50.039 – Theory and Practice of Deep Learning

Alex

Week 04: Initialization

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity, every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

1 Neural Net initialization

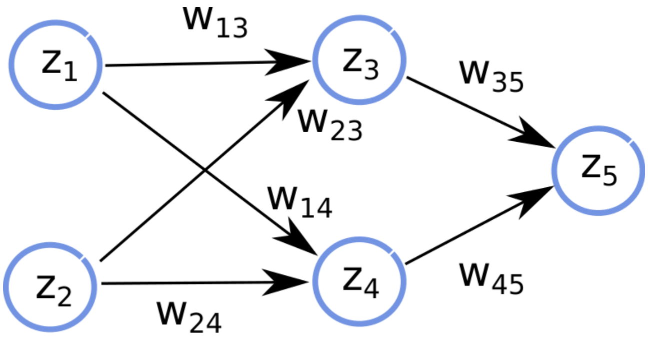
`http://neuralnetworksanddeeplearning.com/chap3.html#weight_initialization`
need to initialize parameters w of a neural network. Three guidelines:

Key take aways

- initialize weight vectors to random numbers for symmetry breaking
- initialize weight vectors with a variance which decreases as a function if inputs and outputs to a neuron. Typical choices are to draw weights from a zero mean normal distribution with variance equal to $\frac{1}{D_{in}}$ or $\frac{2}{D_{in}+D_{out}}$ (Glorot and Bengio, 2010, Understanding the difficulty of training deep feedforward neural networks)
- biases are set usually to zero
- later: use transfer learning

1.1 Neural Net initialization: symmetry breaking

The weights of different neurons should receive different values. The goal is to allow different neurons to learn to become detectors for different structures.
Consider a fully symmetrically initialized neural network:



If $w_{13} = w_{14}$ and $w_{23} = w_{24}$, then the neuron activations of z_3 and z_4 are the same. If now also $w_{35} = w_{45}$, then we would have identically gradient updates for w_{13} versus w_{14}, as well as for w_{23} versus w_{24}. This would imply, that the weights of w_{13} versus w_{14} change in the same way, and all the neurons will never start to produce different outputs.

$$
\begin{aligned}
\frac{\partial L}{\partial w_{13}} &= \frac{\partial L}{\partial z_5}\frac{\partial z_5}{\partial z_3}\frac{\partial z_3}{\partial w_{13}} \\
\frac{\partial L}{\partial w_{23}} &= \frac{\partial L}{\partial z_5}\frac{\partial z_5}{\partial z_4}\frac{\partial z_4}{\partial w_{14}} \\
\frac{\partial z_5}{\partial z_3} &= \sigma'(w_{35}z_3 + w_{45}z_4)w_{35} \\
\frac{\partial z_5}{\partial z_4} &= \sigma'(w_{35}z_3 + w_{45}z_4)w_{34} \\
w_{35} = w_{34} \Rightarrow \frac{\partial z_5}{\partial z_3} &= \frac{\partial z_5}{\partial z_4} \;!! \\
\frac{\partial z_3}{\partial w_{13}} &= \sigma'(w_{13}z_1 + w_{23}z_2)z_1 \\
\frac{\partial z_4}{\partial w_{14}} &= \sigma'(w_{14}z_1 + w_{24}z_2)z_1 \\
w_{13} = w_{14}, w_{23} = w_{24} \Rightarrow \frac{\partial z_3}{\partial w_{13}} &= \frac{\partial z_4}{\partial w_{14}} \\
\Rightarrow \frac{\partial L}{\partial w_{13}} &= \frac{\partial L}{\partial w_{23}}
\end{aligned}
$$

1.2 Neural Net initialization: choice of variance

How to arrive at drawing weights from a zero mean normal distribution with variance equal to $\frac{1}{D_{in}}$?
consider a linear function

$$y = \sum_{d=1}^{D_{in}} w_d x_d$$

The idea is that one wants to have the output variance equal to the input variance: $Var(y) = Var(X)$, while assuming that the inputs are normalized to have zero mean: $E[x] = 0$, and we intend to have weights drawn from a zero-mean gaussian, thus $E[w] = 0$

Lets consider the variance of y as a function of variances of weights and inputs x_d.
If the terms $w_d x_d$ would be statistically independent and identically distributed, then

$$Var(y) = Var(\sum_{d=1}^{D_{in}} w_d x_d) = \sum_{d=1}^{D_{in}} Var(w_d x_d) = D_{in} Var(w_1 x_1)$$

D_{in} appears already, but we have $Var(w_1 x_1)$ in an entangled term. What is $Var(w_1 x_1)$? Lets assume w and x are independent, so that $E[wx] = E[w]E[x]$

$$\begin{aligned} Var(wx) &= E[(wx)^2] - (E[wx])^2 = E[w^2 x^2] - E[w]^2 E[x]^2 = E[w^2]E[x^2] - E[w]^2 E[x]^2 \\ &= Var(W)Var(X) + Var(W)E[X]^2 + Var(X)E[w]^2 \end{aligned}$$

Now assume that the inputs are zero mean, thus $E[X] = 0$, and we intend to initialize the weights W as gaussians with zero mean, thus $E[W] = 0$. then we arrive at:

$$Var(y) = D_{in} Var(w_1 x_1) = D_{in} Var(W) Var(X)$$

Now if one wants to have the output variance equal to the input variance: $Var(y) = Var(X)$, ones arrives at

$$1 = D_{in} Var(W)$$

How to arrive at $\frac{2}{D_{in}+D_{out}}$?

Note, that this is the harmonic mean of $\frac{1}{D_{in}}$ and $\frac{1}{D_{out}}$, a compromise between these two terms. How to get to $\frac{1}{D_{out}}$? For this one needs to look at the variance in backpropagation:

$$\frac{\partial L}{\partial z_k} = \sum_{d=1}^{D_{out}} \frac{\partial L}{\partial z_d} \frac{\partial z_d}{\partial z_k} = \sum_{d=1}^{D_{out}} \frac{\partial L}{\partial z_d} z_d'(z) w_{kd}$$

In case of independence of terms we arrive at

$$Var(\frac{\partial L}{\partial z_k}) = D_{out} Var(\frac{\partial L}{\partial z_d} z_d'(z) w_{kd})$$