Week 1 ML 学习总结 机器学习的环境与数学基础

机器学习环境搭建

笔者采用的环境/框架是 `Python` `CUDA` `Torch`

具体步骤

1. 安装Anaconda，部署Python, Jupyter Notebook等的环境。
2. **在Conda的Console中使用`conda install`命令安装`pytorch`的GPU版本。**

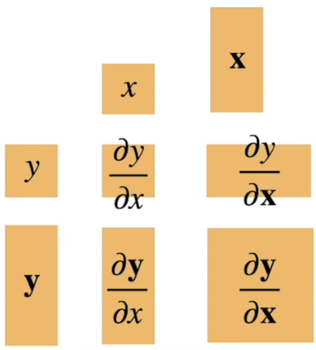
   ```
   # for windows
   conda install pytorch torchvision torchaudio cudatoolkit=11.3 -c pytorch
   ```

   完成。

定义

梯度

梯度是对导数的扩充。



标量函数对向量的梯度

$$\boldsymbol{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \frac{\partial y}{\partial \boldsymbol{x}}=\left[\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n}\right] \tag{1}$$

梯度指向了值变化最大的方向。

实例

$$\frac{\partial(x_1^2+2x_2^2)}{\partial \boldsymbol{x}}=[2x_1, 4x_2] \tag{2}$$

标量函数对向量的梯度

例子

$$
\begin{array}{c|cccc}
y & a & au & sum(\boldsymbol{x}) & \|\boldsymbol{x}\|^2 \\
\hline
\frac{\partial y}{\partial \boldsymbol{x}} & \mathbf{0}^T & a\frac{\partial u}{\partial \boldsymbol{x}} & \mathbf{1}^T & 2\boldsymbol{x}^T
\end{array} \tag{3}
$$

$$
\begin{array}{c|ccc}
y & u+v & uv & \langle \boldsymbol{u}, \boldsymbol{v} \rangle \\
\hline
\frac{\partial y}{\partial \boldsymbol{x}} & \frac{\partial u}{\partial \boldsymbol{x}} + \frac{\partial v}{\partial \boldsymbol{x}} & \frac{\partial u}{\partial \boldsymbol{x}} v + \frac{\partial v}{\partial \boldsymbol{x}} u & \boldsymbol{u}^T \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}} + \boldsymbol{v}^T \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}
\end{array} \tag{4}
$$

一些证明

$$
\begin{aligned}
\frac{\partial \| \boldsymbol{x} \|^2}{\partial \boldsymbol{x}} &= \frac{\partial x_1^2 + x_2^2 + \cdots + x_n^2}{\partial \boldsymbol{x}} \\
&= \begin{bmatrix} 2x_1 & 2x_2 & \cdots & 2x_n \end{bmatrix} \\
&= 2\boldsymbol{x}^T
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
&\frac{\partial \langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\partial \boldsymbol{x}} \\
&= \frac{\partial \boldsymbol{u}_1 \boldsymbol{v}_1 + \boldsymbol{u}_2 \boldsymbol{v}_2, \cdots + \boldsymbol{u}_n \boldsymbol{v}_n}{\partial \boldsymbol{x}} \\
&= \boldsymbol{u}_1 \frac{\partial \boldsymbol{v}_1}{\partial \boldsymbol{x}} + \boldsymbol{v}_1 \frac{\partial \boldsymbol{u}_1}{\partial \boldsymbol{x}} + \boldsymbol{u}_2 \frac{\partial \boldsymbol{v}_2}{\partial \boldsymbol{x}} + \boldsymbol{v}_2 \frac{\partial \boldsymbol{u}_2}{\partial \boldsymbol{x}} + \cdots + \boldsymbol{u}_n \frac{\partial \boldsymbol{v}_n}{\partial \boldsymbol{x}} + \boldsymbol{v}_1 \frac{\partial \boldsymbol{u}_n}{\partial \boldsymbol{x}} \\
&= \boldsymbol{u}^T \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}} + \boldsymbol{v}^T \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}
\end{aligned} \tag{6}
$$

向量函数对标量的梯度

$$
\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad \frac{\partial \boldsymbol{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix} \tag{7}
$$

向量函数对向量的梯度

$$
\begin{gathered}
\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \\
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} \\ \frac{\partial y_2}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial y_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}
\end{gathered} \tag{8}
$$

例子

$$
\begin{array}{c|cccc}
y & \boldsymbol{a} & \boldsymbol{x} & \boldsymbol{A}\boldsymbol{x} & \boldsymbol{x}^T\boldsymbol{A} \\
\hline
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} & \mathbf{0} & \boldsymbol{I} & \boldsymbol{A} & \boldsymbol{A}^T
\end{array} \tag{9}
$$

$$
\begin{array}{c|ccc}
y & a\boldsymbol{u} & \boldsymbol{A}\boldsymbol{u} & \boldsymbol{u}+\boldsymbol{v} \\
\hline
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} & a\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} & A\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} & \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}
\end{array} \tag{10}
$$

一些证明

$$
\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}}=\begin{bmatrix}\frac{\partial x_1}{\partial \boldsymbol{x}}\\ \frac{\partial x_2}{\partial \boldsymbol{x}}\\ \vdots\\ \frac{\partial x_m}{\partial \boldsymbol{x}}\end{bmatrix}=\begin{bmatrix}\frac{\partial x_1}{\partial x_1} & \frac{\partial x_1}{\partial x_2} & \cdots & \frac{\partial x_1}{\partial x_n}\\ \frac{\partial x_2}{\partial x_1} & \frac{\partial x_2}{\partial x_2} & \cdots & \frac{\partial x_2}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial x_n}{\partial x_1} & \frac{\partial x_n}{\partial x_2} & \cdots & \frac{\partial x_n}{\partial x_n}\end{bmatrix}=\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}=I_{n\times n} \tag{11}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \frac{\partial\begin{bmatrix}a_{1,1}x_1+\cdots+a_{1,n}x_n\\ a_{2,1}x_1+\cdots+a_{2,n}x_n\\ \vdots\\ a_{m,1}x_1+\cdots+a_{m,n}x_n\end{bmatrix}}{\partial \boldsymbol{x}}\\
&=\begin{bmatrix}\frac{\partial a_{1,1}x_1+\cdots+a_{1,n}x_n}{\partial x_1} & \frac{\partial a_{1,1}x_1+\cdots+a_{1,n}x_n}{\partial x_2} & \cdots & \frac{\partial a_{1,1}x_1+\cdots+a_{1,n}x_n}{\partial x_n}\\ \frac{\partial a_{2,1}x_1+\cdots+a_{2,n}x_n}{\partial x_1} & \frac{\partial a_{2,1}x_1+\cdots+a_{2,n}x_n}{\partial x_2} & \cdots & \frac{\partial a_{2,1}x_1+\cdots+a_{2,n}x_n}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial a_{m,1}x_1+\cdots+a_{m,n}x_n}{\partial x_1} & \frac{\partial a_{m,1}x_1+\cdots+a_{m,n}x_n}{\partial x_2} & \cdots & \frac{\partial a_{m,1}x_1+\cdots+a_{m,n}x_n}{\partial x_n}\end{bmatrix}\\
&=\begin{bmatrix}a_{1,1} & a_{1,2} & \cdots & a_{1,n}\\ a_{2,1} & a_{2,2} & \cdots & a_{1,n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m,1} & a_{m,2} & \cdots & a_{m,n}\end{bmatrix}\\
&=A
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{x^T}\boldsymbol{A}}{\partial \boldsymbol{x}} &= \frac{\partial\begin{bmatrix}a_{1,1}x_1+a_{2,1}x_2+\cdots+a_{n,1}x_n\\ a_{1,2}x_1+a_{2,2}x_2+\cdots+a_{n,2}x_n\\ \vdots\\ a_{1,m}x_1+a_{2,m}x_2+\cdots+a_{n,m}x_n\end{bmatrix}^T}{\partial \boldsymbol{x}}\\
&=\begin{bmatrix}\frac{a_{1,1}x_1+a_{2,1}x_2+\cdots+a_{n,1}x_n}{\partial x_1} & \frac{a_{1,1}x_1+a_{2,1}x_2+\cdots+a_{n,1}x_n}{\partial x_2} & \cdots & \frac{a_{1,1}x_1+a_{2,1}x_2+\cdots+a_{n,1}x_n}{\partial x_n}\\ \frac{a_{1,2}x_1+a_{2,2}x_2+\cdots+a_{n,2}x_n}{\partial x_1} & \frac{a_{1,2}x_1+a_{2,2}x_2+\cdots+a_{n,2}x_n}{\partial x_2} & \cdots & \frac{a_{1,2}x_1+a_{2,2}x_2+\cdots+a_{n,2}x_n}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{a_{1,m}x_1+a_{2,m}x_2+\cdots+a_{n,m}x_n}{\partial x_1} & \frac{a_{1,m}x_1+a_{2,m}x_2+\cdots+a_{n,m}x_n}{\partial x_2} & \cdots & \frac{a_{1,m}x_1+a_{2,m}x_2+\cdots+a_{n,m}x_n}{\partial x_n}\end{bmatrix}\\
&=\begin{bmatrix}a_{1,1} & a_{2,1} & \cdots & a_{n,1}\\ a_{1,2} & a_{2,2} & \cdots & a_{n,2}\\ \vdots & \vdots & \ddots & \vdots\\ a_{1,m} & a_{2,m} & \cdots & a_{n,m}\end{bmatrix}\\
&=A^T
\end{aligned} \tag{13}
$$

标量函数对矩阵的梯度

m维行向量函数$\boldsymbol{f}(\boldsymbol{x})=[f_1(\boldsymbol{x}),f_2(\boldsymbol{x}),\cdots,f_m(\boldsymbol{x})]$相对于$n$维实向量**x**的梯度为一$n\times m$矩阵，定义为

$$
\begin{bmatrix}\frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_1}\\ \frac{\partial f_1(\boldsymbol{x})}{} & \frac{\partial f_2(\boldsymbol{x})}{} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{}\end{bmatrix}
$$

$$
\nabla_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) \stackrel{\text{def}}{=} \begin{bmatrix} \overline{\partial x_2} & \overline{\partial x_2} & \cdots & \overline{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_n} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} \tag{14}
$$

（关于为什么列向量变为了行向量：行向量和列向量乘积是标量）

求导

链式法则

标量链式法则

$$
\frac{\partial y}{\partial x} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial x} \tag{15}
$$

向量链式法则

$$
\begin{gathered}
\frac{\partial y}{\partial \boldsymbol{x}} = \frac{\partial y}{\partial u} \frac{\partial u}{\partial \boldsymbol{x}} \\
(1, n) \quad (1)(1, n) \\
\frac{\partial y}{\partial \boldsymbol{x}} = \frac{\partial y}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \\
(1, n) \quad (1, k)(k, n) \\
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \\
(m, n) \quad (m, k)(k, n)
\end{gathered} \tag{16}
$$

例子

$$
\begin{gathered}
z = (\langle \boldsymbol{x}, \boldsymbol{w} \rangle - y)^2 \\
\\
a = \langle \boldsymbol{x}, \boldsymbol{w} \rangle \\
b = a - y \\
z = b^2 \\
\\
\begin{aligned}
\frac{\partial z}{\partial w} &= \frac{\partial z}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial w} \\
&= 2b \cdot 1 \cdot \boldsymbol{x}^T \quad (u, v\text{无关}) \\
&= (2\langle \boldsymbol{x}, \boldsymbol{w} \rangle - y) x^T
\end{aligned}
\end{gathered} \tag{17}
$$

$$
\begin{gathered}
z = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 \\
\\
a = \boldsymbol{X}\boldsymbol{w} \\
b = a - \boldsymbol{y} \\
z = \|b\|^2 \\
\\
\begin{aligned}
\frac{\partial z}{\partial w} &= \frac{\partial z}{\partial b} \frac{\partial b}{\partial a} \frac{\partial a}{\partial w} \\
&= 2b^T \cdot 1 \cdot \boldsymbol{X} \\
&= 2(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T \boldsymbol{X}
\end{aligned}
\end{gathered} \tag{18}
$$

动手做!

[自动求导实验]

线性回归模型：以预测房价为例

概述

$$\begin{aligned}\text{Input: } X &= [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^T \\ \text{Output: } y &= [y_1, y_2, \cdots, y_n]^T\end{aligned} \tag{19}$$

输入关于房产信息的向量x_i，输出房价y_i

假设对于房价的影响由三个因素确定：x_1, x_2, x_3

假设成交价是关键因素的加权和$y = w_1x_1 + w_2x_2 + w_3x_3 + b$

推广

广泛的，可以如此表示线性模型：

$$y = \sum_{i=1}^{n} w_i x_i + b \tag{20}$$

也可以以向量形式表示为：

$$y = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \tag{21}$$

衡量与评估质量

损失函数

$$\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{22}$$

(平方损失)

定义

训练损失

$$\ell(x, y, w, b) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle - b)^2 = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - b\|^2 \tag{23}$$

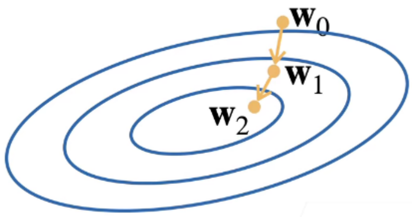
最小化损失来学习参数

$$\boldsymbol{w}^*, \boldsymbol{b}^* = \arg\min_{\boldsymbol{w}, \boldsymbol{b}} \ell(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{w}, b) \tag{24}$$

梯度下降

1. 选取初始值$\boldsymbol{w}_0$
2. 迭代$t = 1, 2, 3 \cdots$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta\frac{\partial \ell}{\partial \boldsymbol{w}_{t-1}} \tag{25}$$



每次向着梯度的反方向前进，会最大的减少损失函数值。

η :学习率 步长的**超参数**

*超参数：在开始学习过程之前设置值的参数

学习率不应该过小，否则梯度下降过慢；学习率过大可能导致震荡

更经济的版本：小批量随机梯度下降

原因：在整个训练集上计算开销过大。

随机采样b个样本来近似损失。

b :超参数，批量大小

$$\frac{1}{b}\sum_{i\in I_b}\ell(\boldsymbol{x}_i, y_i, \boldsymbol{w}) \tag{26}$$

批量不能过小，否则不能最大利用并行资源；

不能过大，增大开销，浪费计算。

动手做！

[linear]