最小二乘法解的矩阵形式

简介

最近在看 NNDL，其中有一个经验风险最小化的例子，即最小二乘法，定义如下：

给定一组包含 N 个训练样本的训练机 $D=\{(\mathbf{x}^{(n)},y^{(n)})\}_{n=1}^{N}$ 。使用线性回归。样本和参数均为列向量。

$$f(\mathbf{x};\mathbf{w})=\mathbf{w}^T\mathbf{x}$$

平方损失函数

经验风险最小化，训练集的风险被定义为，$X=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N]^T$：

$$\begin{aligned}
R(\mathbf{w}) &= \sum_{n=1}^{N}\frac{1}{2}\left(y^n-\mathbf{w}^T\mathbf{x}^{(n)}\right)\\
&=\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2\\
&=\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\\
&=\frac{1}{2}(\mathbf{y}^T-\mathbf{w}^T\mathbf{X}^T)(\mathbf{y}-\mathbf{X}\mathbf{w})\\
&=\frac{1}{2}\left(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}\mathbf{w}-\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}\right)
\end{aligned}$$

损失函数最终是一个标量，可以发现

$$\mathbf{y}^T\mathbf{X}\mathbf{w}=(\mathbf{w}^T\mathbf{X}^T\mathbf{y})^T=scalar$$

两个是一个数字，因此

$$R(\mathbf{w})=\frac{1}{2}(\mathbf{y}^T\mathbf{y}-2(\mathbf{y}^T\mathbf{X}\mathbf{w})+\|\mathbf{X}\mathbf{w}\|^2)$$

对参数求导

首先损失函数是一个凸函数，梯度为 0 的点是全局的最小值。需要对 $\mathbf{w}$ 求导。

$$
\begin{aligned}
\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} &= \frac{1}{2}\frac{\partial(\mathbf{y}^T\mathbf{y} - 2(\mathbf{y}^T\mathbf{X}\mathbf{w}) + \|\mathbf{X}\mathbf{w}\|^2)}{\partial \mathbf{w}} \\
&= \frac{1}{2}(0 - \frac{\partial(2\mathbf{y}^T\mathbf{X}\mathbf{w})}{\partial \mathbf{w}} + \frac{\partial\|\mathbf{X}\mathbf{w}\|^2}{\partial \mathbf{w}}) \\
&= -\frac{\partial \mathbf{y}^T\mathbf{X}\mathbf{w}}{\partial \mathbf{w}} + \frac{1}{2}\frac{\|\mathbf{X}\mathbf{w}\|^2}{\partial \mathbf{w}}
\end{aligned}
$$

分析前半部分，矩阵展开计算依次求导。

$$
\nabla_{\mathbf{w}} f(\mathbf{w}) = \begin{bmatrix} \partial f(\mathbf{w})/\partial w_1 \\ \partial f(\mathbf{w})/\partial w_2 \\ \vdots \\ \partial f(\mathbf{w})/\partial w_N \end{bmatrix} = \mathbf{X}^T\mathbf{y}
$$

对于后半部分

$$
\begin{aligned}
\nabla_{\mathbf{w}} \mathbf{w}^T X^T X \mathbf{w} &= \nabla_{\mathbf{w}} \mathbf{w}^T A \mathbf{w} \\
&= \begin{bmatrix} \partial \mathbf{w}^T \mathbf{A}\mathbf{w}/\partial_{w_1} \\ \partial \mathbf{w}^T \mathbf{A}\mathbf{w}/\partial_{w_2} \\ \vdots \\ \partial \mathbf{w}^T \mathbf{A}\mathbf{w}/\partial_{w_N} \end{bmatrix} \\
&= \begin{bmatrix} 2w_1(A_{11} + A_{12} + A_{13} + \cdots A_{1N}) \\ 2w_2(A_{21} + A_{22} + A_{23} + \cdots A_{2N}) \\ \vdots \\ 2w_N(A_{N1} + A_{N2} + A_{N3} + \cdots A_{NN}) \end{bmatrix} \\
&= 2A\mathbf{w} \\
&= 2X^T X\mathbf{w}
\end{aligned}
$$

所以有

$$
\frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} = -X^T\mathbf{y} + X^T X\mathbf{w}
$$

求解最优参数

让导数为 0， 可得

$$
\begin{aligned}
X^T X\mathbf{w} &= X^T\mathbf{y} \\
\mathbf{w} &= (X^T X)^{-1} X^T\mathbf{y}
\end{aligned}
$$