ml

Tanvir

November 2025

(data + learning algorithm) $\to$ function

1 Supervised learning

1.1 Regression

model, parameters, cost function, objective
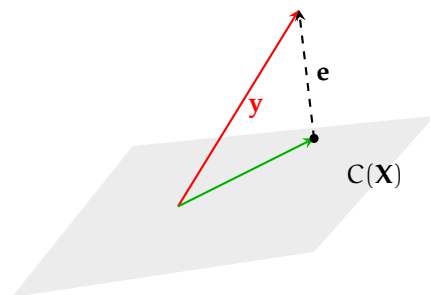
1.1.1 Linear Regression

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

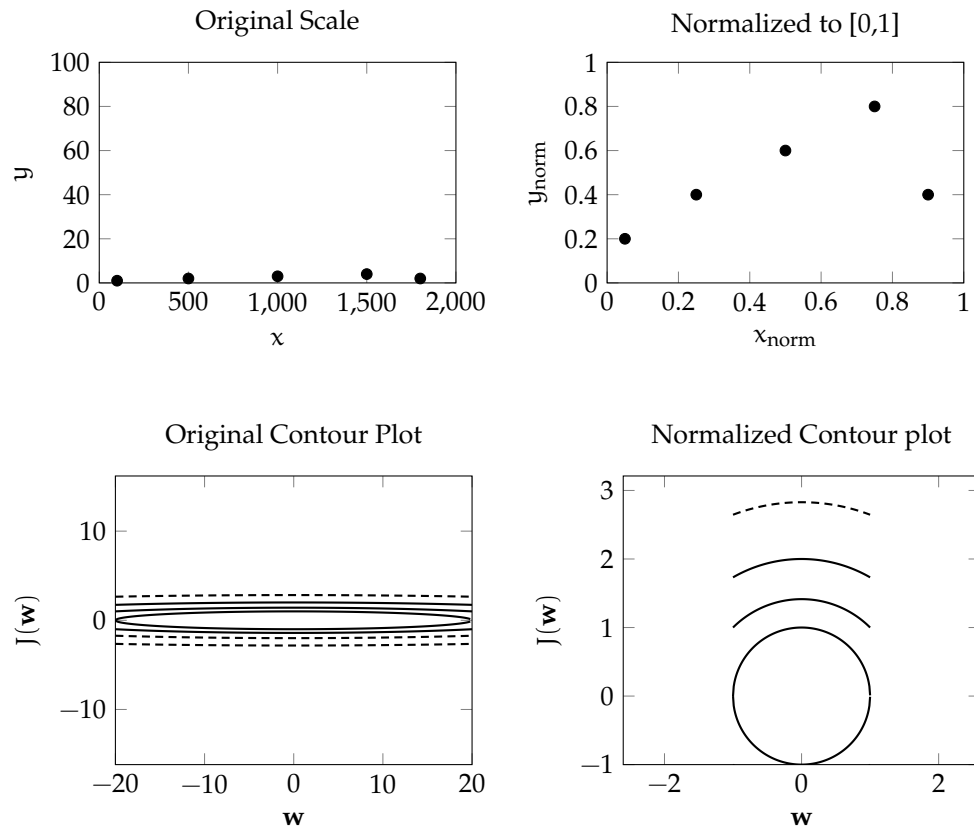$$\mathbf{Xw} \approx \mathbf{y}$$
$$\mathbf{e} = \mathbf{Xw} - \mathbf{y}$$
$$\underset{\mathbf{w}}{\text{minimize }} J(\mathbf{w}) = \|\mathbf{e}\|^2 = (\mathbf{Xw} - \mathbf{y})^\mathsf{T}(\mathbf{Xw} - \mathbf{y})$$
$$\text{gradient descent: } \mathbf{w} = \mathbf{w} - \alpha\frac{\partial J}{\partial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

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.

Original Scale

Normalized to [0,1]

Original Contour Plot

Normalized Contour plot

1.3 Learning rate, α

α too big

2