1 导言(Introduction)

我们中的大部分人是在学校最后一次见到微积分(calculus),但是导数是机器学习的重要部分，尤其在深度神经网络领域，它通过优化loss function 来训练模型。随便一篇机器学习论文，或者深度学习框架的文档，其中涉及的不仅是数值积分，还有矩阵积分，它是线性代数和多变量微分的结合。
通过现代化的机器学习框架，你可以在仅仅掌握数值积分的水平下达到世界级深度学习参与者。如果需要理解这些库的底层实现，或者理解一些前言的训练技术，你需要理解矩阵微积分的特定部分。

xx

如上图所示，对于一个神经网络计算单元，$z(\mathbf{x}) = \sum_i^n w_i x_i + b = \mathbf{w} \cdot \mathbf{x} + b$
函数$F(\mathbf{x})$称为放射函数(affine function),其后跟着一个线性修正单元，也即是激活函数;它将负值修改为0:
$max(0, z(\mathbf{x}))$。
神经网络由这些单元组成，这些单元被组织为$layers$
训练神经网络也即是通过最小化$loss function$选择合适的$\mathbf{w}$和偏置b.优化算法包括
$SGD, SGD\ with\ momentum, Adam$。这里需要获取$activation(\mathbf{x})$相对于$\mathbf{w}$和$b$的导数。
例如，均方差损失如下：

$$\frac{1}{N}\sum_{\mathbf{x}}(\text{target}(\mathbf{x}) - \text{activation}(\mathbf{x}))^2 = \frac{1}{N}\sum_{\mathbf{x}}\left(\text{target}(\mathbf{x}) - \max\left(0, \sum_i^{|x|} w_i x_i + b\right)\right)^2$$

2 数值函数求导法则(Scalar derivative rules)

法则	数学符号	对于x的导数

法则	数学符号	对于x的导数
Constant	c	0
数乘	cf	$c\frac{df}{dx}$
指数	x^n	nx^{n-1}
加法	$f+g$	$\frac{df}{dx}+\frac{dg}{dx}$
减法	$f-g$	$\frac{df}{dx}-\frac{dg}{dx}$
乘法	fg	$\frac{df}{dx}g+\frac{dg}{dx}f$
链式	$f(g(x))$	$\frac{df(u)}{du}\frac{du}{dx}$

3 向量积分和偏导数(vector calculus and partial derivatives)

对于多变量函数$f(x,y)=3x^2y$，它的梯度可以由如下向量表示

$$\nabla f(x,y)=\left[\frac{\partial f(x,y)}{\partial x},\frac{\partial f(x,y)}{\partial y}\right]=[6yx,3x^2]$$

这里处理的是 $\vec{x}$ 向数值 z 的映射，下面的矩阵积分将处理n维向m维的映射

4 矩阵积分(Matrix calculus)

首先引入$g(x,y) = 2x + y^8$

$$\frac{\partial g(x,y)}{\partial x} = \frac{\partial 2x}{\partial x} + \frac{\partial y^8}{\partial x} = 2\frac{\partial x}{\partial x} + 0 = 2 \times 1 = 2$$

$$\frac{\partial g(x,y)}{\partial y} = \frac{\partial 2x}{\partial y} + \frac{\partial y^8}{\partial y} = 0 + 8y^7 = 8y^7$$

对于该函数的梯度表示如下：

$$\nabla g(x,y) = \left[2, 8y^7\right]$$

通过将两个函数的梯度叠放到一个矩阵里面，可以得到如下结果：

$$J = \begin{bmatrix} \nabla f(x,y) \\ \nabla g(x,y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} & \frac{\partial f(x,y)}{\partial y} \\ \frac{\partial g(x,y)}{\partial x} & \frac{\partial g(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 6yx & 3x^2 \\ 2 & 8y^7 \end{bmatrix}$$

这种放置方法成为**numerator layout**,也有很多文献使用**denominator layout**方法，该方法为**numerator layout**的Transpose.

4.1 Jacobian 的推广

对于多元函数，我们可以将其推广到向量方程

$$f(x,y,z) \Rightarrow f(\mathbf{x})$$

黑体表示向量$\mathbf{x}$, 斜体字为数值 x;
假定所有向量为列向量,也即是$n \times 1$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

对于多个标量函数，我们可以将其合并为向量。

$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

其中$|\mathbf{x}| = n$, $|\mathbf{y}| = m$
类似如下表示：

$$
\begin{aligned}
y_1 &= f_1(\mathbf{x}) \\
y_2 &= f_2(\mathbf{x}) \\
&\vdots \\
y_m &= f_m(\mathbf{x})
\end{aligned}
$$

以下是简单的例子

$\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{x}$,对应的标量函数如下

$$
\begin{aligned}
y_1 &= f_1(\mathbf{x}) = x_1 \\
y_2 &= f_2(\mathbf{x}) = x_2 \\
&\vdots \\
y_n &= f_n(\mathbf{x}) = x_n
\end{aligned}
$$

通常来说Jacobian矩阵包含$m \times n$个可能的偏导数，m 对应标量函数的数量，n 对应输入向量的维度

$$
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \nabla f_1(\mathbf{x}) \\ \nabla f_2(\mathbf{x}) \\ \cdots \\ \nabla f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} f_2(\mathbf{x}) \\ \cdots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}) \\ & & \cdots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \frac{\partial}{\partial x_2} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix}
$$

每一个$\frac{\partial}{\partial \mathbf{x}} f_i(\mathbf{x})$对应一个水平的向量。

对于函数$\mathbf{f}(\mathbf{x}) = \mathbf{x}$,也即是 $f_i(\mathbf{x}) = x_i$, 其对应的Jacobian矩阵如下：

$$
\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial \mathbf{x}} f_1(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}} f_2(\mathbf{x}) \\ \cdots \\ \frac{\partial}{\partial \mathbf{x}} f_m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}) \\ & & \cdots & \\ \frac{\partial}{\partial x_1} f_m(\mathbf{x}) & \frac{\partial}{\partial x_2} f_m(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} f_m(\mathbf{x}) \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{\partial}{\partial x_1} x_1 & \frac{\partial}{\partial x_2} x_1 & \cdots & \frac{\partial}{\partial x_n} x_1 \\ \frac{\partial}{\partial x_1} x_2 & \frac{\partial}{\partial x_2} x_2 & \cdots & \frac{\partial}{\partial x_n} x_2 \\ & \cdots & & \\ \frac{\partial}{\partial x_1} x_n & \frac{\partial}{\partial x_2} x_n & \cdots & \frac{\partial}{\partial x_n} x_n \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} x_1 & 0 & \cdots & 0 \\ 0 & \frac{\partial}{\partial x_2} x_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \frac{\partial}{\partial x_n} x_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}
$$

4.2 element-wise 二元操作

element-wise 二元操作是类似向量逐个元素相加。例如 $max(\mathbf{w}, \mathbf{x})$或者$\mathbf{w} > \mathbf{x}$.

当然，我们也可以推广元素级别的操作，使用如下符号表示$\mathbf{y} = \mathbf{f}(\mathbf{w}) \bigcirc \mathbf{g}(\mathbf{x})$, 这也意味着 输出向量$\mathbf{y}$ 同输入向量$\mathbf{x}$ 维度相同均为n.

展开如下

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x}) \\ f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x}) \end{bmatrix}$$

关于w的Jacobian矩阵如下

$$J_{\mathrm{w}} = \frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial}{\partial w_1}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) & \frac{\partial}{\partial w_2}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial w_n}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) \\ \frac{\partial}{\partial w_1}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) & \frac{\partial}{\partial w_2}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial w_n}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) \\ & \cdots & & \\ \frac{\partial}{\partial w_1}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) & \frac{\partial}{\partial w_2}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial w_n}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) \end{bmatrix}$$

类似，可以得到关于$\mathbf{x}$的矩阵:

$$J_{\mathrm{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial}{\partial x_1}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) & \frac{\partial}{\partial x_2}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial x_n}\left(f_1(\mathbf{w}) \bigcirc g_1(\mathbf{x})\right) \\ \frac{\partial}{\partial x_1}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) & \frac{\partial}{\partial x_2}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial x_n}\left(f_2(\mathbf{w}) \bigcirc g_2(\mathbf{x})\right) \\ & \cdots & & \\ \frac{\partial}{\partial x_1}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) & \frac{\partial}{\partial x_2}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) & \cdots & \frac{\partial}{\partial x_n}\left(f_n(\mathbf{w}) \bigcirc g_n(\mathbf{x})\right) \end{bmatrix}$$

由于element-wise 操作的性质，可以得到$\hat{f}_i(w_i) = f_i(\mathbf{w})$，也即是 $\frac{\partial}{\partial x_i}(f_k(\mathbf{w}) \bigcirc f_k(\mathbf{x})) = \frac{\partial}{\partial x_i}(f_k(w_j) \bigcirc f_k(x_j))$, 由导数相关法则可知, 若 $i \neq j$， 则 $\frac{\partial}{\partial x_i}(f_k(w_j) \bigcirc f_k(x_j)) = 0$, 上述公式可以简化为如下:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \begin{bmatrix} \frac{\partial}{\partial w_1}(f_1(w_1) \bigcirc g_1(x_1)) & & & \mathbf{0} \\ & \frac{\partial}{\partial w_2}(f_2(w_2) \bigcirc g_2(x_2)) & & \\ & & \cdots & \\ \mathbf{0} & & & \frac{\partial}{\partial x_n}(f_n(w_n) \bigcirc g_n(x_n)) \end{bmatrix}$$

也可以使用如下的简洁表示方式

$$\frac{\partial \mathbf{y}}{\partial \mathbf{w}} = \operatorname{diag}\left(\frac{\partial}{\partial w_1}\left(f_1\left(w_1\right) \bigcirc g_1\left(x_1\right)\right), \frac{\partial}{\partial w_2}\left(f_2\left(w_2\right) \bigcirc g_2\left(x_2\right)\right), \ldots, \frac{\partial}{\partial w_n}\left(f_n\left(w_n\right) \bigcirc g_n\left(x_n\right)\right)\right)$$

关于$\mathbf{x}$可以有类似表达

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \operatorname{diag}\left(\frac{\partial}{\partial x_1}\left(f_1\left(w_1\right) \bigcirc g_1\left(x_1\right)\right), \frac{\partial}{\partial x_2}\left(f_2\left(w_2\right) \bigcirc g_2\left(x_2\right)\right), \ldots, \frac{\partial}{\partial x_n}\left(f_n\left(w_n\right) \bigcirc g_n\left(x_n\right)\right)\right)$$

4.3涉及标量的导数

依据之前的推论：

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \operatorname{diag}\left(\ldots \frac{\partial}{\partial x_i}\left(f_i\left(x_i\right) \bigcirc g_i(z)\right) \ldots\right)$$

假设$\mathbf{f}(\mathbf{x})=\mathbf{x}, \mathbf{g}(z)=\overrightarrow{1} z$

例如， 对于加法： $\mathbf{y}=\mathbf{x}+z$
由：

$$\frac{\partial}{\partial x_i}\left(f_i\left(x_i\right)+g_i(z)\right)=\frac{\partial\left(x_i+z\right)}{\partial x_i}=\frac{\partial x_i}{\partial x_i}+\frac{\partial z}{\partial x_i}=1+0=1$$

可得：

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}+z)=diag(\overrightarrow{1})=\mathbf{I}$$

对于标量z, 可得如下结果

$$\frac{\partial}{\partial z}(f_i(x_i)+g_i(z))=\frac{\partial(x_i+z)}{\partial z}=\frac{\partial x_i}{\partial z}+\frac{\partial z}{\partial z}=0+1=1$$

所以

$$\frac{\partial}{\partial z}(\mathbf{x}+z)=\overrightarrow{1}$$

对于乘法： $\mathbf{y}=\mathbf{x} z$
关于$\mathbf{x}$的导数

$$\frac{\partial}{\partial x_i}\left(f_i\left(x_i\right) \otimes g_i(z)\right)=x_i \frac{\partial z}{\partial x_i}+z \frac{\partial x_i}{\partial x_i}=0+z=z$$

所以:

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x} z)=\operatorname{diag}(\overrightarrow{\mathbf{1}} z)=I z$$

关于z的导数

$$\frac{\partial}{\partial z}\left(f_i\left(x_i\right) \otimes g_i(z)\right)=x_i \frac{\partial z}{\partial z}+z \frac{\partial x_i}{\partial z}=x_i+0=x_i$$

所以：

$$\frac{\partial}{\partial z}(\mathbf{x} z)=\mathbf{x}$$

4.4 向量求和导数

假设 $y = sum(\mathbf{f}(\mathbf{x})) = \sum_{i=1}^{n} f_i(\mathbf{x})$.
对于$\mathbf{x}$的导数如下：

$$\begin{aligned}\frac{\partial y}{\partial \mathbf{x}} &= \left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \ldots, \frac{\partial y}{\partial x_n}\right] \\ &= \left[\frac{\partial}{\partial x_1}\sum_i f_i(\mathbf{x}), \frac{\partial}{\partial x_2}\sum_i f_i(\mathbf{x}), \ldots, \frac{\partial}{\partial x_n}\sum_i f_i(\mathbf{x})\right] \\ &= \left[\sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_1}, \sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_2}, \ldots, \sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_n}\right]\end{aligned}$$

例如，对于$y = sum(\mathbf{x})$, $f_i(\mathbf{x}) = x_i$,其导数如下：

$$\nabla y = \left[\sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_1}, \sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_2}, \ldots, \sum_i \frac{\partial f_i(\mathbf{x})}{\partial x_n}\right] = \left[\sum_i \frac{\partial x_i}{\partial x_1}, \sum_i \frac{\partial x_i}{\partial x_2}, \ldots, \sum_i \frac{\partial x_i}{\partial x_n}\right]$$

又因为，$\frac{\partial}{\partial x_j} x_i = 0$，对于$j \neq i$
所以：

$$\nabla y = \left[\begin{array}{ccc} \frac{\partial x_1}{\partial x_1}, & \frac{\partial x_2}{\partial x_2}, & \ldots, \frac{\partial x_n}{\partial x_n} \end{array}\right] = [1, 1, \ldots, 1] = \overrightarrow{1}^T$$

4.5 链式法则(Chain Rules)

4.5.1 单变量链式法则

对于 $y = f(g(x))$，链式法则如下：

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}, \quad u = g(x)$$

- forward differentiation 参数 如何影响 函数输出

$$\frac{dy}{dx} = \frac{du}{dx}\frac{dy}{du}, \quad u = g(x)$$

- backward differentiation: 函数输出 如何影响 参数

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}, \quad u = g(x)$$

- **单变量链式法则适用场景**
  - 参数x到输出y只有一条数据流路径, 例如对于$y = sin(x^2)$
    xx

4.5.2 单变量全微分链式法则

需要考虑x变化影响输出y的所有路径, 例如对于$y = x + x^2$
xx

对于$y=x+x^2$我们可以视其为

$$
\begin{aligned}
u_1(x) &= x^2 \\
u_2(x,u_1) &= x+u_1 \quad (y=f(x)=u_2(x,u_1))
\end{aligned}
$$

y 对应 x 的全微分为

$$
\frac{dy}{dx}=\frac{\partial f(x)}{\partial x}=\frac{\partial u_2}{\partial x}+\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial x}=1+2x
$$

推广公式如下：

$$
\frac{\partial f(x,u_1,...,u_n)}{\partial x}=\frac{\partial f}{\partial x}+\frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x}+\frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x}+\ldots+\frac{\partial f}{\partial u_n}\frac{\partial u_n}{\partial x}=\frac{\partial f}{\partial x}+\sum_{i=1}^{n}\frac{f}{\partial u_i}\frac{\partial u_i}{\partial x}
$$

令$u_{n+1}=x$

$$
\frac{\partial f(u_1,...,u_{n+1})}{\partial x}=\sum_{i=1}^{n+1}\frac{f}{\partial u_i}\frac{\partial u_i}{\partial x}
$$

这里可以看出类似,向量总和的形式$\frac{\partial f}{\partial \mathbf{u}}\cdot\frac{\partial \mathbf{u}}{\partial x}$

4.5.3 向量链式法则

$\mathbf{y}=\mathbf{f}(x)$:

$$
\begin{bmatrix} y_1(x) \\ y_2(x) \end{bmatrix}=\begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix}=\begin{bmatrix} \ln\left(x^2\right) \\ \sin(3x) \end{bmatrix}
$$

引入中间变量，$\mathbf{y}=\mathbf{f}(\mathbf{g}(x))$:

$$
\begin{bmatrix} g_1(x) \\ g_2(x) \end{bmatrix}=\begin{bmatrix} x^2 \\ 3x \end{bmatrix}
$$

$$
\begin{bmatrix} f_1(\mathbf{g}) \\ f_2(\mathbf{g}) \end{bmatrix}=\begin{bmatrix} \ln(g_1) \\ \sin(g_2) \end{bmatrix}
$$

$\mathbf{y}$对于x的导数如下：

$$
\frac{\partial \mathbf{y}}{\partial x}=\begin{bmatrix} \frac{\partial f_1(\mathbf{g})}{\partial x_1} \\ \frac{\partial f_2(\mathbf{g})}{\partial x} \end{bmatrix}=\begin{bmatrix} \frac{\partial f_1}{\partial g_1}\frac{\partial g_1}{\partial x}+\frac{\partial f_1}{\partial g_2}\frac{\partial g_2}{\partial x_2} \\ \frac{\partial f_2}{\partial g_1}\frac{\partial_1}{\partial x}+\frac{\partial f_2}{\partial g_2}\frac{1}{\partial x} \end{bmatrix}=\begin{bmatrix} \frac{1}{g_1}2x+0 \\ 0+\cos(g_2)\,3 \end{bmatrix}=\begin{bmatrix} \frac{2x}{x^2} \\ 3\cos(3x) \end{bmatrix}=\begin{bmatrix} \frac{2}{x} \\ 3\cos(3x) \end{bmatrix}
$$

可以作如下变换

$$\begin{bmatrix} \frac{\partial f_1}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_1}{\partial g_2}\frac{\partial g_2}{\partial x} \\ \frac{\partial f_2}{\partial g_1}\frac{\partial g_1}{\partial x} + \frac{\partial f_2}{\partial g_2}\frac{\partial g_2}{\partial x} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_1}{\partial g_2} \\ \frac{\partial f_2}{\partial g_1} & \frac{\partial f_2}{\partial g_2} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial x} \\ \frac{\partial g_2}{\partial x} \end{bmatrix} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x}$$

也即是

$$\frac{\partial}{\partial x}\mathbf{f}(\mathbf{g}(x)) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial x}$$

对于参数为向量的情况：

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = \frac{\partial \mathbf{f}}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{x}}$$

展开为矩阵形式如下：

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{f}(\mathbf{g}(\mathbf{x})) = \begin{bmatrix} \frac{\partial f_1}{\partial g_1} & \frac{\partial f_1}{\partial g_2} & \dots & \frac{\partial f_1}{\partial g_k} \\ \frac{\partial f_2}{\partial g_1} & \frac{\partial f_2}{\partial g_2} & \dots & \frac{\partial f_2}{\partial g_k} \\ \frac{\partial f_m}{\partial g_1} & \frac{\partial f_m}{\partial g_2} & \dots & \frac{\partial f_m}{\partial g_k} \end{bmatrix} \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \dots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \dots & \frac{\partial g_2}{\partial x_n} \\ \frac{\partial g_k}{\partial x_1} & \frac{\partial g_k}{\partial x_2} & \dots & \frac{\partial g_k}{\partial x_n} \end{bmatrix}$$

5 神经元激活函数

这里并没有难理解的地方，主要是前面element-wise微分， 向量微分的应用， 还有分段函数微分，**里面提到的广播机制我没觉得有什么用**。
神经元函数表达式

$$activation(\mathbf{x}) = max(0, \mathbf{w} \cdot \mathbf{x} + b)$$

引入中间变量得到如下表达：

$$z(\mathbf{w}, b, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

$$activation(z) = max(0, z)$$

对于函数$max(0, z)$,我们得到如下分段积分

$$\frac{\partial}{\partial z}\max(0, z) = \begin{cases} 0 & z \le 0 \\ \frac{dz}{dz} = 1 & z > 0 \end{cases}$$

根据链式法则

$$\frac{\partial activation}{\partial \mathbf{w}} = \frac{\partial activation}{\partial z}\frac{\partial z}{\partial \mathbf{w}}$$

对于$\frac{\partial activation}{\partial z}$

$$\frac{\partial activation}{\partial z} = \begin{cases} 0 & z \leq 0 \\ \frac{dz}{dz} = 1 & z > 0 \end{cases}$$

对于$\frac{\partial z}{\partial \mathbf{w}}$

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\mathbf{w} \cdot \mathbf{x} + \frac{\partial}{\partial \mathbf{w}}b = \frac{\partial}{\partial \mathbf{w}}\mathbf{w} \cdot \mathbf{x} + \overrightarrow{0}^T$$

对于$\frac{\partial}{\partial \mathbf{w}}\mathbf{w} \cdot \mathbf{x}$，设$y = \mathbf{w} \cdot \mathbf{x}$引入中间变量得到如下结果

$$\begin{aligned} \mathbf{u} &= \mathbf{w} \cdot \mathbf{x} \\ y &= sum(\mathbf{u}) \end{aligned}$$

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}}(\mathbf{w} \cdot \mathbf{x}) = diag(\mathbf{x}) \\ \frac{\partial y}{\partial \mathbf{u}} &= \frac{\partial}{\partial \mathbf{u}}sum(\mathbf{u}) = \overrightarrow{1}^T \end{aligned}$$

所以可得如下结果：

$$\frac{\partial y}{\partial \mathbf{w}} = \frac{\partial y}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{w}} = \overrightarrow{1}^T diag(\mathbf{x}) = \mathbf{x}^T$$

所以

$$\frac{\partial z}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}\mathbf{w} \cdot \mathbf{x} + \frac{\partial}{\partial \mathbf{w}}b = \mathbf{x}^T + \overrightarrow{0}^T = \mathbf{x}^T$$

所以

$$\frac{\partial activaion}{\partial \mathbf{w}} = \begin{cases} 0\frac{\partial z}{\partial \mathbf{w}} & z \leq 0 \\ 1\frac{\partial z}{\partial \mathbf{w}} & z > 0 \end{cases}$$

也即是

$$\frac{\partial activaion}{\partial \mathbf{w}} = \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ \mathbf{x}^T & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

同样的，对于b

$$\frac{\partial activation}{\partial b} = \frac{\partial activation}{\partial z}\frac{\partial z}{\partial b}$$

$$\frac{\partial z}{\partial b} = \frac{\partial}{\partial b}\mathbf{w}\cdot\mathbf{x} + \frac{\partial}{\partial b}b \quad = \quad 0+1 \quad = 1$$

$$\frac{\partial activaion}{\partial b} = \begin{cases} \overrightarrow{0}^T & \mathbf{w}\cdot\mathbf{x}+b \leq 0 \\ \mathbf{x}^T & \mathbf{w}\cdot\mathbf{x}+b > 0 \end{cases}$$

所以

$$\frac{\partial activaion}{\partial b} = \begin{cases} 0 & \mathbf{w}\cdot\mathbf{x}+b \leq 0 \\ 1 & \mathbf{w}\cdot\mathbf{x}+b > 0 \end{cases}$$

6 神经网络损失函数的梯度

假设神经网络输入如下

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$$

标签如下

$$\mathbf{y} = [\operatorname{target}(\mathbf{x}_1), \operatorname{target}(\mathbf{x}_2), \ldots, \operatorname{target}(\mathbf{x}_N)]^T = [y_1, y_2, \cdots, y_N]$$

损失函数

$$C(\mathbf{w}, b, X, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \operatorname{activation}(\mathbf{x}_i))^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \max(0, \mathbf{w}\cdot\mathbf{x}_i + b))^2$$

定义中间变量如下

$$\begin{aligned} u(\mathbf{w}, b, \mathbf{x}) &\quad = \quad \max(0, \mathbf{w}\cdot\mathbf{x}+b) \\ v(y, u) &\quad = \quad y - u \\ C(v) &\quad = \quad \tfrac{1}{N}\textstyle\sum_{i=1}^{N} v^2 \end{aligned}$$

6.1 对于参数 $\mathbf{w}$ 的梯度

因为

$$\frac{\partial}{\partial \mathbf{w}} u(\mathbf{w}, b, \mathbf{x}) = \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ \mathbf{x}^T & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

所以

$$\frac{\partial v(y,u)}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}}(y-u) = \overrightarrow{0}^T - \frac{\partial u}{\partial \mathbf{w}} = -\frac{\partial u}{\partial \mathbf{w}} = \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x} + b \leq 0 \\ -\mathbf{x}^T & \mathbf{w} \cdot \mathbf{x} + b > 0 \end{cases}$$

所以

$$\begin{aligned}
\frac{\partial C(v)}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} v^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \mathbf{w}} v^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{\partial v^2}{\partial v} \frac{\partial v}{\partial \mathbf{w}} \\
&= \frac{1}{N} \sum_{i=1}^{N} 2v \frac{\partial v}{\partial \mathbf{w}} \\
&= \frac{1}{N} \sum_{i=1}^{N} \begin{cases} 2v \overrightarrow{0}^T = \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ -2v\mathbf{x}^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\
&= \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ -2\left(y_i - u\right) \mathbf{x}_i^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\
&= \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ -2\left(y_i - \max\left(0, \mathbf{w} \cdot \mathbf{x}_i + b\right)\right) \mathbf{x}_i^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\
&= \frac{1}{N} \sum_{i=1}^{N} \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ -2\left(y_i - \left(\mathbf{w} \cdot \mathbf{x}_i + b\right)\right) \mathbf{x}_i^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\
&= \begin{cases} \overrightarrow{0}^T & \mathbf{w} \cdot \mathbf{x}_i + b \leq 0 \\ \frac{-2}{N} \sum_{i=1}^{N} \left(y_i - \left(\mathbf{w} \cdot \mathbf{x}_i + b\right)\right) \mathbf{x}_i^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases} \\
&= \begin{cases} \overrightarrow{0}^T & \mathrm{w} \ \cdot \mathbf{x}_i + b \leq 0 \\ \frac{2}{N} \sum_{i=1}^{N} \left(\mathbf{w} \cdot \mathbf{x}_i + b - y_i\right) \mathbf{x}_i^T & \mathbf{w} \cdot \mathbf{x}_i + b > 0 \end{cases}
\end{aligned}$$

令 $e_i = \mathbf{w} \cdot \mathbf{x}_i + b - y_i$
则

$$\frac{\partial C}{\partial \mathbf{w}} = \frac{2}{N} \sum_{i=1}^{N} e_i \mathbf{x}_i^T$$

假定输入向量只有一个，损失值为$2e_1\mathbf{x}_1^T$.如果错误e_1为0,那么损失值为0; 如果e_1为正数，那么梯度方向在$\mathbf{x}_1$方向，如果e_1为负值，那么梯度方向为$\mathbf{x}_1$的负方向
对于梯度下降算法，我们需要向梯度负方向移动：

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial C}{\partial \mathbf{w}}$$

6.2 对于偏置b的微分

$$\begin{aligned} u(\mathbf{w}, b, \mathbf{x}) &= \max(0, \mathbf{w}\cdot\mathbf{x} + b) \\ v(y, u) &= y - u \\ C(v) &= \frac{1}{N}\sum_{i=1}^{N} v^2 \end{aligned}$$

对于函数u：

$$\frac{\partial u}{\partial b} = \begin{cases} 0 & \mathbf{w}\cdot\mathbf{x} + b \leq 0 \\ 1 & \mathbf{w}\cdot\mathbf{x} + b > 0 \end{cases}$$

对于函数v:

$$\frac{\partial v(y,u)}{\partial b} = \frac{\partial}{\partial b}(y-u) = 0 - \frac{\partial u}{\partial b} = -\frac{\partial u}{\partial b} = \begin{cases} 0 & \mathbf{w}\cdot\mathbf{x} + b \leq 0 \\ -1 & \mathbf{w}\cdot\mathbf{x} + b > 0 \end{cases}$$

对于损失函数：

$$
\begin{aligned}
\frac{\partial C(v)}{\partial b} &= \frac{\partial}{\partial b}\frac{1}{N}\sum_{i=1}^{N} v^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial b}v^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial v^2}{\partial v}\frac{\partial v}{\partial b} \\
&= \frac{1}{N}\sum_{i=1}^{N}2v\frac{\partial v}{\partial b} \\
&= \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 0 & \mathbf{w}\cdot\mathbf{x}+b\leq 0 \\ -2v & \mathbf{w}\cdot\mathbf{x}+b>0 \end{cases} \\
&= \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 0 & \mathbf{w}\cdot\mathbf{x}+b\leq 0 \\ -2\left(y_i-\max\left(0,\mathbf{w}\cdot\mathbf{x}_i+b\right)\right) & \mathbf{w}\cdot\mathbf{x}+b>0 \end{cases} \\
&= \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 0 & \mathbf{w}\cdot\mathbf{x}+b\leq 0 \\ 2\left(\mathbf{w}\cdot\mathbf{x}_i+b-y_i\right) & \mathbf{w}\cdot\mathbf{x}+b>0 \end{cases} \\
&= \begin{cases} 0 & \mathbf{w}\cdot\mathbf{x}_i+b\leq 0 \\ \frac{2}{N}\sum_{i=1}^{N}\left(\mathbf{w}\cdot\mathbf{x}_i+b-y_i\right) & \mathbf{w}\cdot\mathbf{x}_i+b>0 \end{cases}
\end{aligned}
$$

与之前类似

$$
\frac{\partial C}{\partial b} = \frac{2}{N}\sum_{i=1}^{N} e_i
$$

参数优化方式如下

$$
b_{t+1} = b_t - \eta\frac{\partial C}{\partial b}
$$

矩阵求导公式参考

wiki百科