DS-GA 1008: Deep Learning, Spring 2019
Homework Assignment 1

Lekha Iyengar

February 2019

1 Backprop

1.1 Warm-up

$$y = Wx + b$$

$\frac{\partial L}{\partial W}$ =? and $\frac{\partial L}{\partial b}$ =?
L is a function of y and y is a function of W We use chain rule to compute the derivative of L with respect to W:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} . \frac{\partial y}{\partial W}$$

$$y_i = \sum_{j=1}^{N} W_{i,j} x_j + b_i$$

$$\frac{\partial L}{\partial y} = \left[\frac{\partial L}{\partial y_1}, \frac{\partial L}{\partial y_2}, \ldots \frac{\partial L}{\partial y_T} \right]$$

The derivative of y_i w.r.t each element in W is x_j when the element is in row i and 0 otherwise If we split the index of W to i and j we get

$$D_{i,j} y_t = \frac{\partial (\sum_{j=1}^{N} (W_{t,j} x_j + b_t))}{\partial W_{i,j}}$$

$$D_{i,j} y_t = \begin{cases} x_j & i = t \\ 0 & i \neq t \end{cases}$$

Overall we get the Jacobian matrix

$$Dy = \begin{bmatrix} \mathrm{x}_1 & \mathrm{x}_2 & \ldots & \mathrm{x}_N & \ldots & 0 & 0 & \ldots\ 0 \\ 0 & \mathrm{x}_1 & \ldots & \mathrm{x}_{N-1} & \ldots & 0 & 0 & \ldots\ 0 \\ . & . & & . & & . & & \\ . & & . & & . & & . & \\ . & & & . & & . & & . \\ 0 & 0 & \ldots & 0 & \mathrm{x1} & \mathrm{x2} & \ldots & \mathrm{x}_N \end{bmatrix}$$

$$\frac{\partial L}{\partial W} = \begin{bmatrix} \mathrm{x}_1\frac{\partial L}{\partial y_1} & \mathrm{x}_2\frac{\partial L}{\partial y_1} & ... & \mathrm{x}_T\frac{\partial L}{\partial y_1} \\ \mathrm{x}_1\frac{\partial L}{\partial y_2} & \mathrm{x}_2\frac{\partial L}{\partial y_2} & ... & \mathrm{x}_T\frac{\partial L}{\partial y_2} \\ . & & & \\ . & & & \\ . & & & \\ \mathrm{x}_1\frac{\partial L}{\partial y_T} & \mathrm{x}_2\frac{\partial L}{\partial y_T} & ... & \mathrm{x}_T\frac{\partial L}{\partial y_T} \end{bmatrix}$$

$$= \frac{\partial L}{\partial y} \otimes x$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} X^T$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial b}$$

$$y_i = \sum_{j=1}^{N} W_{i,j}x_j + b_i$$

$$\frac{\partial y_i}{\partial b_j} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Dy(b) is an identity matrix with dimension(T,T)

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial y}$$

1.2 Softmax

$$y_j = \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}$$

$$\frac{\partial y_j}{\partial x_i} = \frac{\partial \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}}{\partial x_i}$$

From quotient rule for $f(x) = \frac{g(x)}{h(x)}$

$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h^2(x)}$ where $g(x) = e^{\beta x_j}$ and $h(x) = \sum_i e^{\beta x_i}$

$h'(x)$ will always be $\beta e^{\beta x_i}$ as it will always have $e^{\beta x_i}$ term. $g'(x)$ will be $\beta e^{\beta x_i}$ only if $i = j$ otherwise 0.

if $i = j$,

$$\frac{\partial y_j}{\partial x_i} = \frac{\partial \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}}{\partial x_i}$$

$$= \frac{\beta e^{\beta x_j}\sum_i e^{\beta x_i} - \beta e^{\beta x_i}e^{\beta x_j}}{(\sum_i e^{\beta x_i})^2}$$

$$= \frac{\beta e^{\beta x_j}}{\sum_i e^{\beta x_i}} \frac{(\sum_i e^{\beta x_i} - e^{\beta x_i})}{\sum_i e^{\beta x_i}}$$

$$= \beta y_i (1 - \frac{e^{\beta x_i}}{\sum_i e^{\beta x_i}})$$

$$= \beta y_i (1 - \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}})$$

$$= \beta y_i (1 - y_j)$$

if $i \neq j$,

$$\frac{\partial y_j}{\partial x_i} = \frac{0 - \beta e^{\beta x_i} e^{\beta x_j}}{(\sum_i e^{\beta x_i})^2}$$

$$= -\beta \frac{e^{\beta x_i}}{\sum_i e^{\beta x_i}} \frac{e^{\beta x_j}}{\sum_i e^{\beta x_i}}$$

$$= -\beta y_i y_j$$

Using Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\frac{\partial y_j}{\partial x_i} = \beta y_i (\delta_{ij} - y_j)$$