The Exercises of Mid-term

韩沅伯　15300180032

December 12, 2017

1 Linear Regression

1.1

Proof.

$$R^{emp}(\beta) = \sum_{i=1}^{n} \frac{1}{2} \left(y_i - x_i^T \beta\right)^2 = \frac{1}{2} \left(Y - X\beta\right)^T \left(Y - X\beta\right)$$

$$\frac{\partial R^{emp}(\beta)}{\partial \beta} = -X^T \left(Y - X\beta\right)$$

Set $\frac{\partial R^{emp}(\beta)}{\partial \beta} = 0$, we get $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$. □

1.2

Proof.

$$\begin{aligned} R^{emp}(\beta) &= \frac{C}{2}||\beta||_2^2 + \sum_{i=1}^{n} \frac{1}{2} \left(y_i - x_i^T \beta\right)^2 \\ &= \frac{C}{2}||\beta||_2^2 + \frac{1}{2} \left(Y - X\beta\right)^T \left(Y - X\beta\right) \\ \frac{\partial R^{emp}(\beta)}{\partial \beta} &= C\beta - X^T \left(Y - X\beta\right) \end{aligned}$$

Again set $\frac{\partial R^{emp}(\beta)}{\partial \beta} = 0$, we get $\hat{\beta} = \left(CI + X^T X\right)^{-1} X^T Y$. □

1.3

$$\begin{aligned} \hat{Y} = \Phi\hat{\beta} &= \Phi \left(CI + \Phi^T Phi\right)^{-1} \Phi^T Y \\ &= \Phi \left[C^{-1} I - C^{-2} \Phi^T \left(I + C^{-1} \Phi\Phi^T\right)^{-1} \Phi\right] \Phi^T Y \\ &= C^{-1} \left[G - C^{-1} G \left(I + C^{-1} G\right)^{-1} G\right] Y \\ &= \left(I + C^{-1} G\right)^{-1} Y, \qquad \text{where } G = \Phi\Phi^T. \end{aligned}$$

2 SVM- Fitting an SVM classifier by hand

2.1

$$\phi(x_1) = [1,0,0]^T, \phi(x_2) = [1,2,2]^T$$
$$\overrightarrow{\phi(x_1)\phi(x_2)} = [0,2,2]^T$$

$y+z=2$ is a decision boundary, so $[0,2,2]^T$ is parallel to the optimal w.

2.2

$$margin = \min_{i=1,2}\left\{\frac{y_i\left(w^T\phi(x_i)+w_0\right)}{||w||}\right\}$$

2.3

$$margin = \min_{i=1,2}\left\{\frac{y_i\left(w^T\phi(x_i)+w_0\right)}{||w||}\right\} = \frac{1}{||w||}$$

We shall solve

$$\begin{cases}\operatorname{argmin}_{w,w_0}\frac{1}{2}||w||^2\\ y_i\left(w^T\phi(x_i)+w_0\right)\geq 1 \quad i=1,2\end{cases}$$

Solving by inequality of arithmetic and geometric means,

$$w = [0,\frac{1}{2},\frac{1}{2}]^T$$

2.4

$$\begin{cases}-w_0\geq 1\\ 2+w_0\geq 1\end{cases} \implies w_0=-1$$

2.5

$$f(x) = -1+\frac{\sqrt{2}}{2}x+\frac{1}{2}x^2$$

3 Neural Network

3.1

Proof.

$$\begin{aligned}\frac{\partial\left(y_i \log \hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\right)}{\partial W_k^o} &= \frac{y_i}{\hat{y}_i}\frac{\partial \hat{y}_i}{\partial W_k^o}+\frac{1-y_i}{1-\hat{y}_i}\left(-\frac{\partial \hat{y}_i}{\partial W_k^o}\right)\\ &= (y_i-\hat{y}_i)\frac{1}{\hat{y}_i(1-\hat{y}_i)}\frac{\partial \hat{y}_i}{\partial W_k^o}\\ &= (y_i-\hat{y}_i)\,h_{ik}\end{aligned}$$

Thus,

$$\frac{\partial J}{\partial W_k^o}=\sum_{i=1}^{n}(y_i-\hat{y}_i)h_{ik}+CW_k^o$$

$$\frac{\partial J}{\partial W_{jk}^h}=\frac{\partial J}{\partial W_k^o}\frac{\partial W_k^o}{\partial W_{jk}^h}=CW_{jk}^h+\sum_{i=1}^{n}(\hat{y}_i-y_i)W_k^o h_{ik}(1-h_{ik})x_{ij}$$

□

3.2

Let $\hat{y}_h^l$ denote the value of h-th unit in l-th layer, W_{kh}^{l+1} denote the weight from y_h^l to the k-th unit in$(l+1)$-th layer, and $u_k^{l+1}=\sum_{h=1}^{H}W_{kh}^{l+1}\hat{y}_h^l$.

$$\frac{\partial J}{\partial W_{kh}^l}=\frac{\partial J}{\partial \hat{y}_k^l}\frac{\partial \hat{y}_k^l}{\partial u_k^l}\frac{u_k^l}{\partial W_{kh}^l}$$

Using chain rule repeatedly, we get

$$\frac{\partial J}{\partial W_{kh}^l}=\delta_k^l\hat{y}_h^{l-1}$$

where

$$\delta_k^l=\begin{cases}(\hat{y}_k^l-y_k)u_k^l(1-u_k^l) & l \text{ is the output layer,}\\ \left[\sum_{k=1}K(w_{kh}^{l+1}\delta_k^{l+1})\right]u_k^l(1-u_k^l) & l \text{ is the hidden layer.}\end{cases}$$