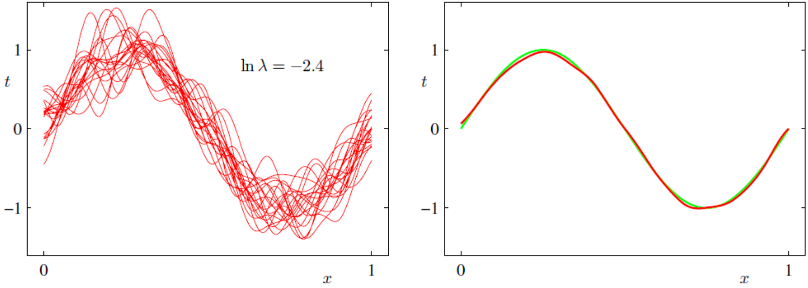
1. Number of datasets (L) =100
2. Number of data points (N) =25
3. Number of Gaussian basis functions=24, i.e $M = 25$

Regularization



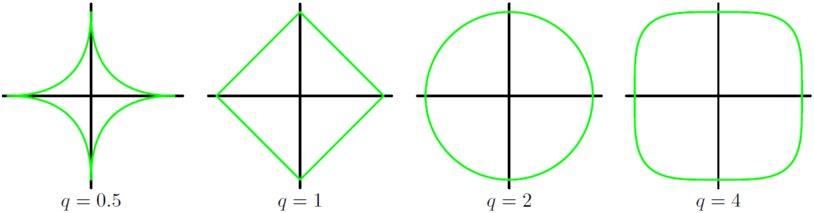
1. Number of datasets (L) =100
2. Number of data points (N) =25
3. Number of Gaussian basis functions=24, i.e $M = 25$

Regularization



1. Number of datasets (L) =100
2. Number of data points (N) =25
3. Number of Gaussian basis functions=24, i.e $M = 25$

Regularization



Contours of the regularization term in for various values of the parameter q.

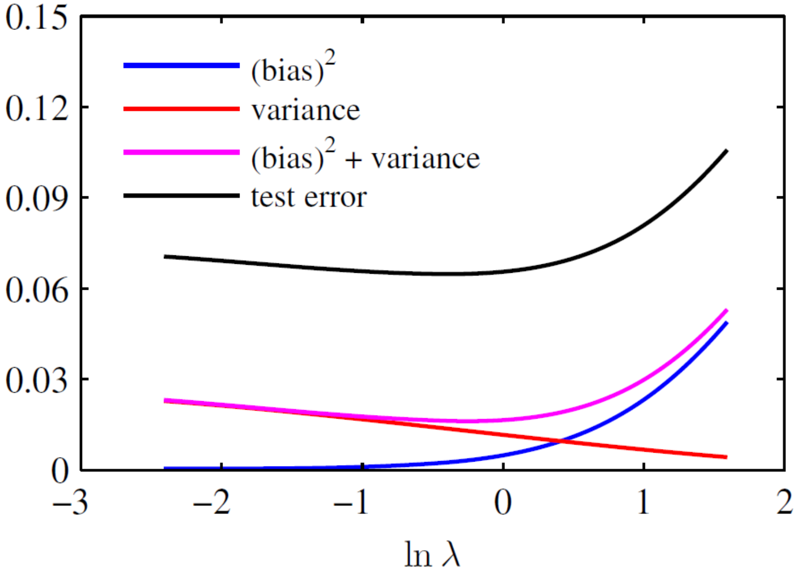
$\text{Bias}^2 + \text{Variance}$

$$\overline{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\overline{y}(x_n) - h(x_n)\}^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \overline{y}(x_n)\right\}^2$$

$Bias^2 + Variance$



1. The observation $t = \mathbf{w}^T \phi(\mathbf{x}) + \epsilon$ is the parametric model.
2. In this, t is the scalar observation corresponding to the input vector $\mathbf{x}$.
3. ϵ is Gaussian distributed with mean zero and variance $\frac{1}{\beta}$
4. Given the training data, establishing the relationship $y(\mathbf{x}) = w^T \phi(\mathbf{x})$ needs estimating the value of $\mathbf{w}$.

Bayes technique to estimate $\mathbf{w}$ in parametric approach based Linear regression.

1. Given the training data, establishing the relationship $y(\mathbf{x}) = w^T\phi(\mathbf{x})$ needs estimating the value of $\mathbf{w}$.
2. $f(\mathbf{w})$ is the prior density function
3. $f(\mathbf{w}/t_1 t_2 \cdots t_N)$ is the posterior density function.
4. $f(t_1 t_2 \cdots t_N/\mathbf{w})$ is the likelihood function.
5. They are related using Bayes as follows.

$$f(\mathbf{t}/\mathbf{w}) = \frac{f(\mathbf{t}/\mathbf{w})f(\mathbf{w})}{f(\mathbf{t})} \tag{10}$$

Bayes technique to estimate $\mathbf{w}$ in parametric approach based Linear regression

1. It is observed that $f(\mathbf{t}/\mathbf{w})$ is modelled as Gaussian distributed with mean $\mathbf{w}^T\phi(x)$ and variance $\frac{1}{\beta}$
2. In Likelihood estimation, $f(\mathbf{w})$ is uniform distributed and hence maximizing $f(\mathbf{w}/\mathbf{t})$ (MAP) is identical as that of maximizing $f(\mathbf{t}/\mathbf{w})$
3. This is known as Maximum Likelihood estimation
4. As log is the increasing function Maximizing $f(\mathbf{t}/\mathbf{w})$ is same as that of maximizing the logarithm of the likelihood function.
5. This ends up solving the matrix $\Phi\mathbf{w} = \mathbf{t}$

Bayes technique to estimate **w** in parametric approach based Linear regression

1. Using pseudo inverse computation **w** is estimated as the following.
$$\mathbf{w} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{x} \tag{11}$$
2. What if **w** is assumed as Multivariate Gaussian density function?

Bayes technique to estimate $\mathbf{w}$ in parametric approach based Linear regression

1. If the prior density function $f(\mathbf{w})$ is assumed as Multivariate Gaussian density function with mean $\mathbf{m_o}$ and co-variance matrix $\mathbf{S_o}$.
2. Then the Aposterior density function of $\mathbf{w}$ given $\mathbf{t}$ is also Gaussian with mean vector $\mathbf{m_N}$ and co-variance matrix $\mathbf{S_N}$ as shown below.

$$\mathbf{m_N} = \mathbf{S_N}(\mathbf{S_o}^{-1} m_o + \beta \mathbf{\Phi}^\mathbf{T} \mathbf{t})$$
$$\mathbf{S_N} = (\mathbf{S_o}^{-1} + \beta(\mathbf{\Phi}^\mathbf{T}\mathbf{\Phi}))^{-1}$$

Bayes technique to estimate $\mathbf{w}$ in parametric approach based Linear regression

1. If the prior density function $f(\mathbf{w})$ is assumed as Multivariate Gaussian density function with mean $\mathbf{m_o}$ and co-variance matrix $\mathbf{S_o}$.
2. Then the Aposterior density function of $\mathbf{w}$ given $\mathbf{t}$ is also Gaussian with mean vector $\mathbf{m_N}$ and co-variance matrix $\mathbf{S_N}$ as shown below.

$$\mathbf{m_N} = \mathbf{S_N}(\mathbf{S_o}^{-1} m_o + \beta \mathbf{\Phi}^\mathbf{T} \mathbf{t})$$
$$\mathbf{S_N} = (\mathbf{S_o}^{-1} + \beta(\mathbf{\Phi}^\mathbf{T} \mathbf{\Phi}))^{-1}$$

3. What is the Conditional mean , Conditional median and the Conditional mode estimate of the posterior density function $f(\mathbf{w}/\mathbf{t})$?

Bayes technique to estimate **w** in parametric approach based Linear regression

1. Consider the case when $\mathbf{m_o}$ is zero vector and the covariance matrix is diagonal as shown below.

$$S_o = \frac{1}{\alpha} I \tag{12}$$

2.

$$\mathbf{m_N} = \beta \mathbf{S_N} \mathbf{\Phi^T t}$$
$$\mathbf{m_N} = \beta((\mathbf{S_o}^{-1} + \beta(\mathbf{\Phi^T \Phi}))^{-1})^{-1} \mathbf{\Phi^T t}$$
$$\mathbf{m_N} = \beta(\alpha \mathbf{I} + \beta(\mathbf{\Phi^T \Phi}))^{-1} \mathbf{\Phi^T t}$$

Bayes technique to estimate **w** in parametric approach based Linear regression

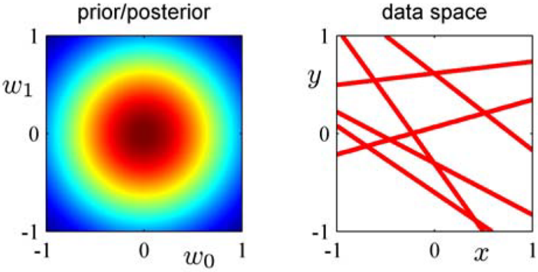
1.

$$\mathbf{m_N} = \beta \mathbf{S_N} \mathbf{\Phi}^\mathbf{T} \mathbf{t}$$
$$\mathbf{m_N} = \beta((\mathbf{S_o}^{-1} + \beta(\mathbf{\Phi}^\mathbf{T}\mathbf{\Phi}))^{-1})\mathbf{\Phi}^\mathbf{T} \mathbf{t}$$
$$\mathbf{m_N} = \beta(\alpha \mathbf{I} + \beta(\mathbf{\Phi}^\mathbf{T}\mathbf{\Phi}))^{-1}\mathbf{\Phi}^\mathbf{T} \mathbf{t}$$
$$\mathbf{m_N} = (\mathbf{\Phi}^\mathbf{T}\mathbf{\Phi} + \frac{\alpha}{\beta}\mathbf{I})^{-1}\mathbf{\Phi}^\mathbf{T} \mathbf{t}$$

Bayes technique to estimate **w** in parametric approach based Linear regression

1. $\mathbf{w} = \mathbf{m_N} = (\mathbf{\Phi^T \Phi} + \frac{\alpha}{\beta}\mathbf{I})^{-1}\mathbf{\Phi^T t}$
2. This solution can be viewed as the Regularized least square solution with $\lambda = \frac{\alpha}{\beta}$

Bayes technique to estimate **w** in parametric approach based Linear regression

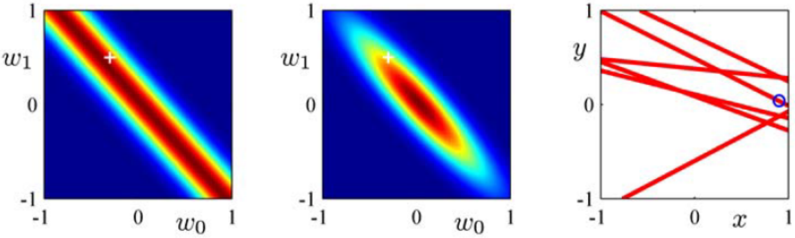


$$y(x, \mathbf{w}) = w_0 + w_1 x$$

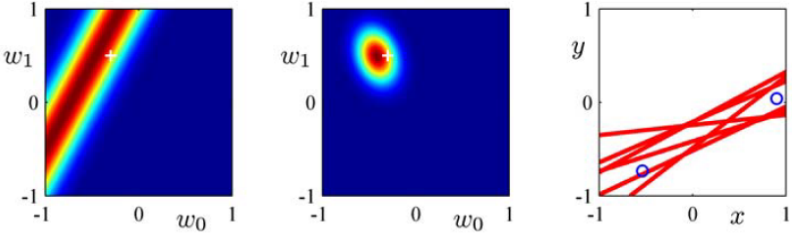
$$\beta = (1/0.2)^2 = 25$$

$$\alpha = 2.0$$

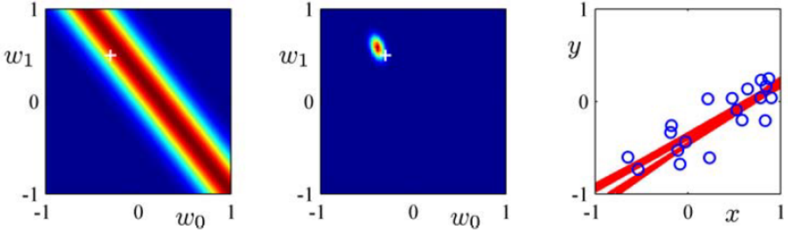
Bayes technique to estimate **w** in parametric approach based Linear regression



Bayes technique to estimate **w** in parametric approach based Linear regression



Bayes technique to estimate **w** in parametric approach based Linear regression



Kernel smoothing

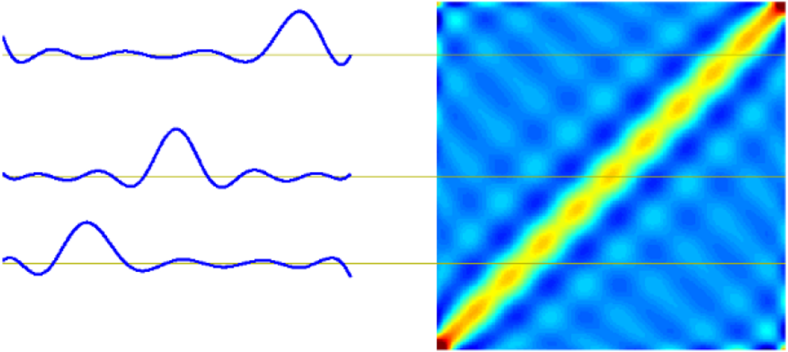
$$y(\mathbf{x}, \mathbf{m}_N) = \mathbf{m}_N^{\mathrm{T}} \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} = \sum_{n=1}^{N} \beta \phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x}_n) t_n$$

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n$$

$$k(\mathbf{x}, \mathbf{x}') = \beta \phi(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \phi(\mathbf{x}')$$

is known as the *smoother matrix* or the *equivalent kernel*.

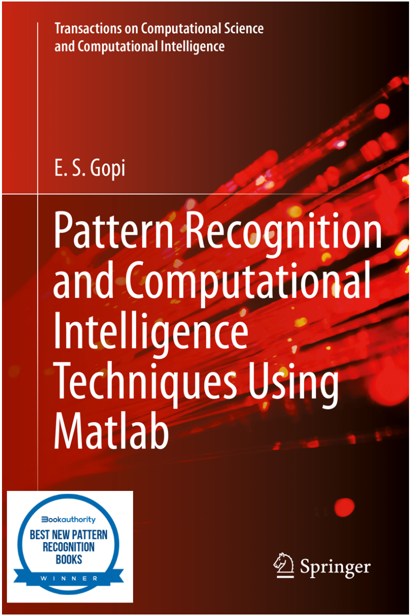
Kernel smoothing



Reference

1. Christopher Bishop,Pattern recognition and Machine Learning, Springer, 2006.
2. E.S.Gopi, Pattern recognition and computational intelligence, Springer, 2020.

Book


Transactions on Computational Science
and Computational Intelligence
E. S. Gopi
Pattern Recognition and Computational Intelligence Techniques Using Matlab
Bookauthority
BEST NEW PATTERN RECOGNITION BOOKS
WINNER
Springer