线性回归作业解答

1，假设训练样本集为 $D=\{(\mathbf{x}_1,y_1)=((2,2)^T,1),(\mathbf{x}_2,y_2)=((4,1)^T,1),(\mathbf{x}_3,y_3)=((1,0)^T,-1)\},$，使用线性回归算法（Linear Regression Algorithm），通过广义逆来求解，并设计这两类的分类函数，判断测试样本$\mathbf{x}=(0,1)^T$属于哪个类别，与 L9 习题 1 设计的分类器相比较，讨论结果。

解：令$D=\{(\mathbf{x}_i,y_i)=\big((1,x_{1i},x_{2i}),y_i\big)\},i=1,2,3$，故可写出：

$$\mathrm{X}=\begin{pmatrix}1&2&2\\1&4&1\\1&1&0\end{pmatrix},\ \mathrm{X}^T=\begin{pmatrix}1&1&1\\2&4&1\\2&1&0\end{pmatrix},\mathrm{Y}=\begin{pmatrix}1\\1\\-1\end{pmatrix}$$

$$\mathrm{X}^T\mathrm{X}=\begin{pmatrix}1&1&1\\2&4&1\\2&1&0\end{pmatrix}\begin{pmatrix}1&2&2\\1&4&1\\1&1&0\end{pmatrix}=\begin{pmatrix}3&7&3\\7&21&8\\3&8&5\end{pmatrix}$$

$$(\mathrm{X}^T\mathrm{X})^{-1}=\begin{pmatrix}1.64&-0.44&-0.28\\-0.44&0.24&-0.12\\-0.28&-0.12&0.56\end{pmatrix}$$

计算广义逆：

$$\mathrm{X}^\dagger=(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T=\begin{pmatrix}0.20&-0.40&1.20\\-0.20&0.40&-0.20\\0.6&-0.20&-0.40\end{pmatrix}$$

因此：

$$\boldsymbol{w}^*=\mathrm{X}^\dagger Y=\begin{pmatrix}0.20&-0.40&1.20\\-0.20&0.40&-0.20\\0.6&-0.20&-0.40\end{pmatrix}\begin{pmatrix}1\\1\\-1\end{pmatrix}=\begin{pmatrix}-1.4\\0.4\\0.8\end{pmatrix}$$

对于样本$\mathbf{x}=(0,1)^T$，所属类别为：

$$sign(\boldsymbol{w}^{*T}\mathbf{x})=sign((-1.4,0.4,0.8)(1,0,1)^T=-1,\quad\therefore\ \ \mathbf{x}\in-1\ 类$$

讨论：同样的训练样本得到不同的分类面，导致测试结果不同。

2，根据向量或矩阵的计算性质，证明：

$$\|\mathbf{X}\boldsymbol{w} - Y\|^2 = \boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} + \mathrm{Y}^T\mathrm{Y}$$

解：

$$\begin{aligned}\|\mathbf{X}\boldsymbol{w} - Y\|^2 &= (\mathbf{X}\boldsymbol{w} - Y)^T(\mathbf{X}\boldsymbol{w} - Y)\\ &= ((\mathbf{X}\boldsymbol{w})^T - \mathrm{Y}^T)(\mathbf{X}\boldsymbol{w} - Y)\\ &= (\boldsymbol{w}^T\mathbf{X}^T - \mathrm{Y}^T)(\mathbf{X}\boldsymbol{w} - Y)\\ &= \boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} - \boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} - \mathrm{Y}^T\mathbf{X}\boldsymbol{w} + \mathrm{Y}^T\mathrm{Y}\\ &= \boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} - \boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} - (\mathbf{X}\boldsymbol{w})^T\mathrm{Y} + \mathrm{Y}^T\mathrm{Y}\\ &= \boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} - \boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} - \boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} + \mathrm{Y}^T\mathrm{Y}\\ &= \boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} - 2\boldsymbol{w}^T\mathbf{X}^T\mathrm{Y} + \mathrm{Y}^T\mathrm{Y}\end{aligned}$$

3，假设在线性回归问题中，$p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\mathbf{x},\sigma^2)$，其中方差$\sigma^2$是一个确定值，请写出参数$\boldsymbol{w}$的似然函数$\mathcal{L}(\boldsymbol{w})$，并证明在这个情况下线性回归的最大化似然等价于最小二乘法，并通过推导似然函数$\mathcal{L}(\boldsymbol{w})$的梯度，求出参数$\boldsymbol{w}$的最大似然估计$\boldsymbol{w}^*$

解：

因为：$p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{w}^T\mathbf{x},\sigma^2)$

说明：

给定输入**x和模型中的各项参数**，y 是正态分布，其均值是对**x**的线性预测，假设训练样本数为 N，则：

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{w}) = \mathcal{L}(D|\boldsymbol{w}) = \sum_{n=1}^{N} log\, p(y_n|\mathbf{x}_{\boldsymbol{n}}, \boldsymbol{w})$$

$$= \sum_{n=1}^{N} log\, (\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\mathbf{x}_n)^2)$$

$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\mathbf{x}_n)^2 - \frac{N}{2} log\, (2\pi\sigma^2)$$

$$= -\frac{1}{2\sigma^2}\|\mathbf{X}\boldsymbol{w} - Y\|^2 - \frac{N}{2} log\, (2\pi\sigma^2)$$

上式的第二项为常数，所以最大化似然就等价于线性回归问题中的最小二乘法:

$$\min_{\boldsymbol{w}} \ell_{in}(\boldsymbol{w}) = \frac{1}{2\sigma^2}\|\mathbf{X}\boldsymbol{w} - Y\|^2$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w})}{\partial\boldsymbol{w}} = \frac{1}{2\sigma^2}2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - Y) = \mathbf{0}$$

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

可见，利用最大似然求得的最佳$\boldsymbol{w}^*$与最小二乘法得到的结果一致。

4，总结梯度下降法、随机梯度下降法、Adagrad、RMSProp、动量法（Momentum）和 Adam 等方法权系数更新表达式。

解：对于任意的损失函数$\ell$，假设任一单个样本 $n$ 的梯度$\nabla\ell_n(\boldsymbol{w})$，$t$ 代表迭代次数

（1）梯度下降法：

$\nabla\ell_{in}(\boldsymbol{w}) = \frac{1}{N}\sum_{n=1}^{N}\nabla\ell_n(\boldsymbol{w})$

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta\nabla\ell_{in}(\boldsymbol{w}_t)$

（2）随机梯度下降法：

$\nabla\ell_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B}\nabla\ell_n(\boldsymbol{w})$，$B$ 代表批量大小，最小可以为 1

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta\nabla\ell_{in}(\boldsymbol{w}_t)$

（3）Adagrad：

$\nabla\ell_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B}\nabla\ell_n(\boldsymbol{w})$

$\sigma_t = \sqrt{\frac{1}{t+1}\sum_{t=0}^{t}(\nabla\ell_{in}(\boldsymbol{w}))^2} + \varepsilon$， $\varepsilon$代表极小量，防止$\sigma_t$为 0

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \frac{\eta}{\sqrt{t+1}\sigma_t}\nabla\ell_{in}(\boldsymbol{w}_t)$

（4）RMSProp：

$\nabla\ell_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B}\nabla\ell_n(\boldsymbol{w})$

$\sigma_{t-1} = \sqrt{\frac{1}{t}\sum_{t=0}^{t-1}(\nabla\ell_{in}(\boldsymbol{w}))^2}$

$\sigma_t = \sqrt{\alpha(\sigma_{t-1})^2 + (1-\alpha)\big(\nabla\ell_{in}(\boldsymbol{w})\big)^2} + \varepsilon$

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \frac{\eta}{\sigma_t}\nabla\ell_{in}(\boldsymbol{w}_t)$

（5）动量法（Momentum）：

$\nabla\ell_{in}(\boldsymbol{w}) = \frac{1}{B}\sum_{n=1}^{B}\nabla\ell_n(\boldsymbol{w})$

$\boldsymbol{m}_{t+1} = \lambda\boldsymbol{m}_t - \eta\nabla\ell_{in}(\boldsymbol{w}_t)$，　　（$\boldsymbol{m}_{\boldsymbol{0}} = 0$）

$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t + \boldsymbol{m}_{t+1}$

（6）Adam

$\boldsymbol{m}_{\boldsymbol{t+1}} = \beta_1\boldsymbol{m}_t - (1-\beta_1)\nabla\ell_{in}(\boldsymbol{w}_{\boldsymbol{t}})$，　　（$\boldsymbol{m}_{\boldsymbol{0}} = 0$）

$\boldsymbol{v}_{\boldsymbol{t+1}} = \beta_2\boldsymbol{v}_t - (1-\beta_2)\big(\nabla\ell_{in}(\boldsymbol{w})\big)^2$，　　（$\boldsymbol{v}_{\boldsymbol{0}} = 0$）

$$\hat{\boldsymbol{m}}_{\boldsymbol{t+1}} = \boldsymbol{m}_{t+1}/(1-\beta_1^{t+1})$$

$$\hat{\boldsymbol{v}}_{\boldsymbol{t+1}} = \boldsymbol{v}_{t+1}/(1-\beta_2^{t+1})$$

$$\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta\hat{\boldsymbol{m}}_{t+1}/(\sqrt{\hat{\boldsymbol{v}}_{t+1}} + \varepsilon)$$