Lecture3 习题作业

1， 假设训练样本集为 $D=\{(\vec{x}_1,y_1)=((0.2,0.7)^T,1),(\vec{x}_2,y_2)=((0.3,0.3)^T,1),(\vec{x}_3,y_3)=((0.4,0.5)^T,1),(\vec{x}_4,y_4)=((0.6,0.5)^T,1),(\vec{x}_5,y_5)=((0.1,0.4)^T,1),(\vec{x}_6,y_6)=((0.4,0.6)^T,-1),(\vec{x}_7,y_7)=((0.6,0.2)^T,-1),(\vec{x}_8,y_8)=((0.7,0.4)^T,-1),(\vec{x}_9,y_9)=((0.8,0.6)^T,-1),(\vec{x}_{10},y_{10})=((0.7,0.5)^T,-1)\}$，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，讨论结果。

解：令 $D=\{(\vec{x}_i,y_i)=((1,x_i^1,x_i^2),y_i)\},i=1\sim10$，故可写出

$$\boldsymbol{X}^T=\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 0.2 & 0.3 & 0.4 & 0.6 & 0.1 & 0.4 & 0.6 & 0.7 & 0.8 & 0.7\\ 0.7 & 0.3 & 0.5 & 0.5 & 0.4 & 0.6 & 0.2 & 0.4 & 0.6 & 0.5\end{bmatrix}$$

$$\boldsymbol{y}=(1,1,1,1,1,-1,-1,-1,-1,-1)$$

进而计算可得

$$\boldsymbol{X}^\dagger=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$
$$=\begin{bmatrix}-0.16 & 0.7 & 0.11 & -0.1 & 0.67 & -0.13 & 0.63 & 0.04 & -0.55 & -0.20\\ -0.53 & -0.39 & -0.16 & 0.25 & -0.78 & -0.14 & 0.20 & 0.43 & 0.67 & 0.45\\ 1.1 & -0.88 & 0.14 & 0.17 & -0.41 & 0.64 & -1.33 & -0.31 & 0.7 & 0.19\end{bmatrix}$$

于是有

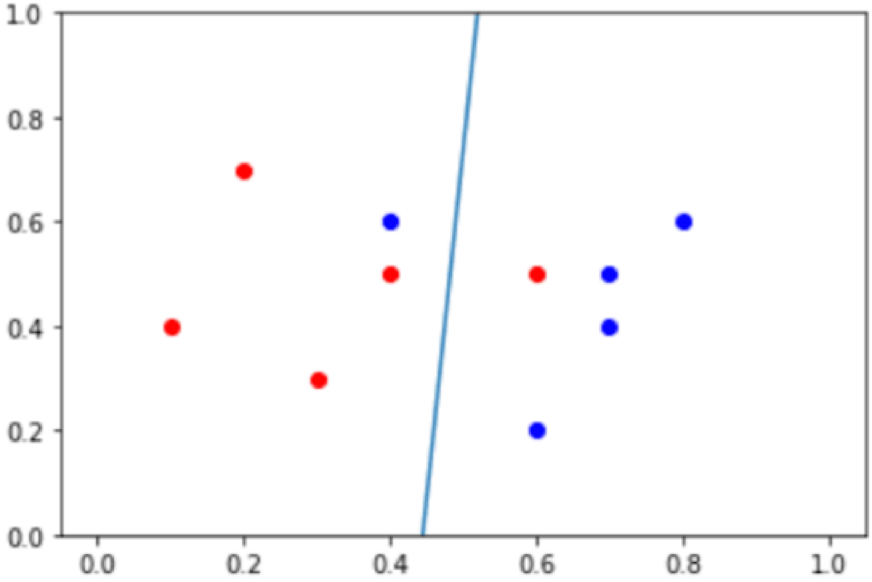
$$\boldsymbol{W}=\boldsymbol{X}^\dagger\boldsymbol{y}$$
$$=(1.43,-3.22,0.24)^T$$

因此这两类的分类函数为

$$h(\boldsymbol{x})=sign(\boldsymbol{W}^T\boldsymbol{x})$$

其中 $\boldsymbol{W}=(1.43,-3.22,0.24)^T$

并且将训练样本集$D=\{(\vec{x}_i,y_i)=((1,x_i^1,x_i^2),y_i)\},i=1\sim10$代入所得的分类函数$h(\boldsymbol{x})=sign(\boldsymbol{W}^T\boldsymbol{x})$可得该分类函数可大致正确分类训练样本。



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：

$$\|\mathbf{X}\boldsymbol{w}-Y\|^2=(\mathbf{X}\boldsymbol{w}-Y)^T(\mathbf{X}\boldsymbol{w}-Y)$$

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

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

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

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

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

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

3，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 L，假设任一单个样本 n 的梯度$\nabla L_n(\boldsymbol{w})$，$t$ 代表迭代次数

（1）梯度下降法：

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

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

（2）随机梯度下降法：

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

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

（3）Adagrad：

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

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

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

（4）RMSProp：

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

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

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

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

（5）动量法（Momentum）：

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

$\boldsymbol{m}_{t+1} = \lambda \boldsymbol{m}_t - \eta \nabla L_{in}(\boldsymbol{w}_t)$，　　（$\boldsymbol{m_0} = 0$）

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

（6）Adam

$\boldsymbol{m_{t+1}} = \beta_1 \boldsymbol{m}_t - (1-\beta_1)\nabla L_{in}(\boldsymbol{w_t})$，　　（$\boldsymbol{m_0} = 0$）

$\boldsymbol{v_{t+1}} = \beta_2 \boldsymbol{v}_t - (1-\beta_2)\left(\nabla L_{in}(\boldsymbol{w})\right)^2$，　　（$\boldsymbol{v_0} = 0$）

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

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

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