# Homework V

12110644 周思呈

# Question 1

**Consider a regression problem involving multiple target variables in which it is assumed that the distribution of the targets, conditioned on the input vector x, is a Gaussian of the form**

$$p(\mathbf{t}|\mathbf{x},\mathbf{w}) = \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{x},\mathbf{w}),\mathbf{\Sigma}) \tag{1}$$

**where $\mathbf{y}(\mathbf{x},\mathbf{w})$ is the output of a neural network with input vector $\mathbf{x}$ and wight vector $\mathbf{w}$, and $\mathbf{\Sigma}$ is the covariance of the assumed Gaussian noise on the targets.**

**(a) Given a set of independent observations of $\mathbf{x}$ and $\mathbf{t}$, write down the error function that must be minimized in order to find the maximum likelihood solution for $\mathbf{w}$, if we assume that $\mathbf{\Sigma}$ is fixed and known.**

The likelihood is

$$p(\mathbf{T} \mid \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}\left(\mathbf{t_n} \mid \mathbf{y}\left(\mathbf{x_n}, \mathbf{w}\right), \mathbf{\Sigma}\right) \tag{2}$$

Take negative logarithm

$$E(\mathbf{w},\mathbf{\Sigma}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\left[\mathbf{y}\left(\mathbf{x_n},\mathbf{w}\right)-\mathbf{t_n}\right]^T \mathbf{\Sigma}^{-1}\left[\mathbf{y}\left(\mathbf{x_n},\mathbf{w}\right)-\mathbf{t_n}\right]\right\} + \frac{N}{2}\ln|\mathbf{\Sigma}| + \text{ const} \tag{3}$$

If $\mathbf{\Sigma}$ is fixed and known

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\left[\mathbf{y}\left(\mathbf{x_n},\mathbf{w}\right)-\mathbf{t_n}\right]^T \mathbf{\Sigma}^{-1}\left[\mathbf{y}\left(\mathbf{x_n},\mathbf{w}\right)-\mathbf{t_n}\right]\right\} + \text{ const} \tag{4}$$

**(b) Now assume that $\mathbf{\Sigma}$ is also to be determined from the data, and write down an expression for the maximum likelihood solution for $\mathbf{\Sigma}$. (Note: The optimizations of $\mathbf{w}$ and $\mathbf{\Sigma}$ are now coupled.)**

By rewriting $E(\mathbf{w},\mathbf{\Sigma})$ we get

$$-\frac{N}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\sum_{n=1}^{N}(\mathbf{t}_n-\mathbf{y}_n)(\mathbf{t}_n-\mathbf{y}_n)^{\mathrm{T}}\right]. \tag{5}$$

We can maximize this by setting the derivative w.r.t. $\mathbf{\Sigma}^{-1}$ to zero, yielding

$$\mathbf{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{t}_n-\mathbf{y}_n)(\mathbf{t}_n-\mathbf{y}_n)^{\mathrm{T}}. \tag{6}$$

# Question 2

**The error function for binary classification problems was derived for a network having a logistic-sigmoid output activation function, so that $0 \leq y(\mathbf{x}, \mathbf{w}) \leq 1$, and data having target values $t \in \{0, 1\}$. Derive the corresponding error function if we consider a network having an output $-1 \leq y(\mathbf{x}, \mathbf{w}) \leq 1$ and target values $t = 1$ for class $\mathcal{C}_1$ and $t = -1$ for class $\mathcal{C}_2$. What would be the appropriate choice of output unit activation function?**

**Hint. The error function is given by:**

$E(\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1 - t_n)\ln(1 - y_n)\}$.

Mapping the original output range to the new output range, we could set

$$y = 2\sigma(a) - 1 \tag{7}$$

The conditional distribution of targets given inputs is

$$p(t \mid \mathbf{x}, \mathbf{w}) = \left[\frac{1 + y(\mathbf{x}, \mathbf{w})}{2}\right]^{(1+t)/2}\left[\frac{1 - y(\mathbf{x}, \mathbf{w})}{2}\right]^{(1-t)/2} \tag{8}$$

Where $\frac{[1+y(\mathbf{x},\mathbf{w})]}{2}$ represents the conditional probability $p\left(C_1 \mid x\right)$. The likelihood is

$$p(\mathbf{T} \mid \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(t_n \mid \mathbf{x}_n, \mathbf{w}_n) \tag{9}$$

Take negative logarithm

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N}\left\{\frac{1 + t_n}{2}\ln\frac{1 + y_n}{2} + \frac{1 - t_n}{2}\ln\frac{1 - y_n}{2}\right\} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\left\{(1 + t_n)\ln\left(1 + y_n\right) + (1 - t_n)\ln\left(1 - y_n\right)\right\} + N\ln 2 \end{aligned} \tag{10}$$

The choice of output unit activation function can be

$$\tanh(a/2) = \frac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} \tag{11}$$

# Question 3

**Verify the following results for the conditional mean and variance of the mixture density network model.**

(a) $\mathbb{E}[\mathbf{t}|\mathbf{x}] = \int \mathbf{t}p(\mathbf{t}|\mathbf{x})\mathrm{d}\mathbf{t} = \sum_{k=1}^{K}\pi_k(\mathbf{x})\mu_k(\mathbf{x})$.

$$\begin{aligned} \mathbb{E}[\mathbf{t} \mid \mathbf{x}] &= \int \mathbf{t}p(\mathbf{t} \mid \mathbf{x})d\mathbf{t} \\ &= \int \mathbf{t}\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\mathbf{t} \mid \boldsymbol{\mu}_k, \sigma_k^2\right)d\mathbf{t} \\ &= \sum_{k=1}^{K}\pi_k\int \mathbf{t}\mathcal{N}\left(\mathbf{t} \mid \boldsymbol{\mu}_k, \sigma_k^2\right)d\mathbf{t} \\ &= \sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k \end{aligned} \tag{12}$$

(b) $s^2(\mathbf{x})=\sum_{k=1}^K \pi_k(\mathbf{x})\{\sigma_k^2(\mathbf{x})+\|\mu_k(\mathbf{x})-\sum_{l=1}^K \pi_l(\mathbf{x})\mu_l(\mathbf{x})\|^2\}$.

$$\begin{aligned} s^2(\mathbf{x}) &= \mathbb{E}\left[\|\mathbf{t}-\mathbb{E}[\mathbf{t}\mid\mathbf{x}]\|^2 \mid \mathbf{x}\right] = \mathbb{E}\left[\left(\mathbf{t}^2-2\mathbf{t}\mathbb{E}[\mathbf{t}\mid\mathbf{x}]+\mathbb{E}[\mathbf{t}\mid\mathbf{x}]^2\right)\mid\mathbf{x}\right] \\ &= \mathbb{E}\left[\mathbf{t}^2\mid\mathbf{x}\right]-\mathbb{E}[2\mathbf{t}\mathbb{E}[\mathbf{t}\mid\mathbf{x}]\mathbf{x}]+\mathbb{E}[\mathbf{t}\mid\mathbf{x}]^2 = \mathbb{E}\left[\mathbf{t}^2\mid\mathbf{x}\right]-\mathbb{E}[\mathbf{t}\mid\mathbf{x}]^2 \\ &= \int\|\mathbf{t}\|^2\sum_{k=1}^K \pi_k\mathcal{N}\left(\boldsymbol{\mu}_k,\sigma_k^2\right)d\mathbf{t}-\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= \sum_{k=1}^K \pi_k\int\|\mathbf{t}\|^2\mathcal{N}\left(\boldsymbol{\mu}_k,\sigma_k^2\right)d\mathbf{t}-\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \end{aligned} \tag{13}$$

Because $\mathbb{E}\left[\|\mathbf{t}\|^2\right]=\int\|\mathbf{t}\|^2\mathcal{N}\left(\mathbf{t}\mid\boldsymbol{\mu},\sigma^2\mathbf{I}\right)d\mathbf{t}=L\sigma^2+\|\boldsymbol{\mu}\|^2$, we have

$$\begin{aligned} \text{above} &= \sum_{k=1}^K \pi_k\left(L\sigma_k^2+\|\boldsymbol{\mu}_k\|^2\right)-\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= L\sum_{k=1}^K \pi_k\sigma_k^2+\sum_{k=1}^K \pi_k\|\boldsymbol{\mu}_k\|^2-\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= L\sum_{k=1}^K \pi_k\sigma_k^2+\sum_{k=1}^K \pi_k\|\boldsymbol{\mu}_k\|^2-2\times\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2+1\times\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= L\sum_{k=1}^K \pi_k\sigma_k^2+\sum_{k=1}^K \pi_k\|\boldsymbol{\mu}_k\|^2-2\left(\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right)\left(\sum_{k=1}^K \pi_k\boldsymbol{\mu}_k\right)+\left(\sum_{k=1}^K \pi_k\right)\left\|\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= L\sum_{k=1}^K \pi_k\sigma_k^2+\sum_{k=1}^K \pi_k\left\|\boldsymbol{\mu}_k-\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2 \\ &= \sum_{k=1}^K \pi_k\left(L\sigma_k^2+\left\|\boldsymbol{\mu}_k-\sum_{l=1}^K \pi_l\boldsymbol{\mu}_l\right\|^2\right) \end{aligned} \tag{14}$$

# Question 4

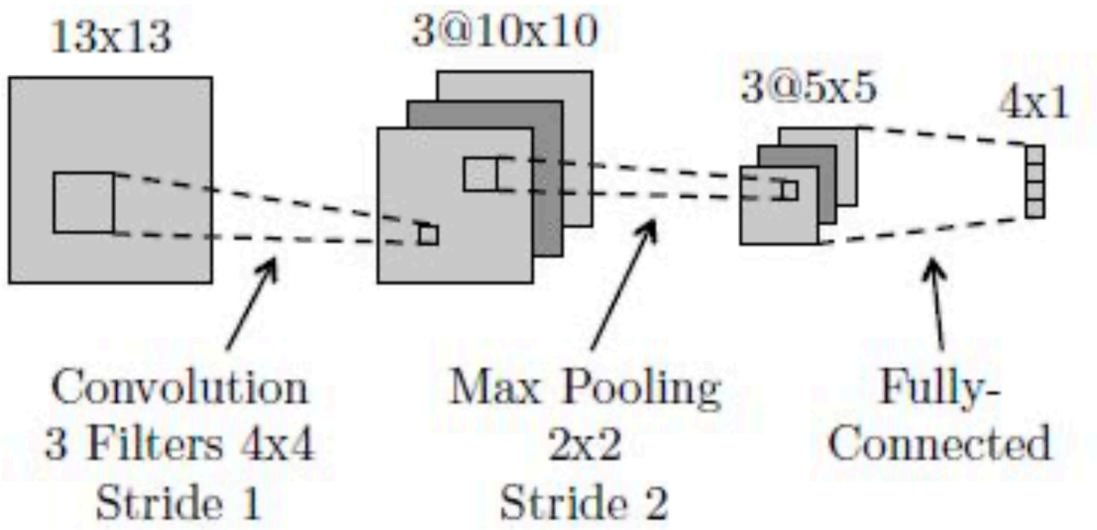
**Can you represent the following boolean function with a single logistic threshold unit (i.e., a single unit from a neural network)? If yes, show the weights. If not, explain why not in 1-2 sentences.**

| A | B | f(A,B) |
|---|---|---|
| 1 | 1 | 0 |
| 0 | 0 | 0 |
| 1 | 0 | 1 |
| 0 | 1 | 0 |

$$F(A,B)=\{A-B-0.5>0\} \tag{15}$$

# Question 5

**Below is a diagram of a small convolutional neural network that converts a 13x13 image into 4 output values. The network has the following layers/operations from input to output: convolution with 3 filters, max pooling, ReLU, and finally a fully-connected layer. For this network we will not be using any bias/offset parameters (b). Please answer the following questions about this network.**




**(a) How many weights in the convolutional layer do we need to learn?**

$$3 \times 4 \times 4 = 48 \tag{16}$$

**(b) How many ReLU operations are performed on the forward pass?**

$$3 \times 5 \times 5 = 75 \tag{17}$$

**(c) How many weights do we need to learn for the entire network?**

$$48 + 75 \times 4 = 348 \tag{18}$$

**(d) True or false: A fully-connected neural network with the same size layers as the above network $(13 \times 13 \rightarrow 3 \times 10 \times 10 \rightarrow 3 \times 5 \times 5 \rightarrow 4 \times 1)$ can represent any classifier?**

True.

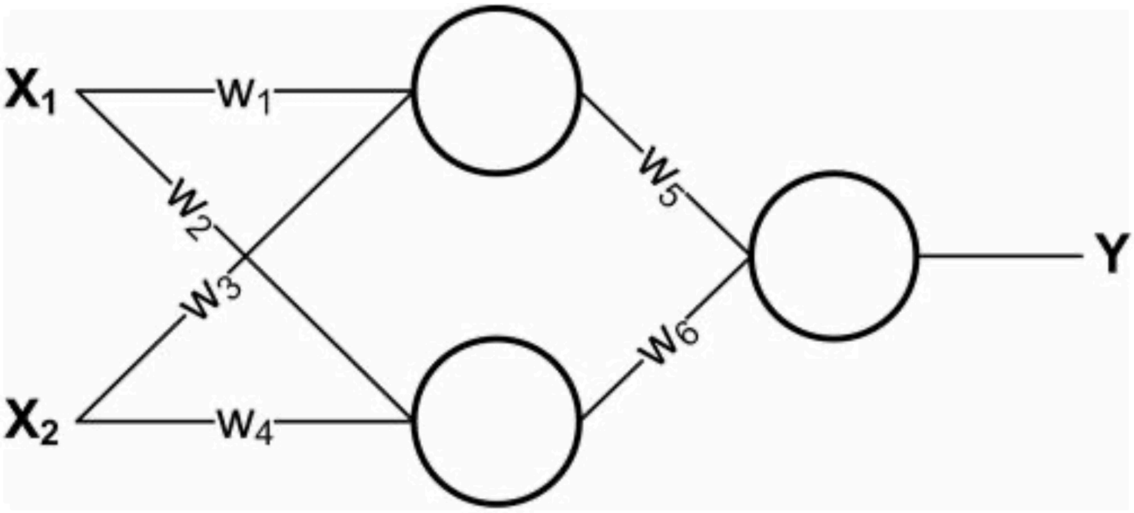
**(e) What is the disadvantage of a fully-connected neural network compared to a convolutional neural network with the same size layers?**

1. **Parameter Efficiency**. CNNs have fewer parameters.
2. **Spatial Hierarchy**. CNNs are designed to handle spatial hierarchies. FCNNs, on the other hand, treat all inputs equally.

# Question 6

**The neural networks shown in class used logistic units: that is, for a given unit $U$, if $A$ is the vector of activations of units that send their output to $U$, and $W$ is the weight vector corresponding to these outputs, then the activation of $U$ will be $(1 + \exp(W^T A))^{-1}$. However, activation functions could be anything. In this exercise we will explore some others. Consider the following neural network,**

**consisting of two input units, a single hidden layer containing two units, and one output unit:**




**(a) Say that the network is using linear units: that is, defining $W$ and and $A$ as above, the output of a unit is $C * W^T A$ for some fixed constant $C$. Let the weight values $w_i$ be fixed. Re-design the neural network to compute the same function without using any hidden units. Express the new weights in terms of the old weights and the constant $C$.**

Connect the input X1 to the output, $weight = C \times (w_5 \times w_1 + w_6 \times w_2)$.

Connect the input X2 to the output, $weight = C \times (w_5 \times w_3 + w_6 \times w_4)$.

**(b) Is it always possible to express a neural network made up of only linear units without a hidden layer? Give a one-sentence justification.**

Yes. We can express all weights to one linear layer weight by linear combination.

**(c) Another common activation function is a theshold, where the activation is $t(W_T A)$ where $t(x)$ is 1 if $x > 0$ and 0 otherwise. Let the hidden units use sigmoid activation functions and let the output unit use a threshold activation function. Find weights which cause this network to compute the XOR of $X_1$ and $X_2$ for binary-valued $X_1$ and $X_2$. Keep in mind that there is no bias term for these units.**

$$w_1 = 2, w_2 = 1, w_3 = 2, w_4 = 1, w_5 = 1 + e^{-4}, w_6 = -(1 + e^{-2}) \tag{19}$$

| $X_1$ | $X_2$ | $a_1 = sigmod(w_1 X_1 + w_3 X_2)$ | $a_2 = sigmod(w_2 X_1 + w_4 X_2)$ | $b = a_1 w_5 + a_2 w_6$ | $y = t(b)$ |
|---|---|---|---|---|---|
| 0 | 0 | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}(e^{-4} - e^{-2})$ | 0 |
| 0 | 1 | $\frac{1}{1+e^{-2}}$ | $\frac{1}{1+e^{-1}}$ | $\frac{1+e^{-4}}{1+e^{-2}} - \frac{1+e^{-2}}{1+e^{-1}}$ | 1 |
| 1 | 0 | $\frac{1}{1+e^{-2}}$ | $\frac{1}{1+e^{-1}}$ | $\frac{1+e^{-4}}{1+e^{-2}} - \frac{1+e^{-2}}{1+e^{-1}}$ | 1 |
| 1 | 1 | $\frac{1}{1+e^{-4}}$ | $\frac{1}{1+e^{-2}}$ | 0 | 0 |