CS231N Assignment 1 - Matrix Calculus, 2-layer NN

1 Matrix Calculus

1.1 Derivative of Vector wrt Matrix

(1.1) We have vector $\boldsymbol{r} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$, matrix $\boldsymbol{A} = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \end{bmatrix}$

(1.2) We have that $\dfrac{\partial \boldsymbol{r}}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial r_1}{\partial \boldsymbol{A}} \\ \frac{\partial r_2}{\partial \boldsymbol{A}} \end{bmatrix}$, where each $\dfrac{\partial r_i}{\partial \boldsymbol{A}}$ is a Scalar-Matrix derivative

(2.1) If we view $\dfrac{\partial \boldsymbol{r}}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial r_1}{\partial \boldsymbol{A}} \\ \frac{\partial r_2}{\partial \boldsymbol{A}} \end{bmatrix} = \begin{bmatrix} \frac{\partial r_1}{\partial w_1} & \frac{\partial r_1}{\partial w_2} & \frac{\partial r_1}{\partial w_3} & \frac{\partial r_1}{\partial w_4} & \frac{\partial r_1}{\partial w_5} & \frac{\partial r_1}{\partial w_6} \\ \frac{\partial r_2}{\partial w_1} & \frac{\partial r_2}{\partial w_2} & \frac{\partial r_2}{\partial w_3} & \frac{\partial r_2}{\partial w_4} & \frac{\partial r_2}{\partial w_5} & \frac{\partial r_2}{\partial w_6} \end{bmatrix}$, this is termed as 'flattening' the matrix

(2.2) Notice that if $\boldsymbol{r} \in \mathbb{R}^{m*1}$ and $\boldsymbol{A} \in \mathbb{R}^{n*k}$, we have that $\dfrac{\partial \boldsymbol{r}}{\partial \boldsymbol{A}} \in \mathbb{R}^{m*(n*k)}$

(3a) Alternatively, we could view $\dfrac{\partial \boldsymbol{r}}{\partial \boldsymbol{A}} = \begin{bmatrix} \frac{\partial r_1}{\partial \boldsymbol{A}} \\ \frac{\partial r_2}{\partial \boldsymbol{A}} \end{bmatrix}$ as a rank-3 tensor

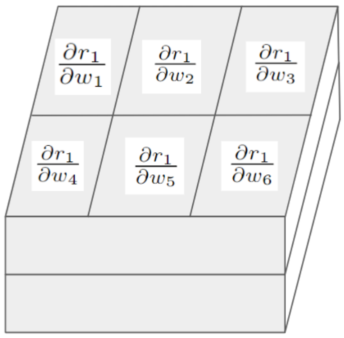

Figure 1: Rank-3 tensor

1.2 Derivative of Matrix wrt Matrix

(1a) We have matrix $\boldsymbol{A} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$, matrix $\boldsymbol{W} = \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \end{bmatrix}$

(1b) We have that $\dfrac{\partial \boldsymbol{A}}{\partial \boldsymbol{W}} = \begin{bmatrix} \frac{\partial a_1}{\partial \boldsymbol{W}} & \frac{\partial a_2}{\partial \boldsymbol{W}} \\ \frac{\partial a_3}{\partial \boldsymbol{W}} & \frac{\partial a_4}{\partial \boldsymbol{W}} \end{bmatrix}$, where each $\dfrac{\partial a_i}{\partial \boldsymbol{A}}$ is a Scalar-Matrix derivative

(1c) Notice that if $\boldsymbol{A} \in \mathbb{R}^{m*n}$ and $\boldsymbol{W} \in \mathbb{R}^{k*l}$; then $\dfrac{\partial a_i}{\boldsymbol{W}} \in \mathbb{R}^{k*l}$, $\dfrac{\partial \boldsymbol{A}}{\partial \boldsymbol{W}} \in \mathbb{R}^{(k*l*m)*(n)}$

1.3 Derivative of Vector wrt Matrix - Alternative methods

Suppose that $\boldsymbol{W} \in \mathbb{R}^{n*m}$ and $\boldsymbol{x} \in \mathbb{R}^{m*1}$. How do we calculate $\frac{d\boldsymbol{W}\boldsymbol{x}}{d\boldsymbol{W}}$?

We know that the quantity in question is a 3^{rd} order tensor.

1.3.1 Index Notation

(1a) We know that $\boldsymbol{f} = \boldsymbol{W}\boldsymbol{x}$

(1b) Define $\boldsymbol{f}_i = \boldsymbol{W}_{ij}\boldsymbol{x}^j$; Note that we are using Einstein summation notation

(1c) $\dfrac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{W}_{mn}} = \dfrac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{W}_{ij}} \dfrac{\partial \boldsymbol{W}_{ij}}{\partial \boldsymbol{W}_{mn}} = \dfrac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{W}_{ij}} \boldsymbol{x}_j = \delta_{im}\delta_{jn}\boldsymbol{x}_j = \delta_{im}\boldsymbol{x}_n$

1.3.2 Vectorization

(1)
$$\begin{aligned} \boldsymbol{f} &= \boldsymbol{W}\boldsymbol{x} \\ &= \boldsymbol{I}\boldsymbol{W}\boldsymbol{x} \\ &= (\boldsymbol{x}^T \otimes \boldsymbol{I})\ vec(\boldsymbol{W}) \\ &= (\boldsymbol{x}^T \otimes \boldsymbol{I})\ \boldsymbol{w} \end{aligned}$$

(2) Thus, $\dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{w}} = (\boldsymbol{x}^T \otimes \boldsymbol{I})$

1.3.3 Special Case

See matrixDifferentiation.pdf for more

2 2-layer NN

2.1 Backpropagation - Bias terms

Note that we are considering backpropagation w.r.t a single **layer**, where a single layer may encapsulate **more than one neuron**

(1a) Input data $\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \end{bmatrix} \in \mathbb{R}^{N*D}$; $N =$ numSamples, $D =$ dataDimension

(1b) Weights $\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \\ w_{41} & w_{42} & w_{43} \end{bmatrix} \in \mathbb{R}^{D*C}$; $C =$ number of classes

(1c) Bias vector $\boldsymbol{b} = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix} \in \mathbb{R}^C$

(1d) Bias matrix $\boldsymbol{B} = \begin{bmatrix} b_1 & b_2 & b_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \in \mathbb{R}^{N*C}$; Bias vector is 'broadcast' once for each data sample

(1e) Upstream term $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{B}$

$$= \begin{bmatrix} xw_{11} & xw_{12} & xw_{13} \\ xw_{21} & xw_{22} & xw_{23} \end{bmatrix} + \begin{bmatrix} b_1 & b_2 & b_3 \\ b_1 & b_2 & b_3 \end{bmatrix}$$

$$= \begin{bmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \end{bmatrix} \in \mathbb{R}^{N*C}$$

(2a) $\dfrac{\partial L_i}{\partial \boldsymbol{b}} = \dfrac{\partial L_i}{\partial \boldsymbol{Z}_i} \cdot \dfrac{\partial \boldsymbol{Z}_i}{\partial \boldsymbol{b}}$; $\boldsymbol{Z}_i$ stands for i^{th} row of $\boldsymbol{Z}$ (corresponding to i^{th} sample), L_i stands for loss of i^{th} sample

$$= \frac{\partial L_i}{\partial \boldsymbol{Z}_i} \cdot \frac{\partial}{\partial \boldsymbol{b}} \begin{bmatrix} z_{i1} & z_{i2} & z_{i3} \end{bmatrix}$$

$$= \frac{\partial L_i}{\partial \boldsymbol{Z}_i} \cdot \begin{bmatrix} \frac{\partial z_{i1}}{\partial b_1} & \frac{\partial z_{i1}}{\partial b_2} & \frac{\partial z_{i1}}{\partial b_3} \\ \frac{\partial z_{i2}}{\partial b_1} & \frac{\partial z_{i2}}{\partial b_2} & \frac{\partial z_{i2}}{\partial b_3} \\ \frac{\partial z_{i3}}{\partial b_1} & \frac{\partial z_{i3}}{\partial b_2} & \frac{\partial z_{i3}}{\partial b_3} \end{bmatrix} = \frac{\partial L_i}{\partial \boldsymbol{Z}_i} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \frac{\partial L_i}{\partial \boldsymbol{Z}_i} \in \mathbb{R}^{1*C}$$

(2b) Thus, $\dfrac{\partial L}{\partial \boldsymbol{b}} = \sum_i \dfrac{\partial L_i}{\partial \boldsymbol{b}} = \sum_i \dfrac{\partial L_i}{\partial z_i}$