ml

Tanvir

November 2025

(data + learning algorithm) $\rightarrow$ function

1 Supervised learning

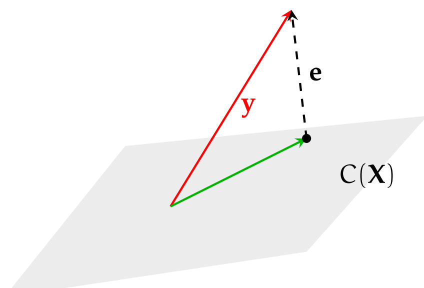
1.1 Regression

model, parameters, cost function, objective

1.1.1 Linear (in w) Regression

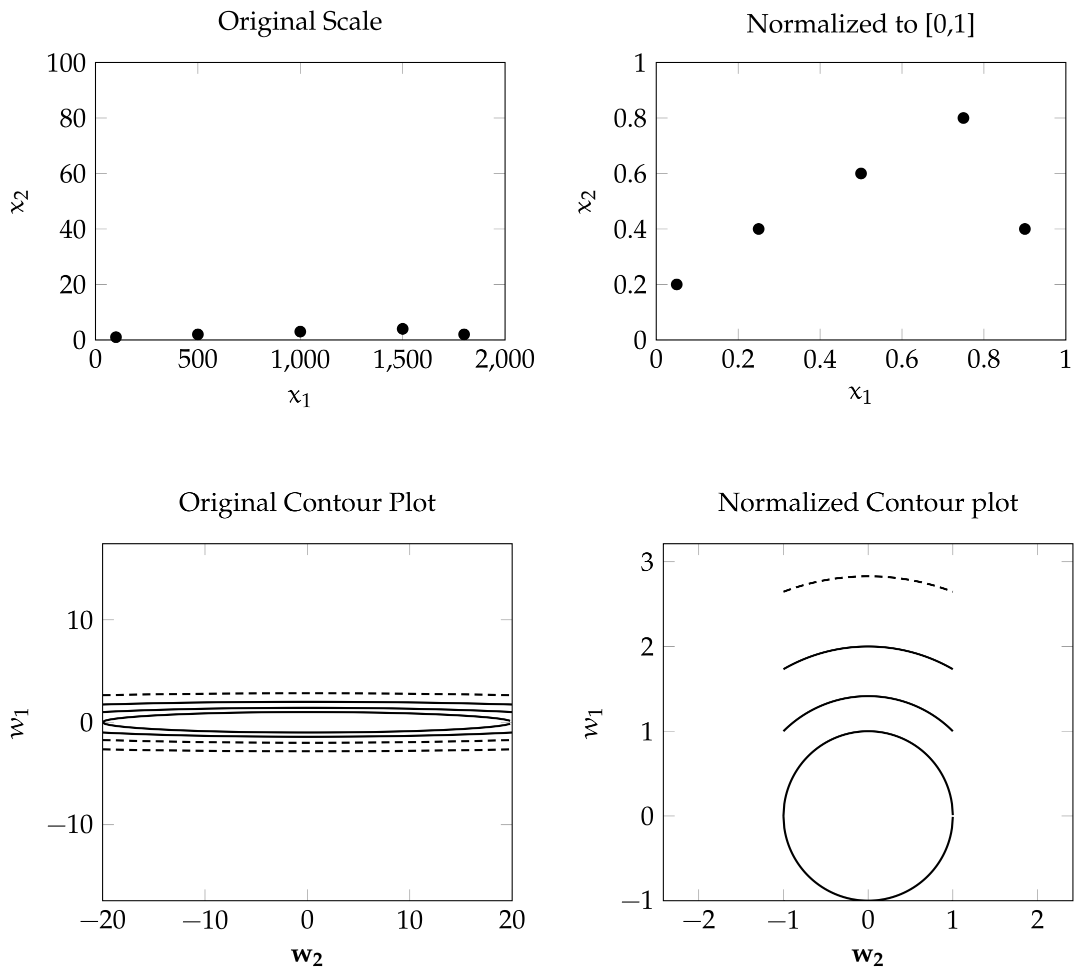
$\mathbf{X}^{m\times n}$ has m examples, each having n features. Usually $m >> n$. $\mathbf{y}^{m\times 1}$ are the corresponding outputs.

$$\begin{aligned}
\mathbf{X}\mathbf{w} &\approx \mathbf{y} \\
\mathbf{e} &= \mathbf{X}\mathbf{w} - \mathbf{y} \\
\underset{\mathbf{w}}{\text{minimize}}\ J(\mathbf{w}) &= \|\mathbf{e}\|^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) \\
\text{gradient descent: } \mathbf{w} &= \mathbf{w} - \alpha\frac{\partial J}{\partial \mathbf{w}} = \mathbf{w} - 2\alpha\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w} - 2\alpha\mathbf{X}^T\mathbf{e}
\end{aligned}$$



1.2 Normalization

Changes scale keeping the shape of distribution same. Gradient descent now works, with more ease, in the world of concentric circular contours.



1.3 Learning rate, α

A good value is between too small (slow convergence) and too big (divergence).