DL Equations

Sivaraman Sivaraj

Normalization of the Data for Training:

- Log Normalization

$$x' = \log(x+1) \;\; if \; x > 0$$
$$x' = -\log(abs(x)+1) \;\; if \; x < 0$$

- Gaussian Normalization

$$x' = \frac{x - \mu}{\sigma}$$

- Mini-Max Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Activation Function:

- ReLU → $f(a) = \max(0, a)$
- Smooth ReLU → $f(a) = \log(1 + e^{a})$
- Leaky ReLU → $f(a) = \begin{Bmatrix} a, & a > 0 \\ \delta . a, & a \leq 0 \end{Bmatrix}$
- Exponential ReLU → $f(a) = \begin{Bmatrix} a, & a > 0 \\ \delta(e^{a} - 1), & a \leq 0 \end{Bmatrix}$
- Hard tanh function → $f(a) = max[\min(a, 1), -1]$
- Sigmoid → $f(a) = \frac{1}{1+e^{-a}}$
- Tanh → $f(a) = \frac{e^{a} - e^{-a}}{e^{a}+e^{-a}}$

Type of Loss Functions:

- Mean Squared Error

  $$L(\theta) = (y_p - y_a)^2$$

- Mean Squared Logarithmic Error

  $$L(\theta) = (\log(y_a + 1) - \log(y_p + 1))^2$$

- Mean Absolute Error

  $$L(\theta) = \left|y_a - y_p\right|$$

- Mean Absolute Percentage Error

  $$L(\theta) = 100 * \left|\frac{(y_a - y_p)}{y_a}\right|$$

- Kullback Leibler Divergence Error

  $$L(\theta) = y_a * log\left(\frac{y_a}{y_p}\right)$$

- Logarithmic Hyperbolic Cosine Error

  $$L(\theta) = \log\left(\cosh\left(y_p - y_a\right)\right) \approx \left({x^2}/_2 \quad (\mathrm{x} \ll), |x| - \log(2)\,(x \gg)\right)$$

- Hinge Loss
  - $L(\theta) = \max(1 - \left(y_p * y_a\right), 0)$

- Poisson Error
  - $L(\theta) = y_p - y_a * \log(y_p)$

- Squared Hinge Loss
  - $L(\theta) = (\max(1 - \left(y_p * y_a\right), 0))^2$

- Huber Loss
  - $L(\theta) = (\frac{x^2}{2})\ if\ x \leq d\ ,\ \left(d * |x| - \left(\frac{d^2}{2}\right)\right) if\ x > d$

Choosing an optimal loss function is important one

Optimizers:

- Stochastic Gradient Descent (SGD):
  - $w(m+1) = w(m) - \eta . \frac{\partial L}{\partial w}$
- Nesterov Accelerated Gradient (NAG):
  - $w(m+1) = w(m) - \eta . g_w(m) - \alpha . \eta . g_w(m-1)$ where, $g_w(m) = (1 + \alpha + \alpha^2 + \; \ldots . + \alpha^m) \frac{\partial L_m}{\partial w}$
- Ada Grad:
  - $w(m+1) = w(m) - \Delta w(m)$
  - $\Delta w(m) = \frac{-\eta}{\epsilon + \sqrt{r_w(m)}} \, g_w(m), \qquad r_w(m) = g_w(0)^2 + g_w(1)^2 + g_w(2)^2 + \cdots \ldots \ldots \ldots \ldots . + g_w(m-1)^2$
- RMS Prob:
  - $r_w(m) = \rho_g \left[\frac{1}{L} \sum_{l=1}^{L} g_w^2(m-l) + (1-\rho_g) g_w^2(m) \right]$
- Ada Delta:
  - $\Delta w(m) = \frac{-\, RMS\; value\; of\; \Delta w}{RMS\; value\; of\; g_w^2 + \epsilon} \quad \Delta w = \sqrt{\rho_\Delta \left[\frac{1}{L} \sum_{l=1}^{L} \Delta w(m-l-1)^2 \right] + (1-\rho_\Delta) \Delta w(m-l)^2}$

Adam's Method:

- $w(m+1) = w(m) - \Delta w(m)$
- $\Delta w(m) = -\eta \, \frac{\hat{q}_w(m)}{\epsilon + \sqrt{\hat{r}_w(m)}}$
- $\hat{r}_w(w) = \frac{r_w(m)}{1-\rho_2^m}$, $\hat{q}_w(m) = \frac{q_m(m)}{1-\rho_1^m}$
- $q_w(m) = \rho_1 \, q_w(m-1) + (1 - \rho_1) g_w(m)$
- $r_w(m) = \rho_2 \, r_w(m-1) + (1-\rho_2) \, g_w^2(m)$