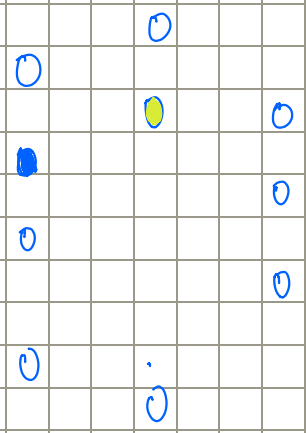
06/10/23

Deep Learning

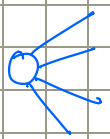
- Recap ✓
- Dropout ✓
- Weight Decay ✓
- ResNet ✓
- 

Dropout: Training – For a given "thinned" network carry out SGD

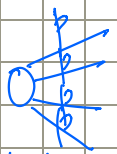
– Choose a new "thinned" network for each training batch.

– Simply don't update weights corresponding to dropped nodes

Test – Multiply the weights at each node with the probability value used to not drop it.



Training: Node retained with prob. p.



Testing: Multiply weights with prob value p.

- Weight decay: $R(\theta) = \sum_{i=1}^{n} d(y^{(i)}, \hat{y}^{(i)})$ – (1)

$$R_{norm}(\theta) = R(\theta) + \frac{\lambda}{2} \|\theta\|_2^2 \quad - (2)$$

Without decay: $\theta^{(r)} = \theta^{(r-1)} - \eta \nabla_{\theta^{(r-1)}} R(\theta)$ – (3) from $R(\theta)$ in (1)

With decay: $\theta^{(r)} = \theta^{(r-1)} - \eta \nabla_{\theta^{(r-1)}} R_{norm}(\theta)$

$$\theta^{(r)} = \theta^{(r-1)} - \eta \nabla_{\theta^{(r-1)}} \left(R(\theta) + \frac{\lambda}{2} \theta^T \theta \right)$$

$$\theta^{(r)} = \underbrace{(1 - \eta\lambda)\theta^{(r-1)}}_{\text{decay of weight}} - \eta \nabla_{\theta^{(r-1)}} R(\theta)$$

Section 7.2 in Goodfellow et al.

— x —

- Residual Network

Regular NN

$$y = \mathcal{H}(x, \theta)$$

$\mathcal{H}$ is the model that captures the relation between x and y

Res Net

$\mathcal{F}(x, \theta) = y - x$ i.e. $\mathcal{F}(x, \theta)$ models the relation between x and the residual $y - x$.

i.e. $\mathcal{H}(x, \theta) = \mathcal{F}(x; \theta) + x$