ML HW 3.

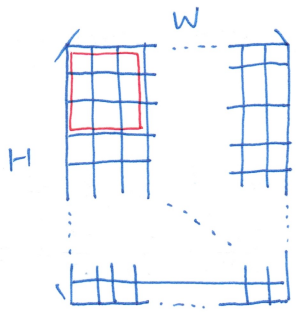
Q1. Given batch of image data with shape (B, W, H, C_{input})

$\Rightarrow$ Conv2D(C_{in}, C_{out}, kernel-size=(k_1, k_2), stride = (s_1, s_2), padding = (P_1, P_2))

(Solution) As the reference of pytorch document

$$H_{out} = \left\lfloor \frac{H_{in} + 2 \times P_1 - (k_1 - 1) - 1}{s_1} + 1 \right\rfloor$$

$$W_{out} = \left\lfloor \frac{W_{in} + 2P_2 - (k_2 - 1)}{s_2} + 1 \right\rfloor$$



if k=(3x3), stride(1,1), padding(0,0) $\Rightarrow$ $W_{new} = W-2$, $H_{new} = H-2$

if k=(5x5), stride(1,1), padding(0,0) $\Rightarrow$ $W_{new} = W-4$, $H_{new} = H-4$

$W_{new} = W - (k_1 - 1)$

$H_{new} = H - (k_2 - 1)$

if $k = k_1 \times k_2$, stride (1×1), padding (P_1, P_2) $\Rightarrow$ $W_{new} = W + 2P_1 - (k_1 - 1)$, $H_{new} = H + 2P_2 - (k_2 - 1)$

~~if k=1x1, Padding(0,0), stride(1x1) $\Rightarrow$ $W_{new} = W$, $H_{new} = H$~~

~~k=1x1, Padding(0,0), stride(2x2) $\Rightarrow$ $W_{new} = \lfloor \frac{W}{2} + 1 \rfloor$, $H_{new} = \lfloor \frac{H}{2} + 1 \rfloor$~~

~~k=1x1, padding(0,0), stride(3,3) $\Rightarrow$ $W_{new} = \lfloor \frac{W}{3} + 1 \rfloor$, $H_{new} = \lfloor \frac{H}{3} + 1 \rfloor$~~

$$W_{new} = \left\lfloor \frac{W + 2P_1 - (k_1 - 1)}{s_1} + 1 \right\rfloor$$

$$H_{new} = \left\lfloor \frac{H + 2P_2 - (k_2 - 1)}{s_2} + 1 \right\rfloor$$

if W=16, k=1, padding=0

$$W_{new} = \begin{cases} 8, & \text{if } s=2. \quad \lfloor \frac{16-1}{2} + 1 \rfloor = 8 \\ 6, & \text{if } s=3 \quad \lfloor \frac{16-1}{3} + 1 \rfloor = 6 \\ 4, & \text{if } s=4 \quad \lfloor \frac{16-1}{4} + 1 \rfloor = 4 \\ 4, & \text{if } s=5 \quad \lfloor \frac{16-1}{5} + 1 \rfloor = 4 \end{cases}$$

$\Rightarrow$

$$W_{new} = \left\lfloor \frac{W_{in} + 2P_1 - (k_1 - 1) - 1}{s_1} + 1 \right\rfloor$$

$$H_{new} = \left\lfloor \frac{H_{in} + 2P_2 - (k_2 - 1) - 1}{s_2} + 1 \right\rfloor$$

Q.E.D.

Q2 Given Input: value of x over a mini-batch: $\mathcal{B} = \{x_{1\dots m}\}$

Output: $y_i = BN_{\gamma,\beta}(x_i)$

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$$

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

$$y_i = \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$$

Before solution, we can review the differentiation w.r.t multi-variable function first.

Given $u(x,y)$ where $x(r,t)$ and $y(r,t)$, then to determine the value of $\frac{\partial u}{\partial r}, \frac{\partial u}{\partial t}$ we need to use multi-variable chain rules that is $\frac{\partial u}{\partial r} = \frac{\partial u}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial r}$

<sol> We have 5 functions

① $l(y)$ ② $y(\hat{x}, \gamma, \beta)$ ③ $\hat{x}(x, \mu_B, \sigma_B^2)$ ④ $\mu_B(x)$ ⑤ $\sigma_B^2(x, \mu_B)$

To determine $\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i}\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial l}{\partial \mu_B}\frac{\partial \mu_B}{\partial x_i} + \frac{\partial l}{\partial \sigma_B^2}\cdot\frac{\partial \sigma_B^2}{\partial x_i}$

$$= \frac{\partial l}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i} + \left(\frac{\partial l}{\partial \hat{x}}\cdot\frac{\partial \hat{x}}{\partial \mu_B} + \frac{\partial l}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial \mu_B}\right)\frac{\partial \mu_B}{\partial x_i} + \frac{\partial l}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial \sigma_B^2}\frac{\partial \sigma_B^2}{\partial x_i}$$

($\hat{x}$ → to $\{x_1 \dots x_m\}$)

$$= \frac{\partial l}{\partial \hat{x}_i}\frac{1}{\sqrt{\sigma_B^2+\varepsilon}} + \left(\sum_{j=1}^{m}\frac{\partial l}{\partial \hat{x}_j}\cdot\frac{-1}{\sqrt{\sigma_B^2+\varepsilon}} + \frac{\partial l}{\partial \sigma_B^2}\cdot\left(\frac{-2}{m}\right)\sum_{i=1}^{m}(x_i-\mu_B)\right)\frac{1}{m} + \sum_{j=1}^{m}\frac{\partial l}{\partial \hat{x}_j}\cdot\left(-\frac{1}{2}\right)\sum_{j=1}^{m}\frac{x_j-\mu_B}{(\sigma_B^2+\varepsilon)^{3/2}}\cdot\frac{2}{m}(x_i-\mu_B)$$

↓ cont. to next page

$$= \frac{\partial \ell}{\partial y_i} \cdot \gamma \left(\frac{1}{\sqrt{\sigma_B^2+\varepsilon}} - \frac{\sum_{j=1}^{m} \frac{\partial \ell}{\partial x_j}}{m\sqrt{\sigma_B^2+\varepsilon}} + \frac{\partial \ell}{\partial \sigma_B^2} \cdot \underbrace{\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_B)} - \frac{1}{m}\sum_{j=1}^{m} \frac{(x_j-\mu_B)}{(\sigma_B^2+\varepsilon)^{3/2}} \frac{\partial \ell}{\partial x_j}(x_i-\mu_B) \right)$$

$$\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} \mu_B$$

$$= \mu_B - \frac{m\mu_B}{m} = 0$$

$$= \frac{\partial \ell}{\partial y_i} \cdot \gamma \left(\frac{1}{\sqrt{\sigma_B^2+\varepsilon}} - \frac{\sum_{j=1}^{m} \frac{\partial \ell}{\partial x_j}}{m\sqrt{\sigma_B^2+\varepsilon}} - \frac{1}{m}\sum_{j=1}^{m} \frac{(x_j-\mu_B)\frac{\partial \ell}{\partial x_j}}{(\sigma_B+\varepsilon)^{3/2}}(x_i-\mu_B) \right)$$

$$= \frac{1}{m\sqrt{\sigma_B^2+\varepsilon}} \left(\frac{\partial \ell}{\partial y_i} \gamma m \frac{\partial \ell}{\partial y} - \sum_{j=1}^{m} \frac{\partial \ell}{\partial x_j} - \sum_{j=1}^{m} (x_j-\mu_B) \frac{\partial \ell}{\partial x_j}(x_i-\mu_B)(\sigma_B+\varepsilon)^{-1} \right)$$

$$= \frac{1}{m\sqrt{\sigma_B^2+\varepsilon}} \left(m \frac{\partial \ell}{\partial y_i}\gamma - \sum_{j=1}^{m} \frac{\partial \ell}{\partial x_i} - \sum_{j=1}^{m} \underbrace{\frac{x_j-\mu_B}{\sqrt{\sigma_B+\varepsilon}}}_{\hat{x}_j} \cdot \frac{\partial \ell}{\partial x_j} \underbrace{\left(\frac{(x_i-\mu_B)}{\sqrt{\sigma_B+\varepsilon}} \right)}_{\hat{x}_i} \right)$$

$$= \boxed{\frac{1}{m\sqrt{\sigma_B^2+\varepsilon}} \left(m \frac{\partial \ell}{\partial y_i}\gamma - \sum_{j=1}^{m} \frac{\partial \ell}{\partial x_i} - \hat{x}_i \sum_{j=1}^{m} \hat{x}_j \frac{\partial \ell}{\partial x_j} \right)}$$ Q.E.D.

$$\frac{\partial \ell}{\partial \gamma} = \frac{\partial \ell}{\partial y} \frac{\partial y}{\partial \gamma} = \boxed{\sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i} \gamma}$$ Q.E.D.

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial y} \frac{\partial y}{\partial \beta} = \boxed{\sum_{i=1}^{m} \frac{\partial \ell}{\partial y_i}}$$ Q.E.D.

Q3 Given N-Dimensional vector of real number

$$\text{Softmax}(\vec{z}) = S(\vec{z}) : \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} \to \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix}$$

$$S_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}} \quad \forall j \in 1 \dots N \Rightarrow S(z_j) = S_j = \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}}$$

$$L(y_i, \hat{y}_i) = -\sum_i y_i \log(\hat{y}_i) \text{ where } y_i \text{ is G.T. and } \hat{y}_i = S(\vec{z}_i)$$

(a) If $i = j$, then $\frac{\partial S(z_i)}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} = \frac{e^{z_i} \sum_{k=1}^{N} e^{z_k} - (e^{z_j})^2}{(\sum_{k=1}^{N} e^{z_k})^2}$

$$= \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} - \left(\frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}}\right)^2 = S(z_i)(1 - S(z_j))$$

If $i \neq j$, then $\frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} = -e^{z_i} (\sum_{k=1}^{N} e^{z_k})^{-2} e^{z_j} = -\frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} \cdot \frac{e^{z_j}}{\sum_{k=1}^{N} e^{z_k}}$

$$= -S(z_i) S(z_j)$$

With respect to these 2 cases, we can create a general form

$$\frac{\partial S(z_i)}{\partial z_j} = S(z_i)(\delta_{ij} - S(z_j)) \text{ where } \delta_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

(b) Derive that $\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$

$$\frac{\partial L}{\partial z_i} = \frac{\partial}{\partial z_i} - \sum_i y_i \log(S(z_i)) = -\sum_i y_i \frac{\partial \log(S(z_i))}{\partial z_i} = -\sum_i \frac{y_i}{\hat{y}_i} \frac{\partial S(z_i)}{\partial z_i}$$

$$= -\frac{y_j}{\hat{y}_j} \frac{\partial S(z_j)}{\partial z_i} - \sum_{i \neq j} \frac{y_i}{\hat{y}_i} \frac{\partial S(z_i)}{\partial z_i} = -\frac{y_j}{\hat{y}_j} S(z_i)(1 - S(z_j)) - \sum_{i \neq j} \frac{y_i}{\hat{y}_i} \cdot (-S(z_i) S(z_j))$$

$$= -y_j + y_j \hat{y}_j + \sum_{i \neq j} y_i \hat{y}_j = -y_j + \sum_i y_i \hat{y}_j = -y_j + \hat{y}_j \times \boxed{\sum_i y_i}$$

$\sum_i y_i = 1$ (It's a probability distribution)

$$= \hat{y}_j - y_j$$

Q.E.D.

Q4. Given $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$ which is $\in \mathbb{R}^{m\times m}$ is a symmetric semi-definite matrix

$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ which is $\in \mathbb{R}^m$

(a) Construct a set of points $\vec{x}_1 \dots \vec{x}_n \in \mathbb{R}^m$ and find the relation between set of points and (μ, Σ) and (μ, Σ) is known. $\vec{x}_1 \dots \vec{x}_n$ are generated identically independent distributed

<sol>

$\because \Sigma$ is a symmetric and semi-definite matrix

$\therefore$ We can do eigen decomposition on it

That is $\Sigma = U\Lambda U^T = [u_1 \dots u_n]\begin{bmatrix}\lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \\ 0 & 0 & \lambda_3 & \\ \vdots & & & \ddots \\ & & & & \lambda_n\end{bmatrix}\begin{bmatrix}u_1 \\ \vdots \\ u_n\end{bmatrix} = \sum_{i=1}^{n} u_i\lambda_i u_i^T$

$= \sum_{i=1}^{n} u_i\sqrt{\lambda_i}\sqrt{\lambda_i}^T u_i^T = \sum_{i=1}^{n}(u_i\sqrt{\lambda_i})(u_i\sqrt{\lambda_i})^T = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$

$\Rightarrow \frac{1}{\sqrt{n}}(x_i-\mu) = u_i\sqrt{\lambda_i} \Rightarrow$ $\boxed{x_i = \mu + \sqrt{n}\,u_i\sqrt{\lambda_i}}$ Q.E.D.

(b). Let $1 \le k \le m$, and objective minimize $\mathrm{Trace}(\Phi^T\Sigma\Phi)$ subject to $\Phi^T\Phi = I_k$, variables $\Phi \in \mathbb{R}^{m\times k}$

<sol> $\mathrm{Trace}\left(\Phi^T\frac{1}{n}XX^T\Phi\right) = \frac{1}{n}\mathrm{Trace}\left(\Phi^TXX^T\Phi\right) = \frac{1}{n}\mathrm{Trace}\left((\Phi^TX)(\Phi^TX)^T\right) = \frac{1}{n}\|\Phi^TX\|_2^F$

which is $\underset{k\times m}{\begin{bmatrix}\phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_m^T\end{bmatrix}}\underset{m\times n}{[x_1 \dots x_n]} = \begin{bmatrix}\phi_1^Tx_1 & \phi_1^Tx_2 & \cdots & \phi_m^Tx_n \\ \vdots & & & \\ \phi_m^Tx_1 & & & \phi_m^Tx_n\end{bmatrix}$

↳ the physical meaning of this column is project x_i to $\phi_1 \dots \phi_m$.

↳ $\phi_1 \ge \phi_2 \ge \phi_3 \ge \cdots \ge \phi_m$ it's a increasing relationship

But now we want to minimize $\mathrm{Trace}(\Phi^T\Sigma\Phi)$ by choosing the value from ϕ_1 to ϕ_N

So, the $\phi_{opt} = [\phi_N, \phi_{N-1} \dots \phi_{N-m+1}]$ which is the opposite concept with PCA

↓ begin from smallest

↳ $\phi_N \le \phi_{N-1} \le \phi_{N-2} \le \cdots \le \phi_{N-m+1}$ which are decreasing relationship

$\Rightarrow$ $\boxed{\frac{1}{n}\sum_{i=1}^{n}\|\Phi^Tx_i\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\text{project } x_i \text{ to span}(\phi_N \dots \phi_{N-m+1})\|}$ which is a problem similar with laplacian eigenmap. *