机器学习 - 周志华 - 线性模型

分子布局

如果我们设 $\boldsymbol{y}=(y_1,y_2,\cdots,y_m)^{\mathrm{T}}$ ，那么对标量 x 求导的符号如下：

$$\frac{\partial \boldsymbol{y}}{\partial x}=\left(\frac{\partial y_1}{\partial x},\cdots,\frac{\partial y_m}{\partial x}\right)^{\mathrm{T}}$$

这种符号也称为分子布局符号. 也就是对每个位置进行标量求导. 相反的即称为分母布局符号.

反过来就是相当于求梯度：对 $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}$ ，

$$\frac{\partial y}{\partial \boldsymbol{x}}=\left(\frac{\partial y}{\partial x_1},\cdots,\frac{\partial y}{\partial x_n}\right)$$

注意这个求导得到的是列向量，一些经典的标量求导：

> **例：范数求导**
>
> 对 $\|\boldsymbol{x}\|^2$ 求导.

使用欧氏范数有

$$\|\boldsymbol{x}\|^2=\sum_{k=1}^{n}x_k^2$$

求导有

$$\frac{\partial\|\boldsymbol{x}\|^2}{\partial \boldsymbol{x}}=(2x_k)=2\boldsymbol{x}^{\mathrm{T}}$$

$\square$

> **例：内积求导**
>
> 对 $\langle\boldsymbol{u},\boldsymbol{v}\rangle$ 求导，$\boldsymbol{u},\boldsymbol{v}$ 均为 $\boldsymbol{x}$ 的函数.

首先 $\langle\boldsymbol{u},\boldsymbol{v}\rangle=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}=\sum_{k=1}^{n}u_kv_k$. 那么我们求导有

$$
\begin{aligned}
\frac{\partial \boldsymbol{u}^{\mathrm{T}} \boldsymbol{v}}{\partial \boldsymbol{x}} &= \left(\sum_{k=1}^{n}\left[v_k \frac{\partial u_k}{\partial x_i} + u_k \frac{\partial v_k}{\partial x_i}\right]\right)_i \\
&= \left(\sum_{k=1}^{n} v_k \frac{\partial u_k}{\partial x_i}\right)_i + \left(\sum_{k=1}^{n} u_k \frac{\partial v_k}{\partial x_i}\right)_i \\
&= \boldsymbol{v}^{\mathrm{T}} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} + \boldsymbol{u}^{\mathrm{T}} \frac{\partial \boldsymbol{v}}{\partial \boldsymbol{x}}
\end{aligned}
$$

□

向量对向量求导

如果我们记 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$ ，$\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^{\mathrm{T}}$ ，那么此时有向量对向量求导：

$$
\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix}
\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\
\vdots & \vdots & & \vdots \\
\frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n}
\end{pmatrix}
$$

相当于就是说：向量对向量求导将会拉伸一个维度到矩阵的维度，我们看接下来的例子就可以大概明白其原理.

下面的 $\boldsymbol{a}$ 和 $\boldsymbol{A}$ 不是 $\boldsymbol{x}$ 的函数，均为常量向量、矩阵. 于是有

- $\frac{\partial}{\partial \boldsymbol{x}}(a\boldsymbol{u}) = a\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$.
- $\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{I}$.
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$.
- $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}) = \boldsymbol{A}^{\mathrm{T}}$.

我们可以拓展上述的内容，矩阵对向量求导将会得到一个三维的**张量**，以此类推有更高维的内容.

链式法则和自动求导

向量、矩阵求导也符合链式法则，我们也和正常求导一样拆开即可.

例：矩阵求导链式法则

设 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ ， $\boldsymbol{w} \in \mathbb{R}^{n}$ ， $\boldsymbol{y} \in \mathbb{R}^{m}$ ，对如下函数求导：

$$z = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2$$

我们依次分解：

$$\boldsymbol{a} = \boldsymbol{X}\boldsymbol{w}, \boldsymbol{b} = \boldsymbol{a} - \boldsymbol{y}, z = \|\boldsymbol{b}\|^2$$

求导有：

$$\begin{aligned}
\frac{\partial z}{\partial \boldsymbol{w}} &= \frac{\partial z}{\partial \boldsymbol{b}}\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{a}}\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}} \\
&= \frac{\partial \|\boldsymbol{b}\|^2}{\partial \boldsymbol{b}}\frac{\partial (\boldsymbol{a}-\boldsymbol{y})}{\partial \boldsymbol{a}}\frac{\partial \boldsymbol{X}\boldsymbol{w}}{\partial \boldsymbol{w}} \\
&= 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{I}\boldsymbol{X} = 2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X} \\
&= 2(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^{\mathrm{T}}\boldsymbol{X}
\end{aligned}$$

□

回归任务

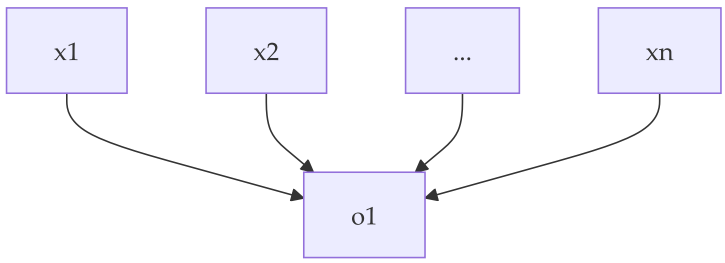
回归任务是通过输入去预测连续数值的任务，线性回归就是通过如下的 n 维输入：

$$\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$$

计算出 n 维权重 $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)^{\mathrm{T}}$ 和标量偏差 b ，从而可得输出为：

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

线性模型可以看作是一个单层的神经网络：



上面的第一层为输入层，第二层为输出层.

衡量回归质量

考虑 MSE ：

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

线性回归任务可以计算其解析解，我们先将训练损失计算为：

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

那么我们实际上就是要求出如下参数：

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

这实际上是一个优化类问题，其解法也比较简单，首先我们将偏差加入到权重当中，我们将偏差对应的项视为新增一个变量 1 ，权重为 b ，也就是说

$$\boldsymbol{X}\to(\boldsymbol{X},1);\boldsymbol{w}\to(\boldsymbol{w},b)$$

因此有

$$\ell(\boldsymbol{X},\boldsymbol{y},w)=\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|^2$$

求导后有

$$\frac{\partial}{\partial\boldsymbol{w}}\ell(\boldsymbol{X},\boldsymbol{y},\boldsymbol{w})=\frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}\boldsymbol{X}$$

损失函数是凸函数，上述导数取 0 即有最值，可得

$$\boldsymbol{w}^*=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$$

这是唯一解. 注意这里我们默认了 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 是可逆的，

线性模型的变化

> **广义线性模型**
>
> 若函数 $g(x)$ 为单调可微函数，则
>
> $$y=g^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)$$
>
> 则称该线性模型为**广义线性模型**.

线性模型的分类任务：对数几率回归

我们讨论的是线性回归模型，但是如果任务是分类的怎么办？考虑二分类任务，其输出标记 $y \in \{0,1\}$,而线性回归模型产生的预测值 $z = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 是实值，于是，我们需将实值转换为 $0/1$ 值.最理想的是“单位阶跃函数“(unit-step function)：

$$y = \begin{cases} 0, & x < 0 \\ 0.5, & x = 0 \\ 1, & x > 0 \end{cases}$$

但是这种函数不可微，因此我们需要选取一种类似的可微函数.