Regularization

- $$\min_{w,b} J(w,b) \tag{1}$$
- $$J=(w,b)=\frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(\hat{y}^{(i)},y^{(i)})+\frac{\lambda}{2m}\|w\|_2^2 \tag{2}$$
- $L2\ regularization$
$$\|w\|_2^2=\sum_{j=1}^{n_x}w_j^2=w^Tw \tag{3}$$
- $L1\ regularization(w\ will\ be\ sparse)$
$$\|w\|_1=\sum_{i=1}^{n_x}|w_i| \tag{4}$$
- $\lambda=regularization\ parameter$

In neural network

- $$J(w^{[1]},b^{[1]},\dots,w^{[L]},b^{[L]})=\frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(\hat{y}^{(i)},y^{(i)})+\frac{\lambda}{2m}\sum_{l=1}^{L}\|w^{[l]}\|^2 \tag{5}$$
- $Frobenius\ norm:\ \|\cdot\|_F^2$
$$\|w^{[l]}\|_F^2=\sum_{i=1}^{n^{[l-1]}}\sum_{j=1}^{n^{[l]}}(w_{ij}^{[l]})^2,\ w:(n^{[l]},n^{[l-1]}) \tag{6}$$
- $now\ in\ backward\ propagation$
$$\begin{aligned} dw^{[l]}&=(from\ backprop)+\frac{\lambda}{m}w^{[l]} \\ w^{[l]}&:=w^{[l]}-\alpha dw^{[l]}=(1-\frac{\alpha\lambda}{m})w^{[l]}-\alpha(from\ backprop)\end{aligned} \tag{7}$$

Why regularization helps

- $\lambda\uparrow,\ w^{[l]}\downarrow$
  - $weaken\ some\ weights\ of\ units\ and\ simplify\ the\ neural\ network$
  - $z^{[l]}\downarrow,\ the\ activation\ functions\ act\ more\ linear,\ thus\ avoiding\ overfitting$
- pay attention to draw the whole term of cost function with the added regularization part