CS 615 – Deep Learning

Chaining Gradients

Slides adapted from material created by E. Alpaydin
Prof. Mordohai, Prof. Greenstadt, Pattern Classification (2nd Ed.),
Pattern Recognition and Machine Learning

Objectives

- Finding gradient rules using the chain rule.

Chain Rule

- We are about to get into our first type of deep learning structure, *artificial neural networks.*
- However, deep learning structures tend to make use of the *chain rule* from calculus in order to determine the gradient rules.
- So let's first review the chain rule as it applies to our gradient based learning!

Chain Rule for LSE

- Let's assume a squared error objective function, and $g(x|\theta)$ is our activation function.
- Our objective function is then:

$$J = \left(y - g(x|\theta)\right)^2$$

- From this we obviously want $\frac{\partial J}{\partial \theta_i}$
- Using the *chain rule* we can write this as:

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial g(x|\theta)} \cdot \frac{\partial g(x|\theta)}{\partial \theta_i}$$

Chain Rule for LSE w/ Linear Function

$$J = \big(y - g(x|\theta)\big)^2, \qquad \frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial g(x|\theta)} \cdot \frac{\partial g(x|\theta)}{\partial \theta_i}$$

- What is $\frac{\partial J}{\partial g(x|\theta)}$?

$$\frac{\partial J}{\partial g(x|\theta)} = -2\big(y - g(x|\theta)\big)$$

- What is $\frac{\partial g(x|\theta)}{\partial \theta_i}$?
  - Depends on $g(x|\theta)$.
  - Let's assume we have a simple linear function: $g(x|\theta) = x\theta$
  - Then $\frac{\partial g(x|\theta)}{\partial \theta_i} = x_i$
- Putting it together, we have

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial g(x|\theta)} \cdot \frac{\partial g(x|\theta)}{\partial \theta_i} = -2\big(y - g(x|\theta)\big)x_i$$

Chain Rule for LSE w/ Logistic Function

$$J = \big(y - g(x|\theta)\big)^2$$

- What if we used a logistic activation function: $g(x|\theta) = \frac{1}{1+e^{-x\theta}}$
- The first term of the chain rule is the same!

$$\frac{\partial J}{\partial \theta_i} = -2\big(y - g(x|\theta)\big) \cdot \frac{\partial g(x|\theta)}{\partial \theta_i}$$

- And recall that the partial of the logistic function is $\frac{\partial g(x|\theta)}{\partial \theta_i} = x_i g(x|\theta)\big(1 - g(x|\theta)\big)$
- Putting it together, we have

$$\frac{\partial J}{\partial \theta_i} = -2\big(y - g(x|\theta)\big)x_i g(x|\theta)\big(1 - g(x|\theta)\big)$$

Chain Rule for Cross-Entropy w/ Softmax

- Let's do one more!
- Softmax activation function with cross-entropy objective function.
- Recall that given our target class a, our cross-entropy objective function is:

$$J = -\ln\left(g\left(x|\theta_{:,a}\right)\right)$$

- Now let's find the partial of this, with respect to $\theta_{i,j}$ leveraging the chain rule!

Chain Rule for Cross-Entropy w/ Softmax

$$J = -\ln\left(g\left(x|\theta_{:,a}\right)\right)$$

- $\frac{\partial J}{\partial \theta_{i,j}} = \frac{\partial J}{\partial g\left(x|\theta_{:,a}\right)} \cdot \frac{\partial g\left(x|\theta_{:,a}\right)}{\partial \theta_{i,j}}$
- For the first term:

$$\frac{\partial J}{\partial g\left(x|\theta_{:,a}\right)} = -\frac{1}{g\left(x|\theta_{:,a}\right)} = -\frac{1}{\hat{y}_a}$$

Chain Rule for Cross-Entropy w/ Softmax

$$J = -\ln\left(g\left(x|\theta_{:,a}\right)\right)$$

$$\frac{\partial J}{\partial \theta_{i,j}} = \frac{\partial J}{\partial g\left(x|\theta_{:,a}\right)} \cdot \frac{\partial g\left(x|\theta_{:,a}\right)}{\partial \theta_{i,j}}$$

- For the second term, recall our softmax activation function:

$$g\left(x|\theta_{:,j}\right) = \frac{e^{x\theta_{:,j}}}{\sum_{k=1}^{K} e^{x\theta_{:,k}}}$$

- And it's two cases:
  - If $j \neq a$:

$$\frac{\partial g\left(x|\theta_{:,a}\right)}{\partial \theta_{i,j}} = -\frac{e^{x\theta_{:,a}} x_i e^{x\theta_{:,j}}}{\left(\sum_{k=1}^{K} e^{x\theta_{:,k}}\right)^2} = -x_i g\left(x|\theta_{:,a}\right) g\left(x|\theta_{:,j}\right)$$

  - If $j = a$:

$$\frac{\partial g\left(x|\theta_{:,a}\right)}{\partial \theta_{i,j}} = \frac{\left(x_i e^{x\theta_{:,j}} \sum_{k=1}^{K} e^{x\theta_{:,k}}\right) - \left(e^{x\theta_{:,j}} x_i e^{x\theta_{:,j}}\right)}{\left(\sum_{k=1}^{K} e^{x\theta_{:,k}}\right)^2} = x_i \left(g\left(x|\theta_{:,j}\right) - g\left(x|\theta_{:,j}\right)^2\right)$$

Chain Rule for Cross-Entropy w/ Softmax

$$\frac{\partial J}{\partial \theta_{i,j}} = \frac{\partial J}{\partial g(x|\theta_{:,a})} \cdot \frac{\partial g(x|\theta_{:,a})}{\partial \theta_{i,j}}$$

- Putting it together:

$$\frac{\partial J}{\partial \theta_{i,j}} = -\frac{1}{g(x|\theta_{:,a})} \cdot \begin{cases} -x_i g(x|\theta_{:,a}) g(x|\theta_{:,j}) & if\ j \neq a \\ x_i \left(g(x|\theta_{:,j}) - g(x|\theta_{:,j})^2 \right) & otherwise \end{cases}$$

$$\frac{\partial J}{\partial \theta_{i,j}} = \begin{cases} x_i g(x|\theta_{:,j}) & if\ j \neq a \\ x_i \left(-1 + g(x|\theta_{:,j}) \right) & otherwise \end{cases}$$

Chain Rule for Cross-Entropy w/ Softmax

$$\frac{\partial J}{\partial \theta_{i,j}} = \begin{cases} x_i g\left(x|\theta_{:,j}\right) & if\ j \neq a \\ x_i\left(-1 + g\left(x|\theta_{:,j}\right)\right) & otherwise \end{cases}$$

- If we have a binary classifier such that $y \in \{0,1\}$ then we can write this as:

$$\frac{\partial J}{\partial \theta_{i,j}} = x_i\left(g\left(x|\theta_{:,j}\right) - y\right)$$