MLHW

Q1: Generative classification model of K classes $C_1 \dots C_K$ ⇒ prior probabilities $P(C_k) = \pi_k$

⇒ conditional probability $p(x|C_k)$ (input feature vector)

Suppose given training dataset $((\vec{x}_i, \vec{t}_i))_{i=1}^N$, $\vec{t}_i = (t_i^1 \dots t_i^K) \in \{0,1\}^K$

<sol> The probability of one data point

Step 1: Construct $P(\vec{x}, \vec{c}) = P(\vec{x}|\vec{c})P(\vec{c}) = \prod_{k=1}^K \left[P(\vec{x}|C_k)P(C_k)\right] = \prod_{k=1}^K \left[P(x|C_k)\pi_k\right]^{t^k}$

(t^k: determine if we consider $P(\vec{x}|C_k)P(C_k)$ or not)

$L(\theta) = f(x_1, x_2, \dots x_n; \theta)$ ↳ Joint Probability Distribution

$$L(\theta) = \prod_{k=1}^K \left[P(x_1|C_k)\pi_k\right]^{t_1^k} \prod_{k=1}^K \left[P(x_2|C_k)\pi_k\right]^{t_2^k} \cdots \prod_{k=1}^K \left[P(x_N|C_k)\pi_k\right]^{t_N^k}$$

$$= \prod_{i=1}^N \prod_{k=1}^K \left[P(x_i|C_k)\pi_k\right]^{t_i^k}$$

Step 2: Apply log

$$\log(L(\theta)) = \log\Big\{ \left[P(x_1|C_1)\pi_1\right]^{t_1^1} \left[P(x_1|C_2)\pi_2\right]^{t_1^2} \cdots \left[P(x_1|C_K)\pi_K\right]^{t_1^K}$$
$$\left[P(x_2|C_1)\pi_1\right]^{t_2^1} \left[P(x_2|C_2)\pi_2\right]^{t_2^2} \cdots \left[P(x_2|C_K)\pi_K\right]^{t_2^K}$$
$$\vdots \qquad \vdots \qquad \vdots \Big\}$$

$$= t_1^1 \log\left[P(x_1|C_1)\pi_1\right] + t_1^2 \log\left[P(x_1|C_2)\pi_2\right] + \cdots + t_1^K \log\left[P(x_1|C_K)\pi_K\right]$$
$$+ t_2^1 \log\left[P(x_2|C_1)\pi_1\right] + t_2^2 \log\left[P(x_2|C_2)\pi_2\right] + \cdots + t_2^K \log\left[P(x_2|C_K)\pi_K\right]$$
$$\vdots \qquad \vdots \qquad \vdots$$

$$= \sum_{k=1}^K t_1^k \log\left[P(x_1|C_k)\pi_k\right] + \sum_{k=1}^K t_2^k \log\left[P(x_2|C_k)\pi_k\right] + \cdots + \sum_{k=1}^K t_N^k$$

$$= \sum_{i=1}^N \sum_{k=1}^K t_i^k \log\left[P(x_i|C_k)\pi_k\right] = \sum_{i=1}^N \sum_{k=1}^K t_i^k \left[\log P(x_i|C_k) + \log \pi_k\right]$$

($\log\left[P(x_N|C_K)\pi_K\right]$)

Step 3: Derivative and let it to be 0

If we have constraint and want to find maximum or minimum value, we can use Lagrange multipliers.

MLHW Q1 cont.

We can think about it that if we don't add constraint in the expression before. differentiate π_k

$$\Rightarrow \frac{\partial}{\partial \pi_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\left[\log\left(P(x_i|C_k)\right)+\log \pi_k\right]\right)$$

$$= \frac{\partial}{\partial \pi_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} \cancel{t_i^k}^{\,0} \log P(x_i|C_k)\right) + \frac{\partial}{\partial \pi_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log \pi_k\right)$$

$$\frac{\partial}{\partial \pi_k}\begin{pmatrix} t_1^1\log\pi_1 + \boxed{t_1^2\log\pi_2} + \cdots + t_1^K\log\pi_K + \\ t_2^1\log\pi_1 + t_2^2\log\pi_2 + \cdots + t_2^K\log\pi_K + \\ \vdots \\ t_N^1\log\pi_1 + t_N^2\log\pi_2 + \cdots + t_N^K\log\pi_K \end{pmatrix}$$

↓ Reserve to differentiate

If apply any value that $\in \{1,\dots K\}$ assume $k=2$

$$\Rightarrow t_1^2\frac{1}{\pi_2} + t_2^2\frac{1}{\pi_2} + t_3^2\frac{1}{\pi_2} + \cdots t_N^2\frac{1}{\pi_2} = \sum_{i=1}^{N} t_i^2 \cdot \frac{1}{\pi_2}$$

$\Rightarrow$ so if we choose $k \in \{1\dots K\} \Rightarrow \sum_{i=1}^{N} t_i^k \frac{1}{\pi_k}$ where $k \in \{1\dots K\}$

We must let $\frac{\partial \mathcal{L}(\theta)}{\partial \pi_k}$ equals to $0 \Rightarrow \sum_{i=1}^{N} t_i^k \frac{1}{\pi_k} = 0$

✡ **Contraction**:

$\sum_{i=1}^{N} t_i^k = N_k$ (the number of data points belong to C_k) >0 and $\pi_k > 0$

then $\sum_{i=1}^{N} t_i^k \frac{1}{\pi_k}$ must greater than 0 but we want the result of differentiation equals to zero. So, we must add a constraint term.

MLHW Q1 cont.

(Method 1) The constraint is $\sum_{k=1}^{K} \pi_k = 1$ because the summation of all class of probability is 1

$$\Rightarrow \frac{\partial}{\partial \pi_k}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log \pi_k + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\right)$$

$$= \sum_{i=1}^{N} t_i^k \frac{1}{\pi_k} + \lambda = 0 \quad \Rightarrow N_k \cdot \frac{1}{\pi_k} + \lambda = 0 \quad \Rightarrow \pi_k = -\frac{N_k}{\lambda}$$

$$\frac{\partial}{\partial \lambda}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log \pi_k + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\right) = \sum_{k=1}^{K}\pi_k - 1 = 0 \Rightarrow 1 = \sum_{k=1}^{K}\pi_k.$$

$$1 = \sum_{k=1}^{K}\pi_k = \sum_{k=1}^{K} -\frac{N_k}{\lambda} = \frac{-N}{\lambda} \Rightarrow \lambda = -N \Rightarrow \boxed{\pi_k = +\frac{N_k}{N}}$$

Q.E.D.

(Method 2) $f(\pi_k) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log \pi_k$ and $g(\pi_k) = \sum_{k=1}^{K}\pi_k = 1$ (constraint function)

$$f_{\pi_k} = \frac{\partial f(\pi_k)}{\partial \pi_k} = \lambda g_{\pi_k} = \lambda \frac{\partial g(\pi_k)}{\partial \pi_k}$$

$$\Downarrow$$

$$\sum_{i=1}^{N} t_i^k \frac{1}{\pi_k} = \lambda \cdot 1 \Rightarrow N_k \cdot \frac{1}{\pi_k} = \lambda \Rightarrow \pi_k = \frac{N_k}{\lambda}$$

apply to $g(\pi_k) = \sum_{k=1}^{K}\frac{N_k}{\lambda} = 1 = \frac{N}{\lambda} \Rightarrow \lambda = N \Rightarrow \boxed{\pi_k = \frac{N_k}{N}}$ Q.E.D

MLHW 2

Q2 (a) Given $\vec{w} \in \mathbb{R}^m$, $\vec{A} \in \mathbb{R}^{m \times m}$ (they're independent)

$$\frac{\partial w^T A w}{\partial w} = A^T w + A w$$

Theorem 1: $\left(\frac{\partial uv}{\partial \vec{x}}\right)_i = \frac{\partial uv}{\partial x_i} = u\frac{\partial v}{\partial x_i} + v\frac{\partial u}{\partial x_i} = u\left(\frac{\partial v}{\partial \vec{x}}\right)_i + v\left(\frac{\partial u}{\partial \vec{x}}\right)_i$

$$= \left(u\frac{\partial v}{\partial \vec{x}} + v\frac{\partial u}{\partial \vec{x}}\right)_i$$

Theorem 2: Refer to theorem 1

$$\frac{\partial \vec{u}^T\vec{v}}{\partial \vec{x}} = \frac{\partial \sum_k u_k v_k}{\partial \vec{x}} = \sum_k \frac{\partial u_k v_k}{\partial \vec{x}} = \sum_k \left(u_k \frac{\partial v_k}{\partial \vec{x}} + v_k \frac{\partial u_k}{\partial \vec{x}}\right)$$

$$= \sum_k \frac{\partial v_k}{\partial \vec{x}} u_k + \sum_k \frac{\partial u_k}{\partial \vec{x}} v_k = \frac{\partial \vec{v}}{\partial \vec{x}}\vec{u} + \frac{\partial \vec{u}}{\partial \vec{x}}\vec{v}$$

Theorem 3: $\left(\frac{\partial \vec{A}\vec{u}}{\partial \vec{x}}\right)_{i,j} = \frac{\partial \sum_k a_{jk} u_k}{\partial x_i} = \sum_k a_{jk}\frac{\partial u_k}{\partial x_i}$

$$= \sum_k \left(\frac{\partial \vec{u}}{\partial \vec{x}}\right)_{ik} \left(\vec{A}^T\right)_{kj} = \left(\frac{\partial \vec{u}}{\partial \vec{x}}\vec{A}^T\right)_{ij}$$

Theorem 4: Refer to theorem 2 and theorem 3

$$\frac{\partial \vec{u}^T\vec{A}\vec{v}}{\partial \vec{x}} = \frac{\partial \vec{u}^T(\vec{A}\vec{x})}{\partial \vec{x}} = \frac{\partial \vec{u}}{\partial \vec{x}}\vec{A}\vec{x} + \frac{\partial(\vec{A}\vec{v})}{\partial \vec{x}}\vec{u} = \frac{\partial \vec{u}}{\partial \vec{x}}\vec{A}\vec{v} + \frac{\partial \vec{v}}{\partial \vec{x}}\vec{A}^T\vec{u}$$

Theorem 5: $\frac{\partial \vec{x}}{\partial \vec{x}} = \vec{I}$

Refer to theorem 4 and theorem 5

$$\frac{\partial \vec{x}^T\vec{A}\vec{x}}{\partial \vec{x}} = \frac{\partial \vec{x}}{\partial \vec{x}}\vec{A}\vec{x} + \frac{\partial \vec{x}}{\partial \vec{x}}\vec{A}^T\vec{x} = \vec{I}\vec{A}\vec{x} + \vec{I}\vec{A}^T\vec{x} = \boxed{(\vec{A}+\vec{A}^T)\vec{x}}$$ Q.E.D

Q2(b) Given $\vec{A} \cdot \vec{B} \in \mathbb{R}^{m \times m}$. Show that $\frac{\partial tr(\vec{A}\vec{B})}{\partial a_{ij}} = b_{ji}$

(Method 1)

$$\vec{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & \ddots & & \\ \vdots & & & \\ a_{m1} & & & a_{mm} \end{bmatrix} \quad \vec{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & \ddots & & \\ \vdots & & & \\ b_{m1} & & & b_{mm} \end{bmatrix}$$

$$\vec{A}\vec{B} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + \cdots a_{1m}b_{m1} & a_{11}b_{12} + a_{12}b_{22} + \cdots a_{1m}b_{m2} & \cdots & a_{11}b_{1m} + \cdots a_{1m}b_{mm} \\ a_{21}b_{11} + a_{22}b_{21} + \cdots a_{2m}b_{m1} & \ddots & & \\ \vdots & & & \\ a_{m1}b_{11} + a_{m2}b_{21} + \cdots a_{mm}b_{m1} & \cdots & & a_{m1}b_{1m} + \cdots a_{mm}b_{mm} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_i a_{1i}b_{i1} & \sum_i a_{1i}b_{i2} & \cdots & \sum_i a_{1i}b_{im} \\ \sum_i a_{2i}b_{i1} & \ddots & & \\ \vdots & & & \\ \sum_i a_{mi}b_{i1} & \cdots & & \sum_i a_{mi}b_{im} \end{bmatrix}$$

$$Trace(\vec{A}\vec{B}) = \sum_i a_{1i}b_{i1} + \sum_i a_{2i}b_{i2} + \sum_i a_{3i}b_{i3} + \cdots \sum_i a_{mi}b_{im}$$

$$= \sum_k \sum_i a_{ki}b_{ik}$$

$$\frac{\partial tr(\vec{A}\vec{B})}{\partial a_{ij}} = \frac{\partial \sum_k \sum_i a_{ki}b_{ik}}{\partial a_{ij}} = \sum_k \sum_i b_{ik} \frac{\partial a_{ki}}{\partial a_{ij}} = \sum_k \sum_i b_{ik} \boxed{\delta_{ik}\delta_{ji}} = b_{ji}$$

$$\delta_{ik}\delta_{ji} = \begin{cases} 1, & \text{if } k=i, i=j \\ 0, & \text{otherwise} \end{cases}$$

Q.E.D

(Method 2) Refer to Q2.(a)

$$\frac{\partial tr(\vec{A}\vec{B})}{\partial a_{ij}} = tr\left(\vec{B}\frac{\partial \vec{A}}{\partial a_{ij}}\right) = tr\left(\vec{B}\vec{e}_i\vec{e}_j^T\right) = tr\left(\vec{e}_j^T\vec{B}\vec{e}_i\right) = \boxed{b_{ji}}$$

Q.E.D

$\vec{e}_i$ is represent i-th element is 1, others element is 0

$e_i e_j^T$ is represent (i, j) element has value 1, the others is 0

Q2 (C) Show that $\frac{\partial \log(\det(\Sigma))}{\partial \sigma_{ij}} = e_j^T \Sigma^{-1} e_i$ where $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ \sigma_{m1} & & & \sigma_{mm} \end{bmatrix} \in \mathbb{R}^{m \times m}$ is a non-singular covariance matrix and $\vec{e}_j$ is the unit vector along the j-th axis (e.g. $\vec{e}_3 = [0,0,1,0,\ldots,0]^T$)

Given $|\vec{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\vec{A}_{ij}| = \sum_{i=1}^{m} (-1)^{i+j} a_{ij} |\vec{A}_{ij}|$

Cramer Rule $\vec{A}\vec{x} = \vec{e}_i = [0 \ldots \underset{\downarrow}{1} \ldots 0]^T \Rightarrow \vec{x} = \vec{A}^{-1}\vec{e}_i$. ($i$-th element)

$$x^{(j)} = e_j^T \vec{A}^{-1} e_i = \frac{(-1)^{i+j} |A_{ij}|}{|\vec{A}|} = \frac{\partial \log |\vec{A}|}{\partial a_{ij}}$$

Proof: $\frac{\partial \log(|\Sigma|)}{\partial \sigma_{ij}} = \frac{1}{|\Sigma|} \times \frac{\partial |\vec{\Sigma}|}{\partial \sigma_{ij}}$

Refer to the given theorem $\Rightarrow \frac{\partial |\vec{\Sigma}|}{\partial \sigma_{ij}} = (-1)^{i+j} |\Sigma_{ij}|$

$$\Rightarrow \frac{\partial \log(|\Sigma|)}{\partial \sigma_{ij}} = \frac{(-1)^{i+j} |\Sigma_{ij}|}{|\vec{\Sigma}|} = \boxed{e_j^T \vec{\Sigma}^{-1} e_i}$$ Q.E.D.

MLHW

Q3: Given conditional probability density function are given by Gaussian Distribution with a shared covariance matrix, namely $p(\vec{x}|C_k) = \mathcal{N}(\vec{x}|\mu_k, \Sigma)$

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Step 1: Construct $L(\theta)$

$$P(\vec{x}, \vec{C}_k) = P(\vec{x}|\vec{C})P(\vec{C}) = \prod_{k=1}^{K}\left[P(\vec{x}|C_k)P(C_k)\right]^{t^k} = \prod_{k=1}^{K}\left[P(\vec{x}|C_k)\pi_k\right]^{t^k}$$

$$= \prod_{i=1}^{N}\prod_{k=1}^{K}\left[P(x_i|C_k)\pi_k\right]^{t_i^k} = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\mathcal{N}(\vec{x}_i|\vec{\mu}_k, \Sigma)\pi_k\right]^{t_i^k}$$

$$= \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\frac{1}{(2\pi)^{p/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right\}\right]^{t_i^k}$$

Step 2: Apply log to $L(\theta)$

$$\log(L(\theta)) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\left[\log\left(\mathcal{N}(\vec{x}_i|\mu_k, \Sigma)\right) + \log\pi_k\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\left[\log\left(\frac{1}{\sqrt{2\pi}\sqrt{|\Sigma|}}\right) + -\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k) + \log\pi_k\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\log\left(\frac{1}{\sqrt{2\pi}\sqrt{|\Sigma|}}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right]t_i^k + \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\log\pi_k$$

Step 3: Differentiate

$$\frac{\partial\log(L(\theta))}{\partial\mu_k} = \frac{\partial}{\partial\mu_k}\left[-\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right]t_i^k\right]$$

$$\Rightarrow \frac{\partial}{\partial\mu_k}\left[\sum_{i=1}^{N} -\frac{t_i^k}{2}\sum_{k=1}^{K}(x_i^T-\mu_k^T)\Sigma^{-1}(x_i-\mu_k)\right] = \frac{\partial}{\partial\mu_k}\left[\sum_{i=1}^{N} -\frac{t_i^k}{2}\sum_{k=1}^{K}\left(x_i^T\Sigma^{-1}x_i - \mu_k^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k\right)\right]$$

$$= \sum_{i=1}^{N}\frac{-t_i^k}{2}\left(\frac{\partial}{\partial\mu_k}\left(x_i^T\Sigma^{-1}x_i - \mu_k^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k\right)\right)$$

$$= \sum_{i=1}^{N}\frac{-t_i^k}{2}\left(0 - \Sigma^{-1}x_i - (x_i^T\Sigma^{-1})^T + \left(\Sigma^{-1} + (\Sigma^{-1})^T\right)\mu_k\right) = \sum_{i=1}^{N}\frac{-t_i^k}{2}\left[-\Sigma^{-1}x_i - (\Sigma^{-1})^Tx_i + 2\Sigma^{-1}\mu_k\right]$$

$$= \sum_{i=1}^{N}\frac{-t_i^k}{2}\left[-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu_k\right] = \sum_{i=1}^{N} t_i^k\left[\Sigma^{-1}x_i - \Sigma^{-1}\mu_k\right] = \sum_{i=1}^{N} t_i^k\Sigma^{-1}(x_i-\mu_k) = 0$$

$$\sum_{i=1}^{N} t_i^k\Sigma^{-1}(x_i-\mu_k) = 0 = \sum_{i=1}^{N} t_i^k x_i - N_k\mu_k \Rightarrow \boxed{\mu_k = \frac{1}{N_k}\sum_{i=1}^{N} t_i^k x_i}$$

* Σ to eliminate Σ^{-1}

** Q.E.D.

MLHW Q3 cont.

Q3(b) $\log(L(\theta)) = \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log\left(\frac{1}{\sqrt{2\pi}\sqrt{|\Sigma|}}\right) - \sum_{i=1}^{N}\sum_{k=1}^{K}\left[\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right]t_i^k + \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log\pi_k$

① $\frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k\left(-\frac{D}{2}\log 2\pi + -\frac{1}{2}\log(\det(\Sigma))\right)\right]$ ② $- \frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\frac{1}{2}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)\right]t_i^k\right]$

① $-\frac{1}{2}\frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \log(\det(\Sigma))\right] = -\frac{1}{2}(\Sigma^{-1})^T \cdot \sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k$

② $-\frac{1}{2}\frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\underbrace{(x_i^T-\mu_k^T)\Sigma^{-1}(x_i-\mu_k)}_{\text{scalar}} t_i^k\right] = -\frac{1}{2}\frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\left(x_i^T\Sigma^{-1}x_i - \mu_k^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k\right)t_i^k\right]$

$= -\frac{1}{2}\frac{\partial}{\partial\Sigma}\left[\sum_{i=1}^{N}\sum_{k=1}^{K}\text{trace}\left((x_i^T-\mu_k^T)\Sigma^{-1}(x_i-\mu_k)\right)t_i^k\right]$

$= +\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \left(\Sigma^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma^{-1}\right)^T\right]$

$= +\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \Sigma^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\Sigma^{-1}\right]$

① + ② = 0

$-\frac{1}{2}\vec{\Sigma}^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k + \frac{1}{2}\left[\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k \vec{\Sigma}^{-1}(x_i-\mu_k)(x_i-\mu_k)^T\vec{\Sigma}^{-1}\right] = 0$

⇓ multiply Σ in 2 times

$-\frac{1}{2}\vec{\Sigma}\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k + \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k (x_i-\mu_k)(x_i-\mu_k)^T = 0$

$\Rightarrow \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k (x_i-\mu_k)(x_i-\mu_k)^T = \frac{1}{2}\vec{\Sigma}\cdot N \Rightarrow \vec{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} t_i^k (x_i-\mu_k)(x_i-\mu_k)^T$

$= \sum_{k=1}^{K}\frac{N_k}{N}\cdot\frac{1}{N_k}\sum_{i=1}^{N} t_i^k (x_i-\mu_k)(x_i-\mu_k)^T$

$= \sum_{k=1}^{K}\frac{N_k}{N}\vec{\Sigma}_k$ where

$\vec{\Sigma}_k = \frac{1}{N_k}\sum_{i=1}^{N} t_i^k (x_i-\mu_k)(x_i-\mu_k)^T$

Q.E.D.

MLHW

Q4. Given n points $x_i \in \mathbb{R}^d$, $i \in \{1 \dots n\}$ / k cluster center $\mu = (\mu_1 \dots \mu_k)$

Objective $\arg\min L(C, \mu) = \sum_{i=1}^{n} \| x_i - \mu_{C(i)} \|_2^2 = \sum_{q=1}^{k} \sum_{i:C(i)=q} \| x_i - \mu_q \|_2^2$

Get the distance between i-th data and the center that i-th data belongs to the cluster q

(a.) $\sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2 \le \sum_{i=1}^{m} \| z_i - z \|_2^2$ where $\bar{z} = \frac{1}{m} \sum_{i=1}^{m} z_i$, points $z_1 \dots z_m$, z is an arbitrary point in same space

(Solved) Expand the expression

$$\sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2 = (z_1 - \bar{z})^T (z_1 - \bar{z}) + (z_2 - \bar{z})^T (z_2 - \bar{z}) + \dots + (z_m - \bar{z})^T (z_m - \bar{z})$$

$$= (z_1^T z_1 - 2\bar{z}^T z_1 + \bar{z}^T \bar{z}) + (z_2^T z_2 - 2\bar{z}^T z_2 + \bar{z}^T \bar{z}) + \dots + (z_m^T z_m - 2\bar{z}^T z_m + \bar{z}^T \bar{z})$$

$$= \sum_{i=1}^{m} \| z_i \|_2^2 - 2\bar{z}^T \sum_{i=1}^{m} z_i + m \| \bar{z} \|_2^2$$

$$\sum_{i=1}^{m} \| z_i - z \|_2^2 = (z_1 - z)^T (z_1 - z) + (z_2 - z)^T (z_2 - z) + \dots + (z_m - z)^T (z_m - z)$$

$$= \sum_{i=1}^{m} \| z_i \|_2^2 - 2z^T \sum_{i=1}^{m} z_i + m \| z \|_2^2$$

$$\sum_{i=1}^{m} \| z_i - z \|_2^2 - \sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2 = m \| z \|_2^2 - 2z^T \sum_{i=1}^{m} z_i - m \| \bar{z} \|_2^2 + 2\bar{z}^T \sum_{i=1}^{m} z_i$$

$$= m \| z \|_2^2 - m \| \bar{z} \|_2^2 - \boxed{2(z^T - \bar{z}^T) \sum_{i=1}^{m} z_i} \rightarrow 2(z^T - \bar{z}^T) m \cdot \frac{1}{m} \sum_{i=1}^{m} z_i$$

$$= 2(z^T - \bar{z}^T) m \cdot \bar{z}$$

$$= m \| z \|_2^2 - m \| \bar{z} \|_2^2 - 2m z^T \bar{z} + 2m \| \bar{z} \|_2^2$$

$$= m \| z \|_2^2 + m \| \bar{z} \|_2^2 - 2m z^T \bar{z} = m (z^T z - 2z^T \bar{z} + \bar{z}^T \bar{z}) = m \| z - \bar{z} \|_2^2$$

$\because m > 0$ and $\| z - \bar{z} \|_2^2 \ge 0$

$\therefore \sum_{i=1}^{m} \| z_i - z \|_2^2 - \sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2 \ge 0 \Rightarrow \boxed{\sum_{i=1}^{m} \| z_i - z \|_2^2 \ge \sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2}$

Q.E.D.

(b) Show that $L(C^{t+1}, \mu^t) < L(C^t, \mu^t)$, that is fix every class center and use the data belongs to new cluster to compute the distance with the center

Given the original loss function $L(C^t, \mu^t) = \sum_{i=1}^{n} \| x_i - \mu_{C(i)} \|_2^2$

(proof) Let C^{t+1} be the new assignment and C^t is the assignment from the previous iteration $\Rightarrow$ $C^{t+1} \in \arg\min_{j=\{1 \dots k\}} \| x_i - \mu_j \|_2^2$

means chose the closest cluster center as x_i's cluster

$$L(C^{t+1}, \mu^t) - L(C^t, \mu^t) = \sum_{i=1}^{n} \| x_i - \mu_{C^{t+1}(i)} \|_2^2 - \sum_{i=1}^{n} \| x_i - \mu_{C^t(i)} \|_2^2$$

$$= \sum_{i=1}^{n} \left(\| x_i - \mu_{C^{t+1}(i)} \|_2^2 - \| x_i - \mu_{C^t(i)} \|_2^2 \right)$$

because $\mu_{C^{t+1}(i)}$ is the closest to x_i, so the distance

$\| x_i - \mu_{C^{t+1}(i)} \|_2^2$ is smaller than $\| x_i - \mu_{C^t(i)} \|_2^2$

$\Rightarrow L(C^{t+1}, \mu^t) - L(C^t, \mu^t) < 0$ Q.E.D.

(c) Use the result from (a). Fix the cluster assignment C^{t+1} and compute the new cluster center μ^{t+1} from all the data in cluster.

Given: $L(C^t, \mu^t) = \sum_{q=1}^{k} \sum_{i : C(i) = q} \| x_i - \mu_q \|_2^2$

(proof) Consider j-th cluster, μ_j^t is previous cluster center and μ_j^{t+1} is new one.

$$\mu_j^{t+1} = \frac{1}{|\{ i : C(i) = j \}|} \sum_{i : C(i) = j} x_i$$

↑ the # of data in cluster j

$$L(C^{t+1}, \mu^{t+1}) - L(C^{t+1}, \mu^t) = \sum_{q=1}^{k} \sum_{i : C(i) = q} \| x_i - \mu_q^{t+1} \|_2^2 - \sum_{q=1}^{k} \sum_{i : C(i) = q} \| x_i - \mu_q^t \|_2^2$$

use the result from (a) $\sum_{i=1}^{m} \| z_i - \bar{z} \|_2^2 - \sum_{i=1}^{m} \| z_i - \bar{\bar{z}} \|_2^2 \geq 0$

Then $L(C^{t+1}, \mu^{t+1}) - L(C^{t+1}, \mu^t) < 0$

Q.E.D.

MLHW Q4 cont.

(d). Use the result that $L(C^{t+1}, \mu^t) \le L(C^t, \mu^t)$ and $L(C^{t+1}, \mu^{t+1}) \le L(C^{t+1}, \mu^t)$

so $L(C^{t+1}, \mu^{t+1}) \le L(C^t, \mu^t)$

(i) use completeness axiom

$\because L(C^t, \mu^t) \ge L(C^{t+1}, \mu^{t+1})$ and $L(C^t, \mu^t) = \sum_{i=1}^{n} \| x_i - \mu_{C(i)} \|_2^2 \ge 0$

$\therefore \exists k$ so that $L(C^t, \mu^t) \ge k \quad \forall t$

implies $\lim_{t \to \infty} L(C^t, \mu^t)$ exist (in real number)

(ii) use approximation theorem

Given $L(C^t, \mu^t)$ has a finite infimum, If $\varepsilon > 0$ then there exist $a \in L(C^t, \mu^t)$

such that $\inf L \le a < \inf L + \varepsilon$

(iii) use montone convergence theorem

because of the theorem and axiom mentioned above,

if $L(C^t, \mu^t)$ is decreasing and bounded below then $\{L\}$ converges to a finite limit

Q.E.D.

(e) The algorithm will stop when no change in loss function occurs during the assignment step. That is, the result of loss function reach the lower bound and won't update the cluster center and loss is a fixable value.

$L(C^t, \mu^t) - L(C^{t+1}, \mu^{t+1}) = \varepsilon_t > 0 \quad \forall t$ and the lower bound is 0 and the initial loss is finite value. that greater than 0

$\Rightarrow$ Repeat: $L(C^t, \mu^t) - \varepsilon_t$ for n times $\Rightarrow$ implies n is a finite step.

($L(C^t, \mu^t)$: >0 and finite; ε_t: >0)

Q.E.D.