Homework week 5

Problem 1 Given

$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu)\right)}, \tag{0.1}$$

where $x, \mu \in \mathbb{R}^k$, Σ is a k-by-k positive definite matrix and $|\Sigma|$ is its determinant. Show that $\int f(x)\,\mathrm{d}x = 1$.

First, we know that

$$\mathrm{I} = \int_{\mathbb{R}} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}. \tag{0.2}$$

Note also that Σ is positive definite matrix, its can be written as follows:

$$\Sigma = Q^T D Q, \qquad Q : \text{orthogonal matrix}, \quad D : \text{diagonal matrix with eigenvalues } \lambda_1, \dots \lambda_k \text{ w.r.t } \Sigma. \tag{0.3}$$

Let $y = Q(x-\mu)$. The integral becomes as follows:

$$\begin{aligned} \frac{1}{((2\pi)^k|\Sigma|)^{\frac{1}{2}}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\,\mathrm{d}x &= \frac{1}{((2\pi)^k|Q\Sigma Q^T|)^{\frac{1}{2}}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}y^T D^{-1} y}\,\mathrm{d}y \\ &= \frac{1}{((2\pi)^k(\Pi_{i=1}^k\lambda_i))^{\frac{1}{2}}} \Pi_{i=1}^k \left(\int_{\mathbb{R}} e^{-\frac{y_i^2}{2\lambda_i}}\,\mathrm{d}y_i \right). \end{aligned} \tag{0.4}$$

Then, applying the result (0.2), we obtain

$$\int_{\mathbb{R}^k} f(x)\,\mathrm{d}x = \frac{1}{((2\pi)^k(\Pi_{i=1}^k\lambda_i))^{\frac{1}{2}}} \Pi_{i=1}^k(\sqrt{2\pi\lambda_i}) = 1. \tag{0.5}$$

Problem 2 Let A, B be n-by-n matrices and x be a n-by-1 vector.

(a) Show that $\frac{\partial}{\partial A}\mathrm{trace}(AB) = B^T$.

(b) Show that $x^T A x = \mathrm{trace}(xx^T A)$.

(c) Derive the maximum likelihood estimators for a multivariate Gaussian.

Proof of (2-a) Observe that

$$\mathrm{trace}(AB) = \sum_{i=1}^n \begin{pmatrix} a_{i1} & a_{i2} & \dots & a_{in} \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \\ \vdots \\ b_{ni} \end{pmatrix} \tag{0.6}$$

Then, we find

$$\frac{\partial}{\partial a_{ij}}\mathrm{trace}(AB) = b_{ji}, \quad \forall i, j = 1, \dots, n. \tag{0.7}$$

This implies that $\frac{\partial}{\partial A}\mathrm{trace}(AB) = B^T$.

Proof of (2-b) Observe that

$$x^T A x = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$= \sum_{i,j=1}^{n} (x_i a_{ij} x_j). \tag{0.8}$$

Next, we know that

$$xx^T = \begin{pmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2x_2 & \dots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \dots & x_nx_n \end{pmatrix}. \tag{0.9}$$

Then

$$\mathrm{trace}(xx^T A) = \sum_{j=1}^{n} (x_1 a_{1j} + x_2 a_{2j} x_j + \dots + x_n a_{nj} x_j) = \sum_{i,j=1}^{n} x_i a_{ij} x_j. \tag{0.10}$$

Therefore, we show that $x^T A x = \mathrm{trace}(xx^T A)$.

Proof of (2-c) Recall the maximum likelihood function

$$L(\theta) = \Pi_{i=1}^{n} \left(\frac{1}{\sqrt{((2\pi)^k |\Sigma|)}} e^{-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu)} \right) \tag{0.11}$$

and let

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \left(-\frac{1}{2}(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu) - \frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| \right). \tag{0.12}$$

Then

$$\frac{\partial}{\partial \mu}\ell(\theta) = -\frac{1}{2}\sum_{i=1}^{n} \frac{\partial}{\partial \mu}\left[(x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu) \right] = \sum_{i=1}^{n} \Sigma^{-1}(x^{(i)}-\mu) = 0. \tag{0.13}$$

This implies that

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}. \tag{0.14}$$

Next, we find

$$\frac{\partial}{\partial \Sigma^{-1}}\ell(\theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\frac{\partial}{\partial \Sigma^{-1}} \left((x^{(i)}-\mu)^T \Sigma^{-1}(x^{(i)}-\mu) \right) + \frac{1}{2}\frac{\partial}{\partial \Sigma^{-1}} \log|\Sigma^{-1}| \right]$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{2} \left((x^{(i)}-\mu)(x^{(i)}-\mu)^T \right)^T + \frac{1}{2}\Sigma \right] = 0, \tag{0.15}$$

where we use the fact $\frac{\partial}{\partial A}\log|A| = A^{-1}$. Therefore, we find

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} (x^{(i)}-\mu)(x^{(i)}-\mu)^T, \qquad \text{as} \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}. \tag{0.16}$$

Problem 3

Question In class, when we formulate the problem of learning $\mathbb{P}(y|x)$, we typically assume that $\mathbb{P}(x|y=0)$ and $\mathbb{P}(x|y=1)$ are Gaussian, while $\mathbb{P}(y=0)$ and $\mathbb{P}(y=1)$ are Bernouli. The teacher also noted that if $\mathbb{P}(x|y=0)$ and $\mathbb{P}(x|y=1)$ are allowed to be a more general distributions, the problem becomes substantially more difficult.

My question is whenever this difficulty is primarily computational (numerical implementation), or does it stem from a lack of applicable theoretical tools?