Homework 5

Questions: https://github.com/SUSTech-ML-Course/Machine-Learning-Course/blob/05b4e0bd4063b9d486f34ce557720b519947cbe3/lesson-homework/hw5/HW5.md

Question1

(a)

The negative logarithm of the likelihood function

$$E(w,\Sigma)=\frac{1}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^T\Sigma^{-1}(y(x_n,w)-t_n)+\frac{N}{2}ln|\Sigma|+C$$

Since Σ is fixed and known, we can get

$$E(w)=\frac{N}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)^T\Sigma^{-1}(y(x_n,w)-t_n)$$

Thus, the error function is the MSE

(b)

Since Σ is determined from the data, we can get the derivative of Σ

$$\frac{\partial E(w,\Sigma)}{\partial\Sigma}=\frac{N}{2}\Sigma^{-1}+\frac{1}{2}\sum_{n=1}^{N}(y(x_n,w)-t_n)'\Sigma^{-2}(y(x_n,w)-t_n)=0$$

Then we can get

$$\Sigma=\frac{1}{N}\sum_{n=1}^{N}(y(x_n,w)-t_n)^T(y(x_n-w)-t_n)$$

Question2

Since $\sigma(a)\in[0,1]$, and the network output $\in[-1,1]$, then we can design

$$h(a)=2\sigma(a)-1\in[-1,1]$$

Since

$$-1\leqslant y(x,w)\leqslant 1, C\in(-1,1)$$

Then we can get

$$\begin{aligned}E(w)&=-\sum_{n=1}^{N}(\frac{1+t_n}{2}\ln\frac{1+y_n}{2}+\frac{1-t_n}{2}\ln\frac{1-y_n}{2})\\&=-\frac{1}{2}\sum_{n=1}^{N}((1+t_n)\ln(1+y_n)+(1-t_n)\ln(1-y_n))+N\ln 2\end{aligned}$$

Question3

(a)

$$
\begin{aligned}
E(t) &= \int tN(t|\mu, \sigma^2 I)dt = \mu \\
E(\|t\|^2) &= \int \|t\|^2 N(t|\mu, \sigma^2 I)dt = L\sigma^2 + \|\mu\|^2
\end{aligned}
$$

where L is the dimension

Then get can get

$$
\begin{aligned}
E(t|x) &= \int tp(t|x)dt \\
&= \int t \sum_{k=1}^{K} \pi_k N(t|\mu_k, \sigma^2)dt \\
&= \sum_{k=1}^{K} \pi_k \int tN(t|\mu_k, \sigma^2)dt \\
&= \sum_{k=1}^{K} \pi_k \mu_k \\
&= \sum_{k=1}^{K} \pi_k(x)\mu_k(x)
\end{aligned}
$$

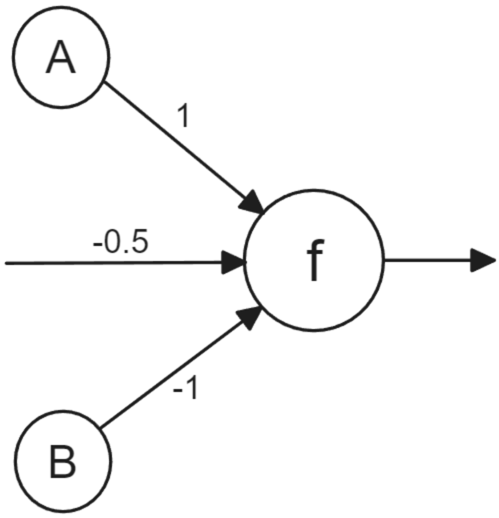
(b)

$$
\begin{aligned}
s^2(x) &= E[\|t - E(t|x)\|^2|x] \\
&= E[(t^2 - 2tE(t|x) + E[t|x]^2)|x] \\
&= E[t^2|x] - E[2tE(t|x)|x] + E[t|x]^2 \\
&= E[t^2|x] - E[t|x] \\
&= \int \|t\|^2 \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2)dt - \left\|\sum_{l=1}^{K} \pi_l, \mu_l\right\| \\
&= \sum_{k=1}^{K} \pi_k (L\sigma_k^2 + \|\mu_k\|^2) - \left\|\sum_{l=1}^{K} \pi_l \mu_l\right\| \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k \|\mu_k\|^2 - 2\left(\sum_{l=1}^{K} \pi_l \mu_l\right)\left(\sum_{k=1}^{k} \pi_k \mu_k\right) + \sum_{k=1}^{K} \pi_k \left\|\sum_{l=1}^{K} \pi_l \mu_l\right\|^2 \\
&= L\sum_{k=1}^{K} \pi_k \sigma_k^2 + \sum_{k=1}^{K} \pi_k \left\|\mu_k - \sum_{l=1}^{K} \pi_l \mu_l\right\|^2 \\
&= \sum_{k=1}^{K} \pi_k \left(L\sigma_k^2 + \left\|\mu_k - \sum_{l=1}^{K} \pi_l \mu_l\right\|^2\right)
\end{aligned}
$$

Question4

Yes, because it is linearly sperable

$$f(A,B)=\begin{cases}1, & A-B-0.5\geqslant 0\\ 0, & A-B-0.5\leqslant 0\end{cases}$$



Question5

(a) # of weights is $3*4*4*1=48$

(b) # of ReLU operations is $3*5*5=75$

(c) # of weights:

$N_{conv}=3*4*4*1=48$

$N_{pool}=0$

$N_{ReLU}=0$

$N_{FCL}=3*5*5*4=300$

$N=48+0+0+300=348$

(d) True

(e) There are too many weights to learn, and it is more likely to overfitting

Question6

(a) First, we define the following symbol for convenience

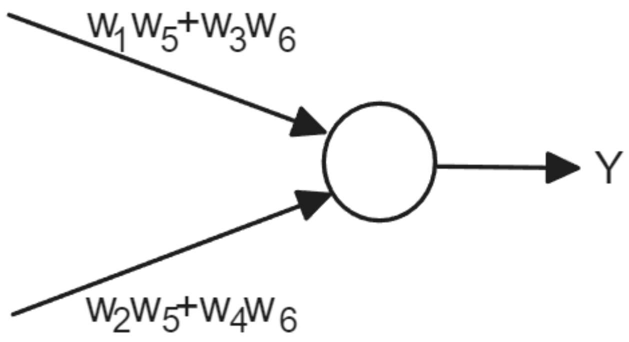
$$W_1=\begin{bmatrix}w_1 & w_3\\ w_2 & w_4\end{bmatrix}\quad W_2=[w_5\quad w_6]\quad A=[X_1\quad X_2]$$

Thus

$$Y=C\cdot W_2^TW_1^TA=C[w_1w_5+w_3w_6,w_2w_5+w_4w_6]\begin{bmatrix}X_1\\ X_2\end{bmatrix}$$

Then we have

$$W = \begin{bmatrix} w_1 w_5 + w_3 w_6 \\ w_2 w_5 + w_4 w_6 \end{bmatrix} \quad A = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$



(b) Yes, because the entire network can be seen to perform a chain of matrix multiplication

(c) Let $w_1 = w_3 = -5,\ w_2 = w_4 = -10,\ w_5 = 5,\ w_6 = -6$, then we have the whole network working like XOR