Q1.

(1) $y = (\sum_i x_i^p)^{\frac{1}{p}}$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x_i}$$

$$\frac{\partial}{\partial x_i}(\sum_i x_i^p) = p \cdot x_i^{p-1}$$

$$\frac{\partial y}{\partial x_i} = \frac{1}{p}(\sum_i x_i^p)^{\frac{1}{p}-1} \cdot \frac{\partial}{\partial x_i}(\sum_i x_i^p)$$

$$= x_i^{p-1} (\sum_i x_i^p)^{\frac{1}{p}-1}$$

$$x_i' = \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x_i} = y' \cdot x_i^{p-1} (\sum_i x_i^p)^{\frac{1}{p}-1}$$

(2).

$$\frac{\partial y}{\partial x_i} = \frac{1}{\beta} \cdot \frac{1}{\frac{1}{n}\sum_i e^{\beta x_i}} \cdot \frac{\partial}{\partial x_i}\left(\frac{1}{n}\sum_i e^{\beta x_i}\right)$$

$$= \frac{1}{\beta} \cdot \frac{1}{\frac{1}{n}\sum_i e^{\beta x_i}} \cdot \frac{1}{n} \cdot e^{\beta x_i} \cdot \beta$$

$$= \frac{e^{\beta x_i}}{\sum_i e^{\beta x_i}}$$

$$x_i' = \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x_i} = y' \cdot \frac{e^{\beta x_i}}{\sum_i e^{\beta x_i}}$$

Q2:

The only difference between Model 1 and Model 2 is about sigmoid function. We can obviously know that whe $y = w^T x + b$, we always get linear input, and we can only solve feature with linear relationship and get linear output at the same time. But if we get sigmoid function, we can solve other complex relationship. We can use it for neural network. And it also can get result as Probabilistic type, make the result in (0, 1), make next process easier and convinient.

(means between 0 to 1)

Q3.

For A:

Root entropy of class outcome $= H(Y) = -\frac{2}{7}\log_2(\frac{2}{7}) - \frac{5}{7}\log_2(\frac{5}{7})$

≈ 0.86

Leaf conditional entropy of class outcome: $H(Y|\text{left}) = 0$

~~IG~~ $H(Y|\text{right}) \approx 0.97$

$IG(\text{splitA}) \approx 0.86 - (\frac{2}{7}\cdot 0 + \frac{5}{7}\cdot 0.97) \approx 0.17$

For B:

Root entropy of class outcome $= H(Y) = -\frac{2}{7}\log_2(\frac{2}{7}) - \frac{5}{7}\log_2(\frac{5}{7})$

≈ 0.86

Leaf conditional entropy of class outcome: $H(Y|\text{left}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

≈ 0.81

$H(Y|\text{right}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$

≈ 0.92

$IG(\text{splitB}) \approx 0.86 - (\frac{4}{7}\cdot 0.81 + \frac{3}{7}\cdot 0.92) \approx 0.006$

$IG(\text{splitA}) > IG(\text{splitB})$. split A is better.

Q4

(1).

Y:

-1	0	-3	6	-5	5	-6	4

(2).

Z:

0	0	0	6	0	5	0	4

(3).

V:

0	6	5	4

(4).

conv: each filter has $3 \times 1 \times 1 + 1 = 4$ params

pool: no parameters.

$\therefore$ 4 parameters.

(5)

consider the kernel move from left to right:

$W_2 = (W_1 - F + 2P)/S + 1$ $\quad F = 3$

$H_2 = (H_1 - F + 2P)/S + 1$

so we get:

$$\frac{8-3+2P}{S} + 1 = 8 \quad ①$$

$$\Rightarrow$$

$$\frac{1-3+2P}{S} + 1 = 1 \quad ②$$

③ → ①: $S = 1$

$\because S \neq 0 \therefore P = 1$

~~consider the kernel~~

Assume that kernel need change line, we should consider the possibility of this kernel move ~~from~~ to next line (if there has zero-padding in the upside and down side)

So we should try it for provement:

this time:

still: $W_2 = (W_1 - F + 2P)/S + 1$ but, $F = 1$

$H_2 = (H_1 - F + 2P)/S + 1$

so we get

$$\frac{8-1+2P}{S} + 1 = 8 \quad ①$$

$$\frac{1-1+2P}{S} + 1 = 1 \quad ②$$

$$\Rightarrow \quad S \neq 0 \Rightarrow P = 0 \quad ③$$

$$③ \rightarrow ①: \quad S = 1$$

so, we don't need zero padding in the upside and down side of sub-system.

We considered 4 side of the sub-system, we can say that stride of this convolution unit must be 1.

Q5:

$$L = \frac{1}{2}(\ln y)^2 + \frac{1}{2}(\ln t)^2 - \ln y \cdot \ln t$$

let $k = \ln y$: $\frac{\partial k}{\partial y} = \frac{1}{y}$ $\quad L = \frac{1}{2}k^2 + \frac{1}{2}(\ln t)^2 - \ln t \cdot k$

$$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial k} \cdot \frac{\partial k}{\partial y} \qquad \frac{\partial L}{\partial k} = k + \frac{1}{2}(\ln t)^2 - \ln t$$

$$= [\ln y + [\tfrac{1}{2}(\ln t)^2] - \ln t] \cdot \frac{1}{y}$$

$$= \frac{1}{y}\ln y + \frac{1}{2y}(\ln t)^2 - \frac{1}{y}\ln t$$

(1) $\bar{y} = \frac{\partial L}{\partial y} = \frac{1}{y}\ln y + \frac{1}{2y}(\ln t)^2 - \frac{1}{y}\ln t$

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial z}$$

$$\frac{\partial y}{\partial z} = \begin{cases} e^z & (z > 0) \\ 0 & (z \le 0) \end{cases}$$

(2) $\bar{z} = \frac{\partial L}{\partial z} = \begin{cases} \bar{y} \cdot e^z & (z > 0) \\ 0 & (z \le 0) \end{cases}$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W_2}$$

$$\frac{\partial z}{\partial W_2} = x^2$$

(3) $$\overline{W_2} = \frac{\partial L}{\partial W_2} = \begin{cases} \overline{y} \cdot e^{z} \cdot x^2 & (z > 0) \\ \cancel{y}\ 0 & (z \le 0) \end{cases}$$

Q6.

(1).

$$L(\theta) = \sum_{i=1}^{N} \log p(z^{(i)}, x^{(i)}) = \sum_{i=1}^{N} \log \{p(x^{(i)} | z^{(i)}) p(z^{(i)})\}$$

$$= \sum_{i=1}^{N} \log \{p(z^{(i)}) \prod_{j=1}^{D} p(x_j^{(i)} | z^{(i)})\}$$

$$= \sum_{i=1}^{N} \left[\log p(z^{(i)}) + \sum_{j=1}^{D} \log p(x_j^{(i)} | z^{(i)})\right]$$

$$= \sum_{i=1}^{N} \log P(z^{(i)}) + \sum_{j=1}^{D} \sum_{i=1}^{N} \log p(x_j^{(i)} | c^{(i)})$$

(2). $$r^{(i)} = \frac{\theta N(x^{(i)}; \mu, \sigma_1)}{\theta N(x^{(i)}; \mu, \sigma_1) + (1-\theta) N(x^{(i)}; \mu, \sigma_0)}$$

(3) $$L(\mu, \sigma_0, \sigma_1, \theta) = \sum_{i=1}^{N} r^{(i)} \log(N(x^{(i)} | \mu, \sigma_1) + r^{(i)} \log \theta$$
$$+ (1-r^{(i)}) \log N(x^{(i)} | \mu, \sigma_0)) + (1-r^{(i)}) \log(1-\theta)$$

(4)

$\frac{\partial L}{\partial \mu} = 0$, we get :

$$\sum_i^N \gamma^{(i)} \frac{(x^{(i)} - \mu)}{\sigma_1^2} + (1-\gamma^{(i)}) \frac{(x^{(i)} - \mu)}{\sigma_0^2} = 0$$

$$\therefore \quad \sum_i^N (x^{(i)} - \mu) \left(\frac{\gamma^{(i)}}{\sigma_1^2} + \frac{(1-\gamma^{(i)})}{\sigma_0^2} \right) = 0$$

$$\therefore \quad \sum_i^N (x^{(i)} - \mu)\left(\sigma_0^2 \gamma^{(i)} + \sigma_1^2 (1-\gamma^{(i)})\right) = 0$$

$$\mu \leftarrow \frac{\sum_{i=1}^N x^{(i)} \left(\gamma^{(i)} \sigma_0^2 + (1-\gamma^{(i)}) \sigma_1^2\right)}{\sum_{i=1}^N \left(\gamma^{(i)} \sigma_0^2 + (1-\gamma^{(i)}) \sigma_1^2\right)}$$