

Chapter 2 Linear Model

Siheng Zhang
zhangsiheng@cvte.com

December 13, 2020

This part corresponds to **Chapter 1,3,4 of PRML, Chapter of UML**, and mainly answers the following questions:

-

Contents

1	Linear classification	2
2	Linear regression	2
2.1	Ridge regression	2
2.2	Lasso	2
3	Generalized linear model	3

1 Linear classification

In the last chapter, we stop at the linear classification of binary classification task, i.e.,

$$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0 \quad (1)$$

in which \mathbf{w} is weight vector, and w_0 is bias. The input vector is assigned to class C_1 iff. $f(\mathbf{x}) \geq 0$ and to class C_2 otherwise.

Consider two points $\mathbf{x}_1, \mathbf{x}_2$ on the decision boundary, i.e., $\mathbf{w}^\top (\mathbf{x}_1 - \mathbf{x}_2) = 0$, hence \mathbf{w} is orthogonal to the decision boundary. And the distance from the origin to the decision boundary is

$$\frac{\mathbf{w}^\top \mathbf{x}}{\|\mathbf{w}\|} = \frac{-w_0}{\|\mathbf{w}\|} \quad (2)$$

It is usually convenient to use a more compact notation in which we introduce an additional input value $x_0 = 1$ and then define $\tilde{\mathbf{w}} = (w_0, \mathbf{w})$ and $\tilde{\mathbf{x}} = (x_0, \mathbf{x})$ so that $y = f(\mathbf{x}) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{x}}$.

1.1 Extend to multiple classes

1.2 Fisher's linear discriminant

2 Linear regression

$$\min_{\mathbf{w}} \sum_{i=1}^N (\mathbf{w}\mathbf{x}_i - y_i)^2 \quad (3)$$

Suppose the fitting error $\epsilon_i = y_i - \mathbf{w}\mathbf{x}_i$ is Gaussian noise, i.e., $\epsilon_i \sim \mathcal{N}(0, \beta)$. Then the log likelihood function of the training sequence is:

$$\log \mathcal{L} = -\frac{N}{2} \log 2\pi\beta - \sum_{i=1}^N \frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\beta} \quad (4)$$

Obviously, MLE is equivalent to linear regression.
over-fitting

2.1 Ridge regression

Ridge regression addresses on over-fitting by penalizing the l_2 -norm of weight vector \mathbf{w} ,

$$\min_{\mathbf{w}} \sum_{i=1}^N (\mathbf{w}\mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

If we assume a Gaussian prior for the weight vector, $\mathbf{w} \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$, then the posterior of the training sequence is:

$$p(\mathbf{w}|S) \propto p(\mathbf{w})p(S|\mathbf{w}) \propto \exp\left(-\frac{\alpha}{2}\mathbf{w}^\top \mathbf{w}\right) \cdot \prod_{i=1}^N \exp\left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\beta}\right) \quad (5)$$

Maximizing the log posterior function is equivalent to the ridge regression.

2.2 Lasso

Lasso addresses on over-fitting by penalizing the l_1 -norm of weight vector \mathbf{w} ,

$$\min_{\mathbf{w}} \sum_{i=1}^N (\mathbf{w}\mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_1$$

If we assume a Laplace prior for the weight vector, $p(\mathbf{w}) = \frac{1}{2\alpha} \exp\left(-\frac{\|\mathbf{w}\|_1}{\alpha}\right)$, then the posterior of the training sequence is:

$$p(\mathbf{w}|S) \propto p(\mathbf{w})p(S|\mathbf{w}) \propto \exp\left(-\frac{\|\mathbf{w}\|_1}{\alpha}\right) \cdot \prod_{i=1}^N \exp\left(-\frac{(y_i - \mathbf{w}\mathbf{x}_i)^2}{2\beta}\right) \quad (6)$$

Maximizing the log posterior function is equivalent to the Lasso model.

3 Generalized linear model