University of Toronto
CSC321, Winter 2018

HW 1: Xiangyu Kong 1002109620 kongxi16

1

1. Given:

$$\begin{aligned}
\mathcal{E}_{\text{Reg}} &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 + \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 + \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}((\sum_{j'=1}^{D} w_{j'}x_{j'}^{(i)} + b) - t^{(i)})^2 + \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2
\end{aligned}$$

We can calculate that:

$$\begin{aligned}
\frac{\partial \mathcal{E}_{\text{Reg}}}{\partial w_j} &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}((\sum_{j'=1}^{D} w_{j'}x_{j'}^{(i)} + b) - t^{(i)}) + \lambda w_j \\
&= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \lambda w_j
\end{aligned}$$

Then the gradient descent update rule for w_j is:

$$w_j \leftarrow w_j - (\frac{\alpha}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \lambda w_j)$$

For b, since there is no new term in $\mathcal{R}(\mathbf{w})$ that involves b, the update rule for b remains the same as before:

$$b \leftarrow b - \alpha\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})$$

This form of regularization is called "weight decay" because weight w_j will decay in proportion to its current size. I.e. the larger the weight is, the larger the decay will be.

2. According to above,

$$\begin{aligned}
\frac{\partial \mathcal{E}_{\text{Reg}}}{\partial w_j} &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \lambda w_{j'} \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_j^{(i)}(\sum_{j'=1}^{D} w_{j'}x_{j'}^{(i)}) - t^{(i)}) + \lambda w_{j'} \\
&= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(\sum_{j'=1}^{D} w_{j'}x_{j'}^{(i)}) - \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}t^{(i)} + \lambda w_{j'} \\
&= \sum_{j'=1}^{D}(\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}x_{j'}^{(i)})w_{j'} + \lambda w_{j'} - (\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}t^{(i)}) \\
&= \sum_{j'=1}^{D}(\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}x_{j'}^{(i)} + \lambda I)w_{j'} - (\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}t^{(i)})
\end{aligned}$$

Then $A_{jj'} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}x_{j'}^{(i)} + \lambda I$ and $c_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}t^{(i)}$ where I is the identity matrix

2

1. According to the formula for $\mathcal{E}$,

$$\mathcal{E} = \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2$$

In this case, plug in the data $(x^{(1)}, t^{(1)})$, $(x^{(2)}, t^{(2)})$ and $(x^{(3)}, t^{(3)})$

$$\begin{aligned}
\mathcal{E} &= \frac{1}{2N}[(y^{(1)} - t^{(1)})^2 + (y^{(2)} - t^{(2)})^2 + (y^{(3)} - t^{(3)})^2] \\
&= \frac{1}{2\times 3}[(2w_1 - 1)^2 + (w_2 - 2)^2 + (w_2 - 0)^2] \\
&= \frac{2}{3}(w_1 - \frac{1}{2})^2 + \frac{1}{3}(w_2 - 1)^2 + \frac{1}{3}
\end{aligned}$$

2. According to above,

$$\mathcal{E} = \frac{2}{3}(w_1 - \frac{1}{2})^2 + \frac{1}{3}(w_2 - 1)^2 + \frac{1}{3}$$

Setting $\mathcal{E} = 1$,

$$\begin{aligned}
\frac{2}{3}(w_1 - \frac{1}{2})^2 + \frac{1}{3}(w_2 - 1)^2 + \frac{1}{3} &= 1 \\
\frac{(w_1 - \frac{1}{2})^2}{1^2} + \frac{(w_2 - 1)^2}{\sqrt{2}^2} &= 1
\end{aligned}$$

See next page for the graph: