# ML2025 Fall Homework Assignment 3
# Handwritten

Yi-Lun Lee
b11901057@ntu.edu.tw

Lu-Fang Chiang
b12901140@ntu.edu.tw

October 2025

**Tools You Need to Know/Learn**

- Data centering, Covariance matrix
- Multivariate Gaussian, KL-Divergence
- Graph construction, Laplacian
- EM Algorithm: E-step, M-step updates

You are expected to know (or learn through this homework) the definitions and usages of the above concepts, and are encouraged to discuss with TAs during TA hour if you encounter difficulties.

**Homework Policy**

- The homework is graded out of 105 points.
- The official submission deadline will follow the schedule announced on NTU COOL.
- Homework may be handwritten or typed (e.g. using LaTeX), but must be submitted in **PDF format**.
- If you discuss the homework with classmates, you should state their student IDs in your submission.
- Plagiarism is strictly prohibited. Serious violations will be dealt with according to NTU regulations.

# Problem 1 (Laplacian Eigenmaps) (25 pts)

Consider an undirected connected graph $G$, which is shown below. We want to utilize Laplacian Eigenmaps method to reduce these 10 points to 3-dimensional space. Here, undirected graph means that edges in the graph do not have a direction, and connected graph means that there is a path from any node to any other node in the graph.

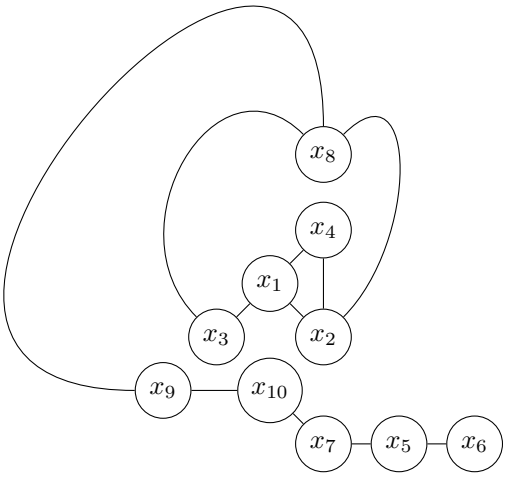


Figure 1: undirected connected graph $G$

1. (2 pts) Write down the adjacency matrix $\boldsymbol{W}$

2. (2 pts) Write down the diagonal matrix $\boldsymbol{D} = \text{diag}(d_1, ..., d_{10})$, where $d_i = \sum_{j=1}^{10} \frac{\boldsymbol{W}_{ij}+\boldsymbol{W}_{ji}}{2}$ and the Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

3. (5 pts) Based on Neighbor Embedding Slide p.7-p.10, use programming tools (e.g., MATLAB or Python) to implement and solve the following optimization problem.

$$\begin{aligned} \text{minimize} \quad & \text{tr}\big(\boldsymbol{\Psi}^\mathsf{T} \boldsymbol{L}\, \boldsymbol{\Psi}\big) \\ \text{subject to} \quad & \boldsymbol{\Psi}^\mathsf{T} \boldsymbol{D}\, \boldsymbol{\Psi} = \boldsymbol{I}_3 \\ \text{variables} \quad & \boldsymbol{\Psi} \in \mathbb{R}^{10\times 3} \end{aligned}$$

Include your code in this file using either (i) Overleaf's `minted` environment (compile with `-shell-escape`) or (ii) a styled screenshot generated with Carbon (`https://carbon.now.sh/`). Also, please plot the reduced points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{10}$ in a 3-D scatter plot.

4. (2 pts) You may find that the minimal eigenvalue of $\boldsymbol{L}$ is 0, and the corresponding eigenvector is

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{1}$$

where $c$ is a constant. Since all the points fall into a plane, the span of these points is $\mathbb{R}^2$. In order to construct $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ such that $\text{span}\{\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}\} = \mathbb{R}^3$, we need choose the second, third, fourth smallest eigenvalue and the corresponding eigenvectors. Please plot the reduced points by the updated $\boldsymbol{z_1}, ..., \boldsymbol{z_{10}}$ in 3-D scatter plot and verify that whether $\text{tr}(\boldsymbol{\Psi}^\mathsf{T} \boldsymbol{L} \boldsymbol{\Psi}) = 1.098$ and $\boldsymbol{\Psi}^\mathsf{T} \boldsymbol{D} \boldsymbol{\Psi} = \boldsymbol{I}_3$.

5. (3 pts) Show that for no matter the graph is, there is an eigenvector of $\boldsymbol{L}$

$$\begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix} \tag{2}$$

where $c$ is a constant, and the corresponding eigenvalue is 0.

6. (3 pts) By Neighbor Embedding Slide p.9, please show that

$$\boldsymbol{f}^\mathsf{T} \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{1 \le i,j \le N} w_{ij} (f_i - f_j)^2, \quad \forall \boldsymbol{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \in \mathbb{R}^N.$$

7. (3 pts) Show that if $\boldsymbol{f}$ is an eigenvector of $\boldsymbol{L}$ which corresponds to eigenvalue 0, then $\boldsymbol{f}^\mathsf{T} \boldsymbol{L} \boldsymbol{f} = 0$.

8. (5 pts) Show that if the graph is connected, the second smallest eigenvalue of $\boldsymbol{L}$ will be nonzero.

## Problem 2 (Principal Component Analysis) (20 pts)

Let $(\mathbf{x}_i)_{i=1}^N$ be $N$ centered data points in $\mathbb{R}^d$ (i.e., $\sum_{i=1}^N \mathbf{x}_i = \mathbf{0}$). Denote the empirical covariance

$$\boldsymbol{\Sigma} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \;\in\; \mathbb{R}^{d\times d}.$$

Throughout, assume $\boldsymbol{\Sigma}$ is symmetric positive semidefinite.

(a) (5 pts) Show that the first principal axis $\mathbf{u}_1$ satisfies

$$\mathbf{u}_1 = \arg\max_{\|\mathbf{u}\|_2=1} \; \mathbf{u}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{u},$$

and that any maximizer must satisfy the eigenvalue equation $\boldsymbol{\Sigma}\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$ with $\lambda_1 = \max_{\|\mathbf{u}\|_2=1} \mathbf{u}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{u}$.

(b) (5 pts) Let $\hat{\mathbf{x}}_i = (\mathbf{u}^{\mathsf{T}}\mathbf{x}_i)\,\mathbf{u}$ be the orthogonal projection of $\mathbf{x}_i$ onto the line spanned by a unit vector $\mathbf{u}$.

Prove the identity

$$\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \;=\; \operatorname{tr}(\boldsymbol{\Sigma}) - \mathbf{u}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{u},$$

and conclude that minimizing the reconstruction error over unit $\mathbf{u}$ is equivalent to part (a).

(c) (5+5 pts) Now take $d = 2$ (i.e., $\mathbf{x}_i = (x_{i,1}, x_{i,2})$) and suppose you are given the aggregated statistics

$$\sum_{i=1}^{N} x_{i,1} = \sum_{i=1}^{N} x_{i,2} = 0, \quad \sum_{i=1}^{N} x_{i,1}^2 = 363, \quad \sum_{i=1}^{N} x_{i,1}x_{i,2} = -60, \quad \sum_{i=1}^{N} x_{i,2}^2 = 482.$$

(i) Compute the eigenvalues and a unit eigenvector $\mathbf{u}_1$ corresponding to the largest eigenvalue.

(ii) Using part (b), express the total reconstruction error $\sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$ along $\mathbf{u}_1$ in closed form.

## Problem 3 (Gradient of the t-SNE Objective) (10pts)

By working on this problem, we hope that students will gain a deeper understanding of the mathematical formulation of the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, particularly the derivation of its gradient from the Kullback–Leibler (KL) divergence objective.

The goal of t-SNE is to find low-dimensional representations $\{\mathbf{y}_i\}_{i=1}^N$ that preserve local similarities of high-dimensional data $\{\mathbf{x}_i\}_{i=1}^N$.

We define the pairwise *high-dimensional similarities* as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad \text{where} \quad p_{j|i} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_i^2}\right)}.$$

The *low-dimensional similarities* are modeled using a Student-$t$ kernel with one degree of freedom:

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq \ell}(1 + \|\mathbf{y}_k - \mathbf{y}_\ell\|^2)^{-1}}, \quad \text{and} \quad q_{ii} = 0.$$

The symmetric t-SNE objective minimizes the KL divergence between these two distributions:

$$C = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

1. (2 pts) Prove that the cost can be simplified to

$$C = \text{constant(to } \mathbf{y}_i) - \sum_{i \neq j} p_{ij} \log q_{ij}.$$

2. (8 pts) Show that the final expression of the gradient is

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}.$$

   Hint: Substitute the definition of $q_{ij}$ into the cost function, and you may use the following useful hint:

$$\log(\frac{A}{B}) = \log A - \log B, \frac{\partial \log(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)}{\partial \mathbf{y}_i} = \frac{2(\mathbf{y}_i - \mathbf{y}_j)}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}.$$

   Carefully account for the contributions where $\mathbf{y}_i$ appears in both indices of the summation.

# Problem 4 (EM for Mixture of Multivariate $t$-Distributions) (30 pts)

**Definitions (first mention).** When the Gamma function and Gamma distribution are first used here, adopt the following standard definitions:

- **Gamma function:** for $a > 0$,
$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt.$$
- **Gamma distribution (shape–rate parametrization):** a random variable $U$ has Gamma$(\alpha, \beta)$ with shape $\alpha > 0$ and rate $\beta > 0$ if its density is
$$f_U(u) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, u^{\alpha-1} e^{-\beta u}, \quad u > 0.$$
(Note you may solve below problem without knowing the above definition, but it's good to know that.)
- **Digamma and trigamma:** the digamma function $\psi(\cdot)$ is the logarithmic derivative of the Gamma function,
$$\psi(a) = \frac{d}{da} \log \Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)},$$
and the trigamma $\psi'(\cdot)$ is its derivative, $\psi'(a) = \frac{d}{da}\psi(a)$.

Consider the generative model parameterized by

$$\theta = \big((\pi_k, \boldsymbol{\mu}_k, \Sigma_k, \nu_k)\big)_{k=1}^K,$$

where each component is a $p$-variate Student-$t$ distribution with center $\boldsymbol{\mu}_k \in \mathbb{R}^p$, positive-definite scale matrix $\Sigma_k \in \mathbb{R}^{p\times p}$, degrees of freedom $\nu_k > 0$, mixing weight $\pi_k > 0$, and $\sum_{k=1}^K \pi_k = 1$.

For an observation $\mathbf{y} \in \mathbb{R}^p$ the mixture density is

$$p_\theta(\mathbf{y}) = \sum_{k=1}^K \pi_k\, t_p(\mathbf{y}; \boldsymbol{\mu}_k, \Sigma_k, \nu_k),$$

where the multivariate Student-$t$ pdf (as in Liu & Rubin (1995)) is

$$t_p(\mathbf{y}; \boldsymbol{\mu}, \Sigma, \nu) = \frac{\Gamma\big(\frac{\nu+p}{2}\big)}{\Gamma\big(\frac{\nu}{2}\big)(\nu\pi)^{p/2}|\Sigma|^{1/2}} \left(1 + \frac{1}{\nu}(\mathbf{y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)^{-\frac{\nu+p}{2}}.$$

Suppose we observe $N$ i.i.d. samples $\mathbf{y}_1, \ldots, \mathbf{y}_N$. The maximum likelihood estimator is

$$\theta^{\text{opt}} = \arg\max_\theta \prod_{i=1}^N p_\theta(y_i).$$

The following problems lead you to derive the E-step and M-step equations of the EM algorithm for mixture of multivariate $t$-distributions.

**EM Algorithm — E-step — posterior expectations and responsibilities.**

As mentioned in class, we can rewrite the MLE as

$$\theta^{\text{opt}} = \arg\max_{\theta}(Q(\theta \mid \theta^{(t)}) + H(\theta \mid \theta^{(t)}))$$

by introducing the latent labels $z_i \in \{1, \ldots, K\}$.

1. (5 pts) Derive the $Q(\theta \mid \theta^{(t)})$ and $H(\theta \mid \theta^{(t)})$. (You can denote $\frac{\pi_k^{(t)}\, t_p(\mathbf{y}_i; \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)}, \nu_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)}\, t_p(\mathbf{y}_i; \boldsymbol{\mu}_\ell^{(t)}, \Sigma_\ell^{(t)}, \nu_\ell^{(t)})}$ as $\delta_{i,k}^{(t)}$ )

2. (4 pts)(you may direct use the result although you do not prove it) if we introduce the latent scale variables $u_{i,k} > 0$ via the standard scale-mixture representation of the multivariate $t$,we are given

$$\mathbf{y}_i \mid (z_i = k, u_{i,k}) \sim \mathcal{N}_p\Big(\boldsymbol{\mu}_k,\ \frac{\Sigma_k}{u_{i,k}}\Big), \qquad u_{i,k} \mid (z_i = k) \sim \text{Gamma}\Big(\frac{\nu_k}{2}, \frac{\nu_k}{2}\Big),$$

   where the Gamma is parameterized by (shape, rate), and define the Mahalanobis distance

$$d_{i,k}(\boldsymbol{\mu}_k, \Sigma_k) \ = \ (\mathbf{y}_i - \boldsymbol{\mu}_k)^{\mathsf{T}} \Sigma_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k).$$

   Show that for each $i, k$, the posterior distribution of $u_{i,k}$ (given $z_i = k$ and $\mathbf{y}_i$) is

$$u_{i,k} \mid (\mathbf{y}_i, z_i = k) \ \sim \ \text{Gamma}\Big(\frac{\nu_k^{(t)} + p}{2},\ \frac{\nu_k^{(t)} + d_{i,k}^{(t)}}{2}\Big),$$

   Hint: multiply the Gaussian likelihood (in $u$) by the Gamma prior and identify the kernel of a Gamma distribution.

3. (3 pts) Show that hence its posterior moments are

$$w_{i,k}^{(t)} \coloneqq \ \mathbb{E}[u_{i,k} \mid \mathbf{y}_i, z_i = k; \theta^{(t)}] \ = \ \frac{\nu_k^{(t)} + p}{\nu_k^{(t)} + d_{i,k}^{(t)}},$$

$$\ell_{i,k}^{(t)} \coloneqq \ \mathbb{E}[\log u_{i,k} \mid \mathbf{y}_i, z_i = k; \theta^{(t)}] \ = \ \psi\Big(\frac{\nu_k^{(t)} + p}{2}\Big) - \log\Big(\frac{\nu_k^{(t)} + d_{i,k}^{(t)}}{2}\Big).$$

**EM Algorithm — M-step — latent-scale representation and closed-form updates.**

1. (12 pts) Using the results in E step , prove the closed-form M-step updates.

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{i,k}^{(t)}.$$

$$\boldsymbol{\mu}_k^{(t+1)} \ = \ \frac{\sum_{i=1}^{N} \delta_{i,k}^{(t)}\, w_{i,k}^{(t)}\, \mathbf{y}_i}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}\, w_{i,k}^{(t)}}$$

and

$$\Sigma_k^{(t+1)} \;=\; \frac{\displaystyle\sum_{i=1}^{N} \delta_{i,k}^{(t)}\, w_{i,k}^{(t)}\, (\mathbf{y}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{y}_i - \boldsymbol{\mu}_k^{(t+1)})^{\mathsf{T}}}{\displaystyle\sum_{i=1}^{N} \delta_{i,k}^{(t)}}$$

by differentiating the $Q$-function w.r.t. $\pi_k, \boldsymbol{\mu}_k$ and $\Sigma_k^{-1}$ (respectively) and setting the derivatives to zero.

Hints: you can simplify $Q(\theta \mid \theta^{(t)})$ and use $w_{i,k}^{(t)}, \ell_{i,k}^{(t)}, \delta_{i,k}^{(t)}, \theta^{(t)}$ to express it.

2. (3 pts) **degrees of freedom $\nu_k$.**
Starting from the $Q$-function (after replacing $u_{i,k}$ by its posterior moments $w_{i,k}^{(t)}$ and $\ell_{i,k}^{(t)}$), derive the stationarity equation for $\nu_k$. That is, show that the maximizing $\nu_k^{(t+1)}$ (solution of $\partial Q/\partial \nu_k = 0$) satisfies the scalar equation

$$\boxed{\log\Big(\frac{\nu_k}{2}\Big) - \psi\Big(\frac{\nu_k}{2}\Big) + 1 \;+\; \frac{1}{\sum_{i=1}^{N} \delta_{i,k}^{(t)}} \sum_{i=1}^{N} \delta_{i,k}^{(t)} \big(\ell_{i,k}^{(t)} - w_{i,k}^{(t)}\big) \;=\; 0,}$$

3. (3 pts) Provide the Newton's method (one-step update) used to solve this equation numerically, i.e.

$$\nu \leftarrow \nu - \frac{f(\nu)}{f'(\nu)},$$

and give explicit expressions for $f(\nu)$ and $f'(\nu)$ (in terms of digamma/trigamma functions).

Hint: set $a = \nu_k/2$ to simplify derivatives of $\log\Gamma(a)$. Use $\frac{d}{da}\log\Gamma(a) = \psi(a)$ and $\frac{d}{da}\psi(a) = \psi'(a)$

**Remark**

- **Degrees-of-freedom control tail strength:** the parameter $\nu_k$ moderates the sensitivity. For fixed $d_{i,k}$, larger $\nu_k$ increases the numerator $\nu_k + p$ but also the denominator $\nu_k + d_{i,k}$ — in the limit $\nu_k \to \infty$ we have $w_{i,k} \to 1$ and the updates reduce to the Gaussian-mixture EM updates. For small $\nu_k$ the down-weighting is stronger: heavy tails mean more robustness.
- **Intuition:** the latent precision $u_{i,k}$ allows each observation to "explain" a different amount of variability. If an observation is unlikely under component $k$ (large Mahalanobis distance), the posterior for $u_{i,k}$ concentrates near small values, effectively increasing the conditional variance for that observation under component $k$ and reducing its influence in parameter estimation. This built-in reweighting is exactly why mixtures of $t$-components are robust to outliers relative to Gaussian mixtures.

## Problem 5 (Expectation Maximization Interpretation behind Semi-Supervised Learning) (20 pts)

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1, \ldots, y_N \in \{0, 1, \ldots, K\}$. Consider the generative model where each sample $\mathbf{x}_i$ is generated independently according to Gaussian mixture model that depends on the label $y_i$, as represented by random variable

$$X_i \sim \begin{cases} \sum_{j=1}^{K} \pi_j \mathcal{N}\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) & , \text{ if } y_i = 0 \\ \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) & , \text{ if } y_i = k \neq 0 \end{cases}$$

where $\pi_1 + \ldots + \pi_K = 1$, and $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with probability density function

$$\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^m \left|\boldsymbol{\Sigma}\right|}} \exp\left(-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right)$$

We would like to apply Expectation Maximization algorithm to find the maximum likelihood estimation of parameters $\theta = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$.

1. (15 pts) Please write down the E-step and M-step and show that the parameters are updated from $\theta^{(t)} = \left\{\left(\pi_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}\right)\right\}_{k=1}^{K}$ to $\theta^{(t+1)} = \left\{\left(\pi_k^{(t+1)}, \boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}\right)\right\}_{k=1}^{K}$ in the following form:

$$\pi_k^{(t+1)} = \frac{\sum_{i:y_i=0} \delta_{ik}^{(t)}}{\sum_{i:y_i=0} 1}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \mathbf{x}_i + \sum_{i:y_i=0} \delta_{ik}^{(t)} \mathbf{x}_i}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i:y_i=k} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^{\mathsf{T}} + \sum_{i:y_i=0} \delta_{ik}^{(t)} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}\right)^{\mathsf{T}}}{N_k + \sum_{i:y_i=0} \delta_{ik}^{(t)}}$$

where $N_k = \sum_{i:y_i=k} 1$ is the number of samples in class k. Please show your derivations.

2. (5 pts) What is the closed form expression of $\delta_{ik}^{(t)}$? Please show your derivations.