We can speed up gradient descent by having each of our input values in roughly the same range. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

The way to prevent this is to modify the ranges of our input variables so that they are all roughly the same. Ideally: −1 ≤ x ≤ 1 or −0.5 ≤ x ≤ 0.5

feature scaling

Divide the features by the range.

mean normalization

Subtract mean from the features and divide it by its range or standard deviation.

octave: sigma = std(v);