CS613 \ Assignment #2 Solutions

Ariel Stolerman

1)

Let x be a multi-dimensional random variable with a joint (multivariate) Gaussian distribution $\mathcal{N}(x|\mu,\Sigma)$ with $\Lambda \equiv \Sigma^{-1}$ and:

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Following are proofs for:

1.1)

$p(x_a|x_b) = \mathcal{N}\left(x\middle|\mu_{a|b}, \Sigma_{a|b}\right)$, where $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$ and $\Sigma_{a|b} = \Lambda_{aa}^{-1}$:

We measure $p(x_a|x_b)$ by looking at $p(x_a, x_b)$ and fixing x_b.

We know that in the Multivariate case:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Looking at the exponent and breaking it down we obtain:

$$-\frac{1}{2}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right)^T \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}\left(\begin{pmatrix} x_a \\ x_b \end{pmatrix} - \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}\right) =$$

$$-\frac{1}{2}\left(\begin{pmatrix} x_a - \mu_a \\ 0 \end{pmatrix}^T + \begin{pmatrix} 0 \\ x_b - \mu_b \end{pmatrix}^T\right)\left(\begin{pmatrix} \Lambda_{aa} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \Lambda_{ab} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \Lambda_{ba} & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Lambda_{bb} \end{pmatrix}\right)\left(\begin{pmatrix} x_a - \mu_a \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_b - \mu_b \end{pmatrix}\right) =$$

$$\underbrace{-\frac{1}{2}\left[(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) + (x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b) + (x_b-\mu_b)^T\Lambda_{ba}(x_a-\mu_a) + (x_b-\mu_b)^T\Lambda_{bb}(x_b-\mu_b)\right]}_{(A):=}$$

The exponent in its general form is:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu + \frac{1}{2}\mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu \underset{\substack{\Sigma\ symmetric \\ \Rightarrow \Sigma^{-1}\ also}}{=} -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu$$

Now we will address x_a as the variable in (A) and x_b as a constant, and observe the second-order terms w.r.t. x_a in (A):

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a$$

And by comparing with the general form:

$$-\frac{1}{2}x^T\Sigma^{-1}x = -\frac{1}{2}x_a^T\Lambda_{aa}x_a \Leftrightarrow \Sigma^{-1} = \Lambda_{aa} \Leftrightarrow \boxed{\Sigma_{a|b} = \Lambda_{aa}^{-1}}$$

Now the terms in (A) that are linear in x_a:

$$-\frac{1}{2}\left[-x_a^T\Lambda_{aa}\mu_a - \mu_a\Lambda_{aa}x_a + x_a^T\Lambda_{ab}x_b - x_a^T\Lambda_{ab}\mu_b + x_b^T\Lambda_{ba}x_a - \mu_b^T\Lambda_{ba}x_a\right] =$$

$$[\text{Note: } \Lambda_{aa} \text{ is symmetric}, \Lambda_{ab} = \Lambda_{ba}^T \text{ and we know that } x^TAy = y^TA^Tx]$$

$$-\frac{1}{2}\left[-2x_a^T\Lambda_{aa}\mu_a + 2x_a^T\Lambda_{ab}x_b - 2x_a^T\Lambda_{ab}\mu_b\right] = x_a^T[\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)]$$

And by comparing with the general form:

$$x^T\Sigma^{-1}\mu = x_a^T[\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)] \Leftrightarrow \Sigma^{-1}\mu = \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b) \Leftrightarrow \mu = \Sigma\left(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)\right)$$

Now we plug in $\Sigma_{a|b} = \Lambda_{aa}^{-1}$ and obtain:

$$\mu_{a|b} = \Lambda_{aa}^{-1}\left(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b - \mu_b)\right) = \boxed{\mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)}$$

Thus we have proven that $p(x_a|x_b) = \mathcal{N}\left(x\middle|\mu_{a|b}, \Sigma_{a|b}\right)$.

1.2)

$p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa})$:

The marginal distribution $p(x_a)$ is given by:

$$p(x_a) = \int p(x_a, x_b)\mathrm{d}x_b$$

Now looking at (A) like before only now x_b is the variable and x_a is addressed as a constant, the terms in the exponent involving x_b are $-\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu$, and plugging in what we have derived for x_a, only now w.r.t. x_b:

$$-\frac{1}{2}x_b^T\Lambda_{bb}x_b + x_b^T[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)]$$

Denote $m \coloneqq [\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)]$; we will now bring this expression to a standard quadratic form of a Gaussian distribution plus some term independent of x_b but dependent on x_a (the following is backwards):

$$-\frac{1}{2}(x_b - \Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1}m) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m =$$

$$-\frac{1}{2}[x_b^T\Lambda_{bb}x_b - x_b^T\Lambda_{bb}\Lambda_{bb}^{-1}m - (\Lambda_{bb}^{-1}m)^T\Lambda_{bb}x_b + (\Lambda_{bb}^{-1}m)^T\Lambda_{bb}\Lambda_{bb}^{-1}m] + \frac{1}{2}m^T\Lambda_{bb}^{-1}m =$$

$$[\text{Note: } (\Lambda_{bb}^{-1}m)^T = m^T(\Lambda_{bb}^{-1})^T = m^T\Lambda_{bb}^{-1}]$$

$$-\frac{1}{2}x_b^T\Lambda_{bb}x_b \underbrace{+\frac{1}{2}x_b^Tm + \frac{1}{2}m^Tx_b}_{equal} \underbrace{-\frac{1}{2}m^T\Lambda_{bb}^{-1}m + \frac{1}{2}m^T\Lambda_{bb}^{-1}m}_{cancel\ each\ other\ out} = -\frac{1}{2}x_b^T\Lambda_{bb}x_b + x_b^Tm$$

Therefore when we take the exponential of the quadratic form we got, the integral is of the form:

$$\int \exp\left[-\frac{1}{2}(x_b - \Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1}m)\right]\mathrm{d}x_b$$

Now we can integrate out x_b and the only term dependent on x_a is:

$$\frac{1}{2}m^T\Lambda_{bb}^{-1}m = \frac{1}{2}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)]^T\Lambda_{bb}^{-1}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)]$$

And combined with the other x_a-dependent terms in (A):

$$\frac{1}{2}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)]^T\Lambda_{bb}^{-1}[\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)] - \frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T[\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b] + C =$$

Where C is a constant independent of x_a. Developing this further, given Λ_{bb} is symmetric and $\Lambda_{ab}^T = \Lambda_{ba}$, more terms are added to the constant part:

$$\frac{1}{2}[\mu_b^T\Lambda_{bb}^T\Lambda_{bb}^{-1}\Lambda_{bb}\mu_b - \mu_b^T\Lambda_{bb}^T\Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a) - (x_a - \mu_a)^T\Lambda_{ba}^T\Lambda_{bb}^{-1}\Lambda_{bb}\mu_b + (x_a - \mu_a)^T\Lambda_{ba}^T\Lambda_{bb}^{-1}\Lambda_{ba}(x_a - \mu_a)] -$$

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T[\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b] + C = \qquad [\text{note: now we get rid of some terms into } C]$$

$$\frac{1}{2}\left[\underbrace{-\mu_b^T\Lambda_{ba}x_a - x_a^T\Lambda_{ab}\mu_b}_{equal} + x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}x_a \underbrace{-x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a - \mu_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}x_a}_{equal}\right] - \frac{1}{2}x_a^T\Lambda_{aa}x_a +$$

$$+x_a^T[\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b] + C =$$

$$-x_a^T\Lambda_{ab}\mu_b + \frac{1}{2}x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}x_a - x_a^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\mu_a - \frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T\Lambda_{aa}\mu_a + x_a^T\Lambda_{ab}\mu_b + C =$$

$$-\frac{1}{2}x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a + x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a + C$$

And now we compare quadratic and linear terms with the general form as before:

$$-\frac{1}{2}x^T\Sigma^{-1}x = -\frac{1}{2}x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})x_a \Leftrightarrow \Sigma^{-1} = \Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba} \Leftrightarrow \boxed{\Sigma = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}}$$

And denote that $\Sigma_a \coloneqq \Sigma = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$

$$x^T\Sigma^{-1}\mu = x_a^T(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a \Leftrightarrow \Sigma^{-1}\mu = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a \underset{plug\ \Sigma_a}{\Longleftrightarrow} \boxed{\mu = \Sigma_a(\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a = \mu_a}$$

Therefore the mean is μ_a. To simplify Σ_a we will use:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$

Where $M = (A - BD^{-1}C)^{-1}$. We know:

$$\begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

Therefore $\Sigma_{aa} = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$, which is exactly Σ_a, therefore $\Sigma_a = \Sigma_{aa}$.

Finally:

$$\boxed{p(x_a) = \mathcal{N}(x_a|\mu_a, \Sigma_{aa})}$$

2)

Let a generative classification model for K classes be defined by prior class probabilities $p(C_k) = \pi_k$ and likelihoods given by Gaussian distributions with a shared covariance matrix:

$$p(x|C_k) = \mathcal{N}(x|\mu_k, \Sigma)$$

Let there be a training set $\{x_n, t_n\}_{n=1}^N$ with t_n as binary target vectors of length K with 1-of-K coding scheme and the data points are drawn independently from this model.

Following are proofs for:

2.1)

For N_k being the number of data points assigned to class C_k, the maximum likelihood for the prior probabilities is $\pi_k = \frac{N_k}{N}$:

Denote $t = (t_1, \dots, t_N)^T$, the likelihood is:

$$p(t|\{\pi_k\}) = \prod_{n=1}^N\prod_{i=1}^K[\pi_i\mathcal{N}(x_n|\mu_i, \Sigma)]^{t_{ni}} \Rightarrow \ln p(t|\{\pi_k\}) = \sum_{n=1}^N\sum_{i=1}^K t_{ni}[\ln\pi_i + \ln\mathcal{N}(x_n|\mu_i, \Sigma)] =$$

$$\sum_{n=1}^N\sum_{i=1}^K t_{ni}\left[\ln\pi_i + \ln\left(\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}(x_n - \mu_i)^T\Sigma^{-1}(x_n - \mu_i)\right]\right)\right] =$$

$$\sum_{n=1}^N\sum_{i=1}^K tn_i\left[\ln\pi_i - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x_n - \mu_i)^T\Sigma^{-1}(x_n - \mu_i)\right]$$

We add a Lagrange multiplier $\lambda(\sum_{i=1}^{K}\pi_i - 1)$ (as $\sum_{i=1}^{K}\pi_i = 1$) and derive with respect to π_k:

$$\frac{\partial \ln p(t|\{\pi_k\})}{\partial \pi_k} = \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda = 0 \Leftrightarrow \sum_{n=1}^{N} t_{nk} = \underbrace{N_k = -\lambda\pi_k}_{(*)} \underset{over\ k}{\overset{sum}{\Longleftrightarrow}} \sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk} = -\lambda\sum_{k=1}^{K}\pi_k \Leftrightarrow N = -\lambda \Leftrightarrow \lambda = -N \Rightarrow$$

$(*)$: $N_k = -\lambda\pi_k \Leftrightarrow \boxed{\pi_k = \frac{N_k}{N}}$

2.2)

The maximum likelihood for the mean of the Gaussian distribution for class C_k (which represents the mean of the data points assigned to class C_k) is $\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}x_n$:

We derive the log of the likelihood we got before, only this time w.r.t. μ_k. We can ignore some terms that will fall out in the derivative, and derive only $A \coloneqq -\frac{1}{2}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)^T\Sigma^{-1}(x_n - \mu_k)$ (disregarding constants):

$$\frac{\partial A}{\partial \mu_k} = \sum_{n=1}^{N} t_{nk}(x_n - \mu_k)^T\Sigma^{-1} = 0 \underset{\cdot\Sigma}{\Leftrightarrow} \sum_{n=1}^{N} t_{nk}(x_n - \mu_k) = 0 \Leftrightarrow \sum_{n=1}^{N} t_{nk}x_n = \underbrace{\sum_{n=1}^{N} t_{nk}}_{=N_k}\mu_k \Leftrightarrow \boxed{\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}x_n}$$

2.3)

The maximum likelihood for the shared covariance matrix is $\Sigma = \sum_{k=1}^{K}\frac{N_k}{N}S_k$ where $S_k = \frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)(x_n - \mu_k)^T$:

We need to show that $\hat{\Sigma}_{MLE} = \sum_{k=1}^{K}\frac{N_k}{N}\frac{1}{N_k}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)(x_n - \mu_k)^T = \boxed{\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)(x_n - \mu_k)^T} =: A$

We derive only terms with Σ and ignore constants. Denote $B \coloneqq \sum_{n=1}^{N}\sum_{i=1}^{K} tn_i\left[-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x_n - \mu_i)^T\Sigma^{-1}(x_n - \mu_i)\right]$:

Notes:

- $-\ln|\Sigma| = \ln|\Sigma|^{-1} = \ln|\Sigma^{-1}|$
- $(x_n - \mu_i)^T\Sigma^{-1}(x_n - \mu_i) = tr\big((x_n - \mu_i)^T\Sigma^{-1}(x_n - \mu_i)\big) = tr(\Sigma^{-1}(x_n - \mu_i)(x_n - \mu_i)^T)$ – because of trace invariance under cyclic permutations. In addition: $\frac{\partial tr(AB)}{\partial A} = B^T$ and of course $(xx^T)^T = (x^T)^Tx^T = xx^T$.

$$\Rightarrow B = \sum_{i=1}^{K}\left[\frac{1}{2}\sum_{n=1}^{N} t_{ni}\ln|\Sigma^{-1}| - \frac{1}{2}\sum_{n=1}^{N} t_{ni}tr(\Sigma^{-1}(x_n - \mu_i)(x_n - \mu_i)^T)\right] \Rightarrow$$

$$\frac{\partial B}{\partial \Sigma^{-1}} = \sum_{i=1}^{K}\left[\frac{1}{2}\sum_{n=1}^{N} t_{ni}\Sigma - \frac{1}{2}\sum_{n=1}^{N} t_{ni}(x_n - \mu_i)(x_n - \mu_i)^T\right] = 0 \Leftrightarrow \sum_{i=1}^{K}\sum_{n=1}^{N} t_{ni}\Sigma = \sum_{i=1}^{K}\sum_{n=1}^{N} t_{ni}(x_n - \mu_i)(x_n - \mu_i)^T \Leftrightarrow$$

$$\Sigma\underbrace{\sum_{i=1}^{K}\sum_{n=1}^{N} t_{ni}}_{=N} = \sum_{i=1}^{K}\sum_{n=1}^{N} t_{ni}(x_n - \mu_i)(x_n - \mu_i)^T \Leftrightarrow \boxed{\Sigma = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} t_{nk}(x_n - \mu_k)(x_n - \mu_k)^T} = A$$

3)

Let $X = \{x_n\}_{n=1}^N$ be a set of data points with a convex hull defined as $x = \sum_{n=1}^N \alpha_n x_n$ where $\alpha_n \geq 0, \sum_{n=1}^N \alpha_n = 1$. Let $Y = \{y_m\}_{m=1}^M$ be another set of data points with their corresponding convex hull $y = \sum_{m=1}^M \beta_m y_m$ where $\beta_m \geq 0, \sum_{m=1}^M \beta_m = 1$ (generalizing from the question, where $Y = \{y_n\}$ implies $|Y| = |X|$).

X and Y are linearly separable if there exists a vector $\hat{w}$ and a scalar w_0 such that:

$$(1) \quad \forall x_n \in X: \ \hat{w}^T x_n + w_0 > 0$$

$$(2) \quad \forall y_m \in Y: \ \hat{w}^T y_m + w_0 < 0$$

Following is a proof that: X and Y intersect ⇒ X and Y are NOT linearly separable:

We will show the contraposition: X and Y are linearly separable ⇒ $X \bigcap Y = \emptyset$.

Assume X and Y are linearly separable, then there exists a vector $\hat{w}$ and a scalar w_0 such that $(1), (2)$ apply.

The linear discriminant for the points in the convex hull of X:

$$f(x) = \hat{w}^T x + w_0 = \hat{w}^T \left(\sum_{n=1}^N \alpha_n x_n \right) + w_0 = \sum_{n=1}^N \alpha_n (\hat{w}^T x_n) + w_0 \underset{\sum_{n=1}^N \alpha_n = 1}{=} \sum_{n=1}^N \alpha_n (\hat{w}^T x_n + w_0)$$

Similarly the linear discriminant for the points in the convex hull of Y:

$$f(y) = \sum_{m=1}^M \beta_m (\hat{w}^T y_m + w_0)$$

Assume that $X \bigcap Y \neq \emptyset$, then there exists $p \in X \bigcap Y$ such that:

$$f(p) = \sum_{n=1}^N \alpha_n (\hat{w}^T p_n + w_0) = \sum_{m=1}^M \beta_m (\hat{w}^T p_m + w_0)$$

But, due to the constraints on α_i, β_i, it is impossible that both $f(p) = \hat{w}^T p_n + w_0 > 0$ and $f(p) = \hat{w}^T p_n + w_0 < 0$, thus X and Y are not linearly separable, by contradiction to the assumption. Therefore there cannot exist such $p \in X \bigcap Y$, and so $X \bigcap Y = \emptyset$, as required to show.