Linear Models for Regression

Yan Yan

Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences

yan.yan@siat.ac.cn

September 30, 2016

Overview

1. Linear Basis Function Models
   - Maximum likelihood and least squares
   - Geometry of least squares
   - Sequential learning
2. Bias-Variance Decomposition
3. Bayesian Linear Regression
   - Parameter distribution
   - Predictive distribution
   - Equivalent kernal
4. Bayesian Model Comparison
5. Evidence Approximation
   - Evaluation of Approximation Function
   - Maximizeing the evidence function

Linear Basis Function Models

The simplest linear model for regression (often simply known as *linear regression*):

$$y(\boldsymbol{x}, \boldsymbol{w}) = \omega_0 + \omega_1 x_1 + \cdots + \omega_D x_D = \sum_{j=0}^{M-1} \omega_j \phi_j(\boldsymbol{x}) = \boldsymbol{\omega}^T \phi(x) \qquad (1)$$

where $\boldsymbol{\omega} = (\omega_0, ..., \omega_{M-1})^T$ and $\boldsymbol{\phi} = (\phi_0, \cdots, \phi_{M-1})^T$. This kinds of models are called linear models.

When ϕ has different types, which means different kinds of *basis function*, we have other kinds of modes, like:

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

or

$$\phi_j(x) = \sigma(\frac{x-\mu_j}{s}) \quad and \quad \sigma(a) = \frac{1}{1+exp(-a)} \qquad (3)$$

Examples of basis function

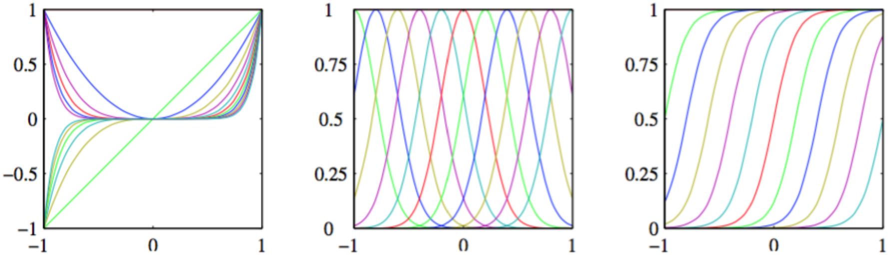

Figure: Examples of basis functions.

Maximum Likelihood and Least Squares I

The target variable t is given by a deterministic function $y(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise so that

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \tag{4}$$

where ϵ is a zero mean Gaussian random variable with precision β.

So, we can write

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \tag{5}$$

Consider a data set of inputs X with corresponding target value vector $\boldsymbol{t}$, make the assumption that these data points are drawn independently from the distribution. So the likelihood function is:

$$lnp(\boldsymbol{t}|\boldsymbol{w}, \beta) = \sum_{n=1}^{N} ln\mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \tag{6}$$

.

Maximum Likelihood and Least Squares II

The sum-of-squares error function is defined by:

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{7}$$

Let the gradient of the log likelihood function equals to 0. Solving for w we obtain (which are known as the *normal equations* for the least squares problem):

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} \tag{8}$$

Here $\boldsymbol{\Phi}$ is an $N \times M$ matrix (the *design matrix*):

$$\begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{bmatrix}$$

Geometry of Least Squares

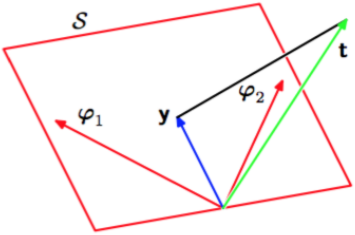

Figure: Examples of basis functions.

In an N-dimensional space whose axes are the value of $t_1, \cdots, t_N$. The least-squares regression function is obtained by finding the orthogonal projection of the data vector $\boldsymbol{t}$ onto the subspace spanned by the basis functions.

Sequential Learning

Batch techniques, such as the maximum likelihood solution which can process large dataset in one go, if the dataset if sufficiently large, it may be worthwhile to use sequential algorithms, known as *on-line* algorithm. Sequential learning is also appropriate for real-time applications.

Stochastic gradient descent is applied. If the error function comprises a sum over data points $E = \sum_n E_n$, the update rule:

$$\boldsymbol{w}^{\tau+1} = \boldsymbol{w}^{\tau} - \eta \nabla E_n \tag{9}$$

For the case of the sum-of-squares error function, this gives:

$$\boldsymbol{w}^{\tau+1} = \boldsymbol{w}^{\tau} - \eta (t_n - \boldsymbol{w}^{(\tau)T} \phi_n) \phi_n \tag{10}$$

Regularized Least Squares I

The regularization terms are added to the error functions. One of the simplest forms of regularizer is given by the sum-of-squares of the weight vector elements:

$$E_W(\boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}) \tag{11}$$

Consider the sum-of-squares error function with regularizer (quadratic regularizer):

$$E = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{12}$$

This choice of regularizer is known as *weight decay*, because in sequential learning algorithms, it encourages weight values to decay towards zero.

Solving fro w as the gradient of (12) equal to zero, we obtain:

$$\boldsymbol{w} = (\lambda\boldsymbol{I} + \boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{t} \tag{13}$$

This represents a simple extension of the least-squares solution (8).

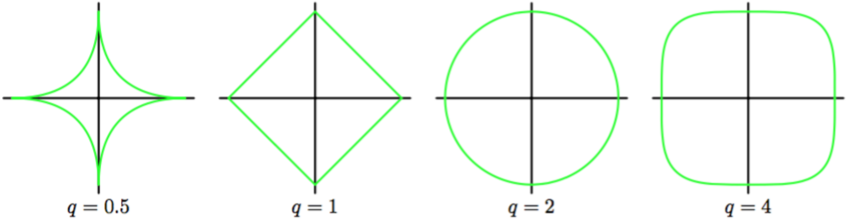

Figure: Contours of the regularization term for various values of the parameter q.

A more general regularizer is

$$\frac{\lambda}{2} \sum_{j=1}^{M} |w_j|^q \tag{14}$$

when q=2 corresponds to the quadratic regularizer in (12).

Regularized Least Squares III

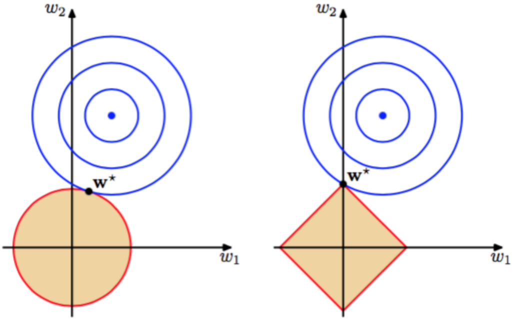

Figure: The error function without regularization and the constraint region for quadratic regularizer (left) and lasso regularizer (right). We can see the w in lasso equals to 0 make a sparse solution.

The case of q=1 is known as the *lasso* which leading to a *sparse* model in which the corresponding basis functions play no role. Regularization allows complex models to be trained on data sets of limited size without severe over-fitting, essentially by limiting the effective model complexity. However, the problem of determining the optimal model complexity is then shifted from one of finding the regularization coefficient λ.

Multiple Outputs

If the target value t turns to a vector $\boldsymbol{t}$ which means multiple outputs. This could be done via multiple, independent regression problems.

However, usually we use the same set of basis functions to model all components of the target value vector. Suppose we take the conditional distribution of the target vector to be an isotropic Gaussian of the form

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{W}, \beta) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{x}), \beta^{-1}\boldsymbol{I}) \tag{15}$$

The log likelihood function is then given by:

$$lnp(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{W}, \beta) = \sum_{n=1}^{N} \mathcal{N}(\boldsymbol{t}_n|\boldsymbol{W}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}\boldsymbol{I}) \tag{16}$$

and maximize this function with respect to $\boldsymbol{W}$.

Test

Parameter Distribution

Predictive Distribution

Equivalent Kernal

Bayesian Model Comparison

Evaluation of Approximation Function

Maximizeing the Evidence Function

Bullet Points

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Aliquam blandit faucibus nisi, sit amet dapibus enim tempus eu
- Nulla commodo, erat quis gravida posuere, elit lacus lobortis est, quis porttitor odio mauris at libero
- Nam cursus est eget velit posuere pellentesque
- Vestibulum faucibus velit a augue condimentum quis convallis nulla gravida

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Block 3

Suspendisse tincidunt sagittis gravida. Curabitur condimentum, enim sed venenatis rutrum, ipsum neque consectetur orci, sed blandit justo nisi ac lacus.

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)
Title of the publication
Journal Name 12(3), 45 – 678.

The End