1. $f(x) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Since Σ is positive definite matrix, there exist a unique positive definite A such that $\Sigma = AA^T$

and $|\Sigma| = |A|^2$

Let $y = A^{-1}(x-\mu) \Leftrightarrow x = \mu + Ay$

Thus, $dx = |\det(A)|\, dy = \sqrt{|\Sigma|}\, dy$

$(x-\mu)^T \Sigma^{-1}(x-\mu) = (Ay)^T \Sigma^{-1}(Ay) = y^T y = \|y\|^2$ $\quad (A^T \Sigma^{-1} A = A^T (AA^T)^{-1} A = A^T (A^T)^{-1} A^{-1} A = I)$

Therefore, $\frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) dx$

$$= \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \int_{\mathbb{R}^k} \exp^{-\frac{1}{2}\|y\|^2} \sqrt{|\Sigma|}\, dy$$

$$= (2\pi)^{-\frac{k}{2}} \int_{\mathbb{R}^k} e^{-\frac{1}{2}\|y\|^2} dy$$

$$= (2\pi)^{-\frac{k}{2}} \cdot (2\pi)^{\frac{k}{2}}$$

$$= 1$$

2. (a)

Let $tr(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} B_{ji}$

Differential for any component A_{pq} :

$$\frac{\partial}{\partial A_{pq}} tr(AB) = B_{qp}$$

After filling in a matrix with all the subscript, we can get B^T

(b)

We have $x^T A x = tr(x^T A x)$ since any scalar is equal to its trace.

By the cyclical property of the trace, we get $tr(ABC) = tr(CAB)$

Thus, $x^T A x = tr(x^T A x) = tr(x x^T A)$

(c)

We have sample $x_1, x_2, \cdots, x_n \in \mathbb{R}^k$ and suppose $x_i \sim$ i.i.d. $N_k(\mu, \Sigma)$, where $\mu \in \mathbb{R}^k$ and Σ is a symmetry positive definite matrix.

PDF :

$$f(x_i|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right)$$

Since x_i is i.i.d., the overall likelihood function is :

$$L(\mu,\Sigma) = \prod_{i=1}^{n} f(x_i|\mu,\Sigma)$$

Thus, $l(\mu,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)$, where we ignore the constant $-\frac{nk}{2}\log(2\pi)$

Define the scatter matrix $S(\mu) = \sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T$.

Then, $l(\mu,\Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}tr(\Sigma^{-1}S(\mu))$

To find the maximum value of μ, let $\frac{\partial l}{\partial \mu} = 0$.

Then, we get $\Sigma^{-1}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$

$\sum_{i=1}^{n} x_i - n\mu = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$

To find the maximum value of Σ, let $\frac{\partial}{\partial\Sigma} l(\hat{\mu}, \Sigma) = 0$

Then, we get $-\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1} S \Sigma^{-1} = 0$

$\Rightarrow \Sigma^{-1} S \Sigma^{-1} = n\Sigma^{-1}$

$\Rightarrow S = n\Sigma$

$\Rightarrow \hat{\Sigma} = \frac{1}{n} S(\hat{\mu})$

Thus, we get $\hat{\mu} = \bar{x}$ and $\hat{\Sigma} = \frac{1}{n} S(\hat{\mu})$

3.

Why can it be assumed that each category is a Gaussian distribution?
If the data distribution is not Gaussian, what will be the consequences of GDA?
How can it be improved?