Machine Learning - Homework 4

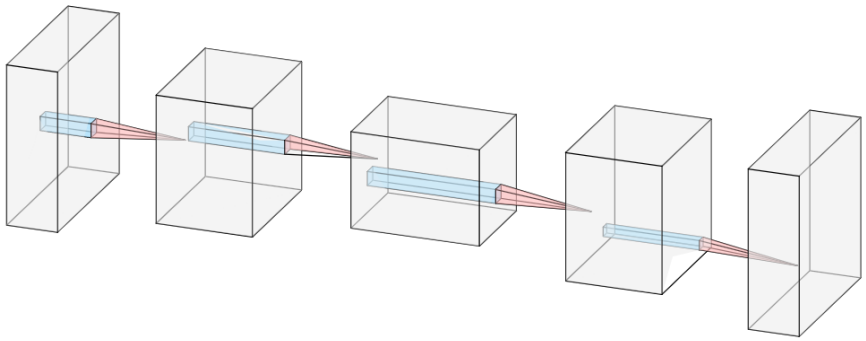
資工四 B05902023 李澤諺

November 22, 2019

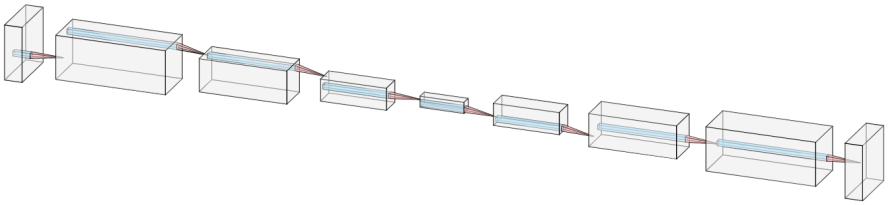
Part 1. Programming Problem

1. (1%) 請使用不同的 Autoencoder model，以及不同的降維方式 (降到不同維度)，討論其 reconstruction loss 和 public/private accuracy。(因此模型需要兩種，降維方法也需要兩種，但 clustering 不用兩種。)

以下爲我於本題中所比較的兩個 Autoencoder 的架構：



Autoencoder 1	
Encoder	Conv2d(3 , 8 , kernel_size = (3 , 3) , stride = (2 , 2) , padding = (1 , 1))
	Conv2d(8 , 16 , kernel_size = (3 , 3) , stride = (2 , 2) , padding = (1 , 1))
Decoder	ConvTranspose2d(16 , 8 , kernel_size = (2 , 2) , stride = (2 , 2))
	ConvTranspose2d(8 , 3 , kernel_size = (2 , 2) , stride = (2 , 2))
	Tanh()



Autoencoder 2	
Encoder	Conv2d(3 , 1024 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxPool2d(2 , return_indices = True)
	Conv2d(1024 , 256 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxPool2d(2 , return_indices = True)
	Conv2d(256 , 64 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxPool2d(2 , return_indices = True)
	Conv2d(64 , 16 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxPool2d(2 , return_indices = True)
Decoder	MaxUnpool2d(2)
	ConvTranspose2d(16 , 64 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxUnpool2d(2)
	ConvTranspose2d(64 , 256 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxUnpool2d(2)
	ConvTranspose2d(256 , 1024 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	MaxUnpool2d(2)
	ConvTranspose2d(1024 , 3 , kernel_size = (3 , 3) , stride = (1 , 1) , padding = (1 , 1))
	Tanh()

兩個 Autoencoder 皆爲使用 Adam 訓練，learning rate 爲 0.0001，使用 l1-loss 作爲 loss function，batch size 爲 256，訓練了 20 個 epoch。

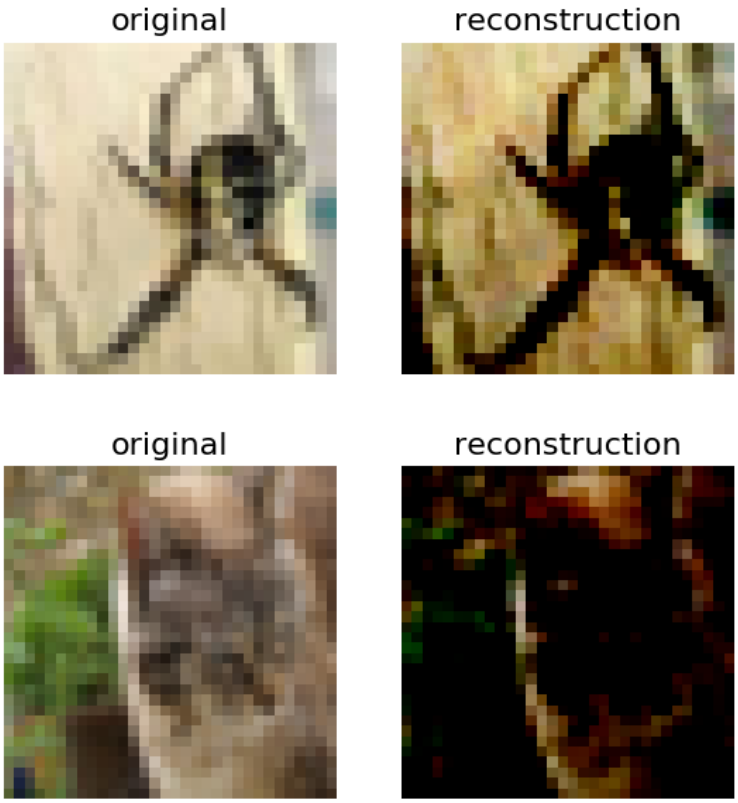
Autoencoder 1 會將圖片降到 1024 維，而 Autoencoder 2 會將圖片降到 64 維，在兩個 Autoencoder 分別做完第一次的降維之後，接著皆會使用 t-SNE 進行第二次的降維，將圖片降到 2 維，最後使用 K-Means 進行 clustering。

以下爲兩種方法分別所得到的 average reconstruction loss 以及 public/private accuracy：

	reconstruction loss	public accuracy	private accuracy
Autoencoder 1	0.00084342	0.69259	0.70063
Autoencoder 2	0.00031634	0.75333	0.75920

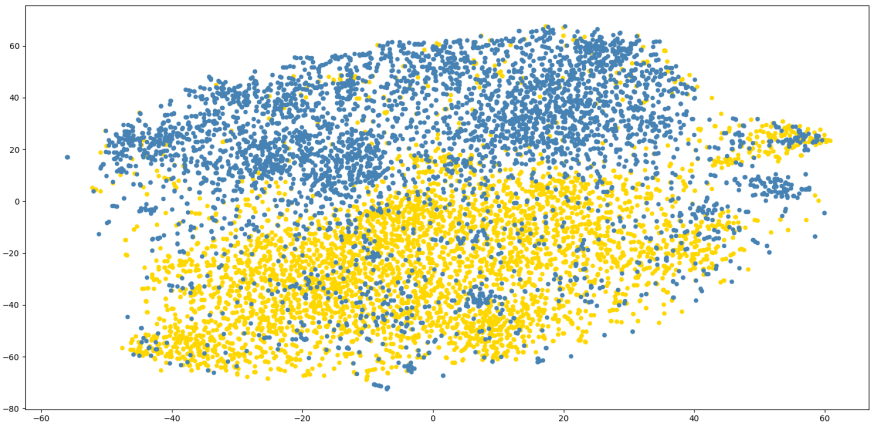
2. (1%) 從 dataset 選出 2 張圖，並貼上原圖以及經過 autoencoder 後 reconstruct 的圖片。

以下爲從 training dataset 中取出前 2 張圖，使用第 1 題訓練出來的 Autoencoder 2，將圖片進行降維並 reconstruct 之後的結果：



3. (1%) 在之後我們會給你 dataset 的 label。請在二維平面上視覺化 label 的分佈。

以下爲使用第 1 題訓練出來的 Autoencoder 2 以及 t-SNE，將圖片進行降維之後的結果：



Part 2. Math Problem

In this problem set, we denote $[\![a, b]\!] = \{i \in \mathbb{Z} : a \leq i \leq b\}$.

1. Principle Component Analysis (1%)

Given 10 samples in 3D: $(1, 2, 3), (4, 8, 5), (3, 12, 9), (1, 8, 5), (5, 14, 2), (7, 4, 1), (9, 8, 9), (3, 8, 1), (11, 5, 6), (10, 11, 7)$.

(a) What are the principal axes?
(b) Please compute the principal components for each sample.
(c) What is the average reconstruction error if reduce dimension to 2D? Here the reconstruction error is defined as the squared loss.

solution

(a)

令題幹中給定的點依序爲 x_1、x_2、$\cdots$、$x_{10} \in \mathbb{R}^3$。
x_1、x_2、$\cdots$、x_{10} 的 mean 和 covariance matrix 分別爲

$$\mu = \frac{1}{10}\sum_{n=1}^{10} x_n = \begin{pmatrix} 5.4 \\ 8 \\ 4.8 \end{pmatrix}$$

$$\Sigma = \frac{1}{10}\sum_{n=1}^{10}(x_n - \mu)(x_n - \mu)^T = \begin{pmatrix} 12.04 & 0.5 & 3.28 \\ 0.5 & 12.2 & 2.9 \\ 3.28 & 2.9 & 8.16 \end{pmatrix}$$

將 Σ 正交對角化爲 $\Sigma = Q\Lambda Q^T$，其中

$$Q = \begin{pmatrix} 0.616596 & 0.678179 & -0.399856 \\ 0.58815 & -0.73439 & -0.337589 \\ 0.522596 & 0.0272856 & 0.852144 \end{pmatrix}$$

$$\Lambda = \begin{pmatrix} 15.2974 & 0 & 0 \\ 0 & 11.6305 & 0 \\ 0 & 0 & 5.47203 \end{pmatrix}$$

因此可得 Σ 的 eigenvector 爲 (依照 eigenvalue 的大小排序)

$$v_1 = \begin{pmatrix} 0.616596 \\ 0.58815 \\ 0.522596 \end{pmatrix} \quad v_2 = \begin{pmatrix} 0.678179 \\ -0.73439 \\ 0.0272856 \end{pmatrix} \quad v_3 = \begin{pmatrix} -0.399856 \\ -0.337589 \\ 0.852144 \end{pmatrix}$$

此即爲 principal axis。 □

(b)

令

$$W = \begin{pmatrix} v_1^T \\ v_2^T \\ v_3^T \end{pmatrix} = \begin{pmatrix} 0.616596 & 0.58815 & 0.522596 \\ 0.678179 & -0.73439 & 0.0272856 \\ -0.399856 & -0.337589 & 0.852144 \end{pmatrix}$$

則 x_1、x_2、$\cdots$、x_{10} 的 principal component 依序爲

$$Wx_1 = \begin{pmatrix} 3.360684 \\ -0.7087442 \\ 1.481398 \end{pmatrix} \quad Wx_2 = \begin{pmatrix} 9.784564 \\ -3.025976 \\ -0.039416 \end{pmatrix}$$

$$Wx_3 = \begin{pmatrix} 13.610952 \\ -6.5365726 \\ 2.41866 \end{pmatrix} \quad Wx_4 = \begin{pmatrix} 7.934776 \\ -5.060513 \\ 1.160152 \end{pmatrix}$$

$$Wx_5 = \begin{pmatrix} 12.362272 \\ -6.8359938 \\ -5.021238 \end{pmatrix} \quad Wx_6 = \begin{pmatrix} 7.191368 \\ 1.8369786 \\ -3.297204 \end{pmatrix}$$

$$Wx_7 = \begin{pmatrix} 14.957928 \\ 0.4740614 \\ 1.36988 \end{pmatrix} \quad Wx_8 = \begin{pmatrix} 7.077584 \\ -3.8132974 \\ -3.048136 \end{pmatrix}$$

$$Wx_9 = \begin{pmatrix} 12.858882 \\ 3.9517326 \\ -0.973497 \end{pmatrix} \quad Wx_{10} = \begin{pmatrix} 16.293782 \\ -1.105508 \\ -1.747031 \end{pmatrix} \quad \square$$

(c)

令

$$\tilde{W} = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = \begin{pmatrix} 0.616596 & 0.58815 & 0.522596 \\ 0.678179 & -0.73439 & 0.0272856 \end{pmatrix}$$

則 average reconstruction error 爲

$$\frac{1}{10}\sum_{n=1}^{10} \|x_n - \tilde{W}^T(\tilde{W}x_n)\|^2 = 6.0681663 \ \square$$

2. Constrained Mahalanobis Distance Minimization Problem (1%)

(a) (0.25%) Let $A \in \mathbb{R}^{m\times n}$, show that AA^T and A^TA are both symmetric and positive semi-definite, and share the same non-zero eigenvalues.

(b) (0.25%) Let $\Sigma \in \mathbb{R}^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in \mathbb{R}^m$. Please construct a set of points $x_1, \cdots, x_n \in \mathbb{R}^m$ such that

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T = \Sigma, \ \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$$

(c) (0.5%) Let $1 \le k \le m$, solve the following optimization problem (and justify with proof):

$$\begin{aligned} &minimize\ Trace(\Phi^T\Sigma\Phi) \\ &subject\ to\ \Phi^T\Phi = I_k \\ &variables\ \Phi \in \mathbb{R}^{m\times k} \end{aligned}$$

solution

(a)

因爲

$$(AA^T)^T = (A^T)^TA^T = AA^T$$
$$(A^TA)^T = A^T(A^T)^T = A^TA$$

所以 AA^T 和 A^TA 皆爲 symmetric。
因爲 $\forall\ x \in \mathbb{R}^m$ 且 $x \neq 0$，以及 $\forall\ y \in \mathbb{R}^n$ 且 $y \neq 0$，皆有

$$x^T(AA^T)x = (x^TA)(A^Tx) = (A^Tx)^T(A^Tx) = \|A^Tx\|^2 \ge 0$$
$$y^T(A^TA)y = (y^TA^T)(Ay) = (Ay)^T(Ay) = \|Ay\|^2 \ge 0$$

所以 AA^T 和 A^TA 皆爲 positive semi-definite。
令 $\lambda \neq 0$ 爲 AA^T 的一個 eigenvalue。
取 $v \in \mathbb{R}^m$ 爲其對應的一個 eigenvector，則有

$$(AA^T)v = \lambda v$$

因此可得

$$(A^TA)(A^Tv) = A^T((AA^T)v) = A^T(\lambda v) = \lambda(A^Tv)$$

故 λ 亦爲 A^TA 的一個 eigenvalue，而 A^Tv 爲其對應的一個 eigenvector。
同理，令 $\mu \neq 0$ 爲 A^TA 的一個 eigenvalue。
取 $u \in \mathbb{R}^n$ 爲其對應的一個 eigenvector，則有

$$(A^TA)u = \mu u$$

因此可得

$$(AA^T)(Av) = A((A^TA)u) = A(\mu u) = \mu(Au)$$

故 μ 亦爲 AA^T 的一個 eigenvalue，而 Au 爲其對應的一個 eigenvector。
由上述可得，AA^T 和 A^TA 有相同的 non-zero eigenvalue。 □

(b)

首先，取 z_1、z_2、$\cdots$、$z_{2m} \in \mathbb{R}^m$ 依序爲

$$\begin{pmatrix} \sqrt{m} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} -\sqrt{m} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{m} \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ -\sqrt{m} \\ \vdots \\ 0 \end{pmatrix} \cdots \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \sqrt{m} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -\sqrt{m} \end{pmatrix}$$

則 z_1、z_2、$\cdots$、z_{2m} 的 mean 爲

$$\frac{1}{2m}\sum_{k=1}^{2m} z_k = 0$$

而 covariance matrix 爲

$$\frac{1}{2m}\sum_{k=1}^{2m}(z_k - 0)(z_k - 0)^T = \frac{1}{2m}\sum_{k=1}^{2m} z_k z_k^T = I_m$$

接著，因爲 $\Sigma \in \mathbb{R}^{m\times m}$ 爲 positive semi-definite，所以 $\exists\ A \in \mathbb{R}^{m\times m}$ ，使得 $\Sigma = AA^T$ (例如可取 $\Sigma = AA^T$ 爲 Σ 的 Cholesky decomposition)。
取 $x_k = Az_k + \mu$，即可得到 x_1、x_2、$\cdots$、x_{2m} 的 mean 爲

$$\frac{1}{2m}\sum_{k=1}^{2m}(Az_k + \mu) = A(\frac{1}{2m}\sum_{k=1}^{2m} z_k) + \mu = A \cdot 0 + \mu = \mu$$

而 covariance matrix 爲

$$\begin{aligned}
&\frac{1}{2m}\sum_{k=1}^{2m}(x_k-\mu)(x_k-\mu)^T \\
=&\frac{1}{2m}\sum_{k=1}^{2m}((Az_k+\mu)-\mu)((Az_k+\mu)-\mu)^T \\
=&\frac{1}{2m}\sum_{k=1}^{2m}(Az_k)(Az_k)^T=\frac{1}{2m}\sum_{k=1}^{2m}(Az_kz_k^TA^T) \\
=&A(\frac{1}{2m}\sum_{k=1}^{2m}z_kz_k^T)A^T=A\cdot I_m\cdot A^T=AA^T=\Sigma\ \square
\end{aligned}$$

(c)

因爲 Σ 和 $\Phi\Phi^T$ 皆爲 symmetric，所以 Σ 和 $\Phi\Phi^T$ 皆可以正交對角化。
令 Σ 的 eigenvalue 爲 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m$，而 $\Phi\Phi^T$ 的 eigenvalue 爲 $\mu_1\geq\mu_2\geq\cdots\geq\mu_m$。
事實上，若 $\Phi=(v_1\ v_2\ \cdots\ v_k)$，因爲 $\Phi^T\Phi=I_k$，所以 v_1、v_2、$\cdots$、v_k 爲 $\mathbb{R}^m$ 中的一組 orthonormal vectors。
取 w_1、w_2、$\cdots$、$w_{m-k}\in\mathbb{R}^m$，使得 v_1、v_2、$\cdots$、v_k、w_1、w_2、$\cdots$、w_{m-k} 構成 $\mathbb{R}^m$ 的一組 orthonormal basis。
因爲

$$(\Phi\Phi^T)v_i=\Phi(\Phi^Tv_i)=\Phi\cdot e_i=v_i=1\cdot v_i$$

所以 v_i 皆爲 $\Phi\Phi^T$ 的 eigenvector，且其對應的 eigenvalue 爲 1 (因此 1 的 multiplicity 爲 k)。
因爲

$$(\Phi\Phi^T)w_i=\Phi(\Phi^Tw_i)=\Phi\cdot 0=0=0\cdot w_i$$

所以 w_i 皆爲 $\Phi\Phi^T$ 的 eigenvector，且其對應的 eigenvalue 爲 0 (因此 0 的 multiplicity 爲 $m-k$)。
由此可得 $\Phi\Phi^T$ 的 eigenvalue 爲 $\mu_1=\mu_2=\cdots=\mu_k=1$，$\mu_{k+1}=\mu_{k+2}=\cdots=\mu_m=0$。
所以

$$\begin{aligned}
Trace(\Phi^T\Sigma\Phi)&=Trace(\Phi^T(\Sigma\Phi))=Trace((\Sigma\Phi)\Phi^T)=Trace(\Sigma(\Phi\Phi^T)) \\
&\geq\sum_{i=1}^{m}\lambda_i\mu_{m-i+1}\ (\textit{Von Neumann's Trace Inequality}) \\
&=\lambda_1\mu_m+\lambda_2\mu_{m-1}+\cdots+\lambda_{m-k}\mu_{k+1}+ \\
&\quad\lambda_{m-k+1}\mu_k+\lambda_{m-k+2}\mu_{k-1}+\cdots+\lambda_m\mu_1 \\
&=\lambda_1\cdot 0+\lambda_2\cdot 0+\cdots+\lambda_{m-k}\cdot 0+ \\
&\quad\lambda_{m-k+1}\cdot 1+\lambda_{m-k+2}\cdot 1+\cdots+\lambda_m\cdot 1 \\
&=\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m
\end{aligned}$$

因此可得 $\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m$，即 Σ 最小的 k 個 eigenvalue 之和，爲 $Trace(\Phi^T\Sigma\Phi)$ 的一個 lower bound。
接著，只要證明 $\exists\ \Phi\in\mathbb{R}^{m\times k}$ 且 $\Phi^T\Phi=I_k$，使得 $Trace(\Phi^T\Sigma\Phi)=\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m$，即可得到 $\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m$ 爲 $Trace(\Phi^T\Sigma\Phi)$ 的最小值。
若 Σ 的正交對角化爲 $\Sigma=Q\Lambda Q^T$，其中 $Q=(u_1\ u_2\ \cdots\ u_m)$ 爲 orthogonal，且 u_i 所對應的 eigenvalue 爲 λ_i。
取 $\Phi=(u_{m-k+1}\ u_{m-k+2}\ \cdots\ u_m)$，即 Φ 的 column vector 爲 Σ 最小的 k 個 eigenvalue 其對應的 eigenvector。
則 $\Phi\in\mathbb{R}^{m\times k}$，滿足 $\Phi^T\Phi=I_k$，且

$$
\begin{aligned}
&Trace(\Phi^T\Sigma\Phi)\\
&=Trace(\Phi^T(\Sigma\Phi))\\
&=Trace\left(\begin{pmatrix}u_{m-k+1}^T\\u_{m-k+2}^T\\\vdots\\u_m^T\end{pmatrix}\cdot\Sigma\begin{pmatrix}u_{m-k+1}\ u_{m-k+2}\ \cdots\ u_m\end{pmatrix}\right)\\
&=Trace\left(\begin{pmatrix}u_{m-k+1}^T\\u_{m-k+2}^T\\\vdots\\u_m^T\end{pmatrix}\cdot\begin{pmatrix}\Sigma u_{m-k+1}\ \Sigma u_{m-k+2}\ \cdots\ \Sigma u_m\end{pmatrix}\right)\\
&=Trace\left(\begin{pmatrix}u_{m-k+1}^T\\u_{m-k+2}^T\\\vdots\\u_m^T\end{pmatrix}\cdot\begin{pmatrix}\lambda_{m-k+1}u_{m-k+1}\ \lambda_{m-k+2}u_{m-k+2}\ \cdots\ \lambda_m u_m\end{pmatrix}\right)\\
&=Trace\left(\begin{pmatrix}\lambda_{m-k+1}\|u_{m-k+1}\|^2&&&\\&\lambda_{m-k+2}\|u_{m-k+2}\|^2&&\\&&\ddots&\\&&&\lambda_m\|u_m\|^2\end{pmatrix}\right)\\
&=\lambda_{m-k+1}\|u_{m-k+1}\|^2+\lambda_{m-k+2}\|u_{m-k+2}\|^2+\cdots+\lambda_m\|u_m\|^2\\
&=\lambda_{m-k+1}\cdot1^2+\lambda_{m-k+2}\cdot1^2+\cdots+\lambda_m\cdot1^2\\
&=\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m
\end{aligned}
$$

故可得 $\lambda_{m-k+1}+\lambda_{m-k+2}+\cdots+\lambda_m$，即 Σ 最小的 k 個 eigenvalue 之和，爲 $Trace(\Phi^T\Sigma\Phi)$ 的最小值，且當 Φ 的 column vector 分別爲 λ_{m-k+1}、λ_{m-k+2}、$\cdots$、λ_m 所對應的 eigenvector 時，$Trace(\Phi^T\Sigma\Phi)$ 即可取到該最小值。 □

3. Multiclass AdaBoost (1%)

Let $\mathcal{X}$ be the input space, $\mathcal{F}$ be a collection of multiclass classifiers that map from $\mathcal{X}$ to $[\![1,K]\!]$, where K denotes the number of classes. Let $\{(x_i,\hat{y}_i)\}_{i=1}^n$ be

the training data set, where $x_i \in \mathbb{R}^m$ and $\hat{y}_i \in [\![1, K]\!]$.

Given $T \in \mathbb{N}$, suppose we want to find functions

$$g_T^k(x) = \sum_{t=1}^{T} \alpha_t^k f_t(x), \ k \in [\![1, K]\!]$$

where $f_t \in \mathcal{F}$ and $\alpha_t^k \in \mathbb{R}$ for all $t \in [\![1, T]\!]$, $k \in [\![1, K]\!]$, by which the aggregated classifier $h : \mathcal{X} \to [\![1, K]\!]$ is defined as

$$h(x) = argmax_{1 \le k \le K} \ g_T^k(x)$$

Please apply gradient boosting to show how the functions f_t and coefficients α_t^k are computed with an aim to minimize the following loss function

$$L(g_T^1, \cdots, g_T^K) = \sum_{i=1}^{n} exp(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} g_T^k(x_i) - g_T^{\hat{y}_i}(x_i))$$

solution