1] Consider the pdf for two jointly Gaussian RVs X and Y, $f_{X,Y}(x,y)$ and the pdf for the multivariate jointly Gaussian RV $Z \in \mathbb{R}^n$ $f_Z(z)$. Suppose $Z = [X, Y]^T$.

a) Find Σ, Σ^{-1} and μ in terms of $m_1, m_2, \sigma_1, \sigma_2, \rho_{XY}$.

↳ Strategy: find the variables using knowledge of probability then prove the result by showing $f_{X,Y}(x,y) = f_Z(z)$.

- $\Sigma = \begin{bmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(Y,X) \\ \mathrm{cov}(X,Y) & \mathrm{cov}(Y,Y) \end{bmatrix}$ by definition.

Claim: $\mathrm{cov}(X,X) = \mathrm{var}(X) = \sigma_1^2$ and $\mathrm{cov}(Y,Y) = \mathrm{var}(Y) = \sigma_2^2$.

Recall, the correlation coefficient ρ_{XY} is defined $\rho_{XY} := \dfrac{\mathrm{cov}(X,Y)}{\sigma_1 \sigma_2}$.

$\Longrightarrow \mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) = \rho_{XY}\sigma_1\sigma_2$

So, it is likely true that $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{XY}\sigma_1\sigma_2 \\ \rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

- Taking the inverse of Σ, which is a 2x2 matrix, obtain

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \rho_{XY}^2\sigma_1^2\sigma_2^2} \begin{bmatrix} \sigma_2^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}$$

- Guess that m_1 is the mean of X and m_2 is the mean of Y.
So by definition, $\mu = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}$

- Given these guesses for Σ, Σ^{-1}, μ, now show $f_Z(z) = f_{X,Y}(x,y)$ which would then prove the relationships are true.

$$f_Z(z) = \frac{\exp\{-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\}}{\sqrt{(2\pi)^n|\Sigma|}}$$

$$= \frac{\exp\left\{-\frac{1}{2}\left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}\right]^T \frac{1}{\sigma_1^2\sigma_2^2 - \rho_{XY}^2\sigma_1^2\sigma_2^2}\begin{bmatrix} \sigma_2^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}\left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}\right]\right\}}{\sqrt{(2\pi)^2(\sigma_1^2\sigma_2^2 - \sigma_1^2\sigma_2^2\rho_{XY}^2)}}$$

$\longrightarrow$

$$= \frac{\exp\left\{-\frac{1}{2(1-\rho_{XY}^2)\sigma_1^2\sigma_2^2}\begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix}\begin{bmatrix} \sigma_2^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}\begin{bmatrix} x-m_1 \\ y-m_2 \end{bmatrix}\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}}$$

$$= \frac{\exp\left\{-\frac{1}{2(1-\rho_{XY}^2)\sigma_1^2\sigma_2^2}\begin{bmatrix} x-m_1 & y-m_2 \end{bmatrix}\begin{bmatrix} \sigma_2^2(x-m_1) - \rho_{XY}\sigma_1\sigma_2(y-m_2) \\ -\rho_{XY}\sigma_1\sigma_2(x-m_1) + \sigma_1^2(y-m_2) \end{bmatrix}\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}}$$

$$= \frac{\exp\left\{\frac{-1}{2(1-\rho_{XY}^2)\sigma_1^2\sigma_2^2}\left[(x-m_1)\sigma_2^2(x-m_1) - (x-m_1)\rho_{XY}\sigma_1\sigma_2(y-m_2) - (y-m_2)\rho_{XY}\sigma_1\sigma_2(x-m_1) + \sigma_1^2(y-m_2)\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}}$$

$$= \frac{\exp\left\{\frac{-1\ \sigma_1^2\sigma_2^2}{2(1-\rho_{XY}^2)\sigma_1^2\sigma_2^2}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho_{XY}\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{XY}^2}}$$

$$= \frac{\exp\left\{\frac{-1}{2(1-\rho^2_{XY})}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho_{XY}\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2\right]\right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2_{XY}}}$$

$$= f_{X,Y}(x,y)$$

Hence, the following is true:

$$\boxed{\begin{aligned} \Sigma &= \begin{bmatrix} \sigma_1^2 & \rho_{XY}\sigma_1\sigma_2 \\ \rho_{XY}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ \Sigma^{-1} &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{XY}^2)}\begin{bmatrix} \sigma_2^2 & -\rho_{XY}\sigma_1\sigma_2 \\ -\rho_{XY}\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} \\ \mu &= \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \end{aligned}}$$

b) Suppose $\rho_{XY} = 0$.

In this case, $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$.

Then, $$f_{X,Y}(x,y) = \frac{\exp\{-\frac{1}{2}\left[\left(\frac{x-m_1}{\sigma_1}\right)^2 + \left(\frac{y-m_2}{\sigma_2}\right)^2\right]}{2\pi\sigma_1\sigma_2}$$

$$= \frac{\exp\{-\frac{1}{2}\left(\frac{x-m_1}{\sigma_1}\right)^2\} \exp\{-\frac{1}{2}\left(\frac{y-m_2}{\sigma_2}\right)^2\}}{2\pi\sigma_1\sigma_2}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}\right)\left(\frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}}\right)$$

$$= f_x(x)\, f_y(y).$$

$$\longrightarrow f_{x,y}(x,y) = f_x(x)\, f_y(y).$$

Because we can write $f_x(x,y)$ as the product of two single variate Gaussian distributions $f_x(x)$, $f_y(y)$, it follows that X and Y are independent.

2] Suppose Naive Bayes Assumption (NB) applies to GDA. Show $\Sigma_{ij} = 0 \ \forall i \neq j$, i.e. all off diagonal elements of Σ equal 0.

Suppose the NB assumption holds.
Then, $P(x_i | Y, x_j) = P(x_i | Y) \ \forall j \neq i$.
Hence, $P(x_i, x_j | Y) = P(x_i|Y) P(x_j | Y, x_i) = P(x_i|Y) P(x_j|Y) \ \forall j \neq i$.
$\Rightarrow x_i|Y, x_j|Y$ are independent $\forall j \neq i$.

Claim: If $x_i|Y$ and $x_j|Y$ are independent, then $\text{cov}(x_i|Y, x_j|Y) = 0$.

$\hookrightarrow \text{cov}(x_i|Y, x_j|Y) = E[(x_i|Y - E[x_i|Y])(x_j|Y - E[x_j|Y])]$ by definition

$= E[x_i|Y - E[x_i|Y]] E[x_j|Y - E[x_j|Y]]$ by independence

$= (E[x_i|Y] - E[E[x_i|Y]])(E[x_j|Y] - E[E[x_j|Y]])$ by linearity of E

$= (E[x_i|Y] - E[x_i|Y])(E[x_j|Y] - E[x_j|Y])$ b/c $E[X] = E[E[X]]$

$= 0$

The matrix Σ is defined such that $\Sigma_{ij} = \text{cov}(x_i|Y, x_j|Y) \ \forall i, j$.
Since $x_i|Y$ and $x_j|Y$ are independent $\forall j \neq i$, by the claim,

$\Sigma_{ij} = \text{cov}(x_i|Y, x_j|Y) = 0 \ \forall j \neq i$.

Hence, the off diagonal elements of Σ equal 0, as desired.

□

$\boxed{3}$ $$P(C_0|x) = \frac{P(x|C_0)P(C_0)}{P(x|C_0)P(C_0) + P(x|C_1)P(C_1)}$$

a) Show $P(C_0|x) = \sigma(a)$ where a defined in terms of $P(x|C_0), P(x|C_1), P(C_0), P(C_1)$.

$$P(C_0|x) = \frac{P(x|C_0)P(C_0)}{P(x|C_0)\cancel{P(C_0)}\left(1 + \frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right)}$$

$$= \frac{1}{1+\exp(-a)} \quad \text{with } \exp(-a) = \frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}.$$

$$\exp(-a) = \frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)} \iff -a = \ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_0)P(C_0)}\right)$$

$$\iff a = \ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right)$$

i.e. $P(C_0|x) = \sigma(a)$ with $\underline{a = \ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right)}$.

b) Show that $a = w^T x + b$ for some w, b. Report w and b.

$$a = \ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right) = \ln\left(\frac{\frac{P(C_0)}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}{\frac{P(C_1)}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \ln\left(\frac{\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))}{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \left(-\frac{1}{2}(x^T\Sigma^{-1} - \mu_0^T\Sigma^{-1})(x-\mu_0) + \frac{1}{2}(x^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})(x-\mu_1)\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \Big(-\frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0)$$
$$+ \frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1)\Big)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \left(\frac{1}{2}x^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_0^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}x^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}x + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \left(\frac{1}{2}x^T(\Sigma^{-1}\mu_0 - \Sigma^{-1}\mu_1) + (\mu_0^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})\frac{1}{2}x + \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \left(\frac{1}{2}x^T\Sigma_1^{-1}(\mu_0-\mu_1) + (\mu_0^T-\mu_1^T)\Sigma_1^{-1}\frac{1}{2}x + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)\right)$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_1^{-1}\mu_0) + \frac{1}{2}x^T\Sigma_1^{-1}(\mu_0-\mu_1) + \frac{1}{2}(\mu_0-\mu_1)^T\Sigma_1^{-1}x$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_1^{-1}\mu_0) + \frac{1}{2}((\mu_0-\mu_1)^T\Sigma_1^{-1}x)^T + \frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}x$$

$$= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + ((\mu_0-\mu_1)^T\Sigma_1^{-1}x)$$

$\therefore$ $a = w^Tx + b$ with the definitions

$$\boxed{\begin{aligned} w &= \Sigma_1^{-1}(\mu_0-\mu_1) \\ b &= \ln\left(\frac{P(C_0)}{P(C_1)}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_1^{-1}\mu_0) \end{aligned}}$$

c) Show that $a = x^TAx + w^Tx + b$ for some A, w, b. Report A, w, b.

$$a = \ln\left(\frac{P(x|C_0)P(C_0)}{P(x|C_1)P(C_1)}\right) = \ln\left(\frac{\frac{P(C_0)}{(2\pi)^{n/2}|\Sigma_0|^{1/2}}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0))}{\frac{P(C_1)}{(2\pi)^{n/2}|\Sigma_1|^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}\right)$$

$$= \ln\left(\frac{P(C_0)\,|\Sigma_1|^{1/2}}{P(C_1)\,|\Sigma_0|^{1/2}}\right) + \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)$$

$$= \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) + \frac{1}{2}(x^T\Sigma_1^{-1} - \mu_1^T\Sigma_1^{-1})(x-\mu_1) - \frac{1}{2}(x^T\Sigma_0^{-1} - \mu_0^T\Sigma_0^{-1})(x-\mu_0)$$

$$\begin{aligned} = \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) &+ \frac{1}{2}(x^T\Sigma_1^{-1}x - x^T\Sigma_1^{-1}\mu_1 - \mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1) \\ &- \frac{1}{2}(x^T\Sigma_0^{-1}x - x^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0) \end{aligned}$$

$$\begin{aligned} = \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) &+ \frac{1}{2}x^T\Sigma_1^{-1}x - \frac{1}{2}x^T\Sigma_0^{-1}x + \frac{1}{2}x^T\Sigma_0^{-1}\mu_0 - \frac{1}{2}x^T\Sigma_1^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma_0^{-1}x - \frac{1}{2}\mu_1^T\Sigma_1^{-1}x \\ &+ \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) \end{aligned}$$

$$\begin{aligned} = \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) &+ \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) + x^T\frac{1}{2}(\Sigma_1^{-1} - \Sigma_0^{-1})x \\ &+ \frac{1}{2}x^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) + (\mu_0^T\Sigma_0^{-1} - \mu_1^T\Sigma_1^{-1})\frac{1}{2}x \end{aligned}$$

$$\begin{aligned} = x^T(\tfrac{1}{2}(\Sigma_1^{-1} - \Sigma_0^{-1}))x &+ (\mu_0^T\Sigma_0^{-1} - \mu_1^T\Sigma_1^{-1})x \\ &+ \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) \end{aligned}$$

$$\boxed{\begin{aligned} A &= \frac{1}{2}(\Sigma^{-1} - \Sigma_0^{-1}) \\ w &= \Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1 \\ b &= \ln\left(\frac{P(C_0)|\Sigma_1|^{1/2}}{P(C_1)|\Sigma_0|^{1/2}}\right) + \frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0) \end{aligned}}$$

4] Consider GDA on a training set $\{(x^i, y^i) ; i = \{1, \ldots, m\}\}$, $x^i \in \mathbb{R}^n$ al $y^i \in \{0,1\}$.

$\hookrightarrow \phi = P(y=1) \quad 1-\phi = P(y=0)$

$$P(x|y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)$$

$$P(x|y=1) = \text{"} \qquad \mu_1 \qquad \mu_1 \text{"}$$

$$\rightarrow L(\phi, \mu_0, \mu_1, \Sigma) = \ln P(x^1, \ldots, x^m, y^1, \ldots, y^m) = \ln \prod_{i=1}^{m} P(x^i|y^i)P(y^i).$$

a) we can explicitly write the following:

- $P(x^1, \ldots, x^m, y^1, \ldots, y^m)$

$$= \prod_{i=1}^{m} \left[\frac{1-\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y^i-\mu_0)^T\Sigma^{-1}(x^i-\mu_0)\right)\right]^{1-y^i} \left[\frac{\phi}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^i-\mu_1)^T\Sigma^{-1}(x^i-\mu_1)\right)\right]^{y^i}$$

- $L(\phi, \mu_0, \mu_1, \Sigma)$

$$= \sum_{i=1}^{m} \Big\{(1-y^i)\left[\ln(1-\phi) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^i-\mu_0)^T\Sigma^{-1}(x^i-\mu_0)\right]$$
$$+ y^i\left[\ln\phi - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(x^i-\mu_1)^T\Sigma^{-1}(x^i-\mu_1)\right]\Big\}$$

b) Find maximum likelihood estimate for ϕ. Show it is the "best", not "worst" estimate.

$$\frac{\partial L}{\partial \phi} = \frac{\partial}{\partial \phi}\left(\sum_{i=1}^{m}\{(1-y^i)\ln(1-\phi) + y^i \ln\phi\}\right)$$

$$= \sum_{i=1}^{m}\left[\frac{y^i}{\phi} - \frac{(1-y^i)}{1-\phi}\right] = \frac{1}{\phi}\sum_{i=1}^{m} y^i - \frac{1}{1-\phi}\sum_{i=1}^{m}(1-y^i)$$

$$= 0$$

$$\Longrightarrow \frac{1}{\phi}\sum_{i=1}^{m} y^i = \frac{1}{1-\phi}\sum_{i=1}^{m}(1-y^i)$$

$$\longleftrightarrow \sum_{i=1}^{m} y^i - \phi\sum_{i=1}^{m} y^i = \phi\sum_{i=1}^{m}(1-y^i)$$

$$\longleftrightarrow \phi\Big(\underbrace{\sum_{i=1}^{m} y^i + \sum_{i=1}^{m}(1-y^i)}_{m}\Big) = \sum_{i=1}^{m} y^i$$

$$\therefore \boxed{\text{MLE for } \phi \text{ is } \hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}}$$

$\longrightarrow$

If $\frac{\partial^2 L}{\partial \phi^2} < 0$, then the choice of ϕ falls at a maximum of L, and is thus the "best" choice of ϕ.

$$\frac{\partial^2 L}{\partial \phi^2} = \frac{\partial}{\partial \phi}\left(\sum_{i=1}^{m} \left[y^i \phi^{-1} - (1-y^i)(1-\phi)^{-1} \right] \right)$$

$$= \sum_{i=1}^{m} \Big[\underbrace{-y^i \phi^{-2}}_{\leq 0} + \underbrace{(-1)(1-y^i)(1-\phi)^{-2}}_{\leq 0} \Big]$$

b/c $y^i, (1-y^i), \phi^{-2}, (1-\phi)^{-2}$ are ≥ 0.

But, at least 1 term is < 0 b/c $y^i \phi^{-2}$ or $(1-y^i)(1-\phi)^{-2}$ is > 0 for at least 1 datapoint.

$$\Rightarrow \sum_{i=1}^{m} \left[-y^i \phi^{-2} + (-1)(1-y^i)(1-\phi)^{-2} \right] < 0$$

$\Rightarrow \frac{\partial^2 L}{\partial \phi^2} < 0$, so ϕ is the "best" as it maximizes L.

c) Find maximum likelihood estimate for μ_0. Show it is the "best".

$$\frac{\partial L}{\partial \mu_0} = \frac{\partial}{\partial \mu_0}\left(\sum_{i=1}^{m} \left\{ -\frac{1}{2}(1-y^i)(x^i-\mu_0)^T \Sigma^{-1} (x^i - \mu_0) \right\} \right)$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \left\{ (1-y^i) \frac{\partial}{\partial \mu_0} (x^i-\mu_0)^T \Sigma^{-1} (x^i-\mu_0) \right\}$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \left\{ (1-y^i)(\Sigma^{-1} + \Sigma^{-T})(x^i-\mu_0)(-1) \right\}$$ b/c $\frac{\partial}{\partial w}(w^T A w) = (A + A^T) w$

$$= \frac{1}{2}(\Sigma^{-1} + \Sigma^{-T}) \sum_{i=1}^{m} (1-y^i)(x^i - \mu_0)$$

$$= 0$$

$$\Rightarrow \sum_{i=1}^{m} (1-y^i)(x^i-\mu_0) = 0$$

$$\longleftrightarrow \sum_{i=1}^{m} (1-y^i) x^i - \sum_{i=1}^{m} (1-y^i) \mu_0 = 0$$

$$\longleftrightarrow \mu_0 = \frac{\sum_{i=1}^{m} (1-y^i) x^i}{\sum_{i=1}^{m} (1-y^i)}$$

$\therefore$ MLE for μ_0 is

$$\boxed{\hat{\mu}_0 = \frac{1}{\sum_{i=1}^{m}(1-y^i)} \sum_{i=1}^{m} x_i (1-y^i)}$$

→

Need to show that the estimate for μ_0 is the "best" estimate.

μ_0 is the best estimate if it maximizes L.

μ_0 maximizes L if the Hessian Matrix of L wrt μ_0 is negative definite.

↳ Goal: Show Hessian of L wrt μ_0 is negative definite,
then we're done.

$$\frac{\partial^2 L}{\partial \mu_0^2} = \frac{\partial}{\partial \mu_0}\left(\frac{1}{2}\left(\Sigma^{-1} + \Sigma^{-T}\right) \sum_{i=1}^{m} (1-y^i)(x^i - \mu_0)\right)$$

$$= -\frac{1}{2}(2\Sigma^{-1}) \sum_{i=1}^{m} (1-y^i)$$

$$= -\Sigma^{-1} \sum_{i=1}^{m} (1-y^i)$$

- $\sum_{i=1}^{m} (1-y^i)$ is a positive scalar
- Σ is positive definite for multivariate Gaussian distribution.
  - $\Rightarrow \Sigma^{-1}$ is positive definite
  - $\Rightarrow -\Sigma^{-1}$ is negative definite
- Hence, $-\Sigma^{-1} \sum_{i=1}^{m} (1-y^i)$ is negative definite.

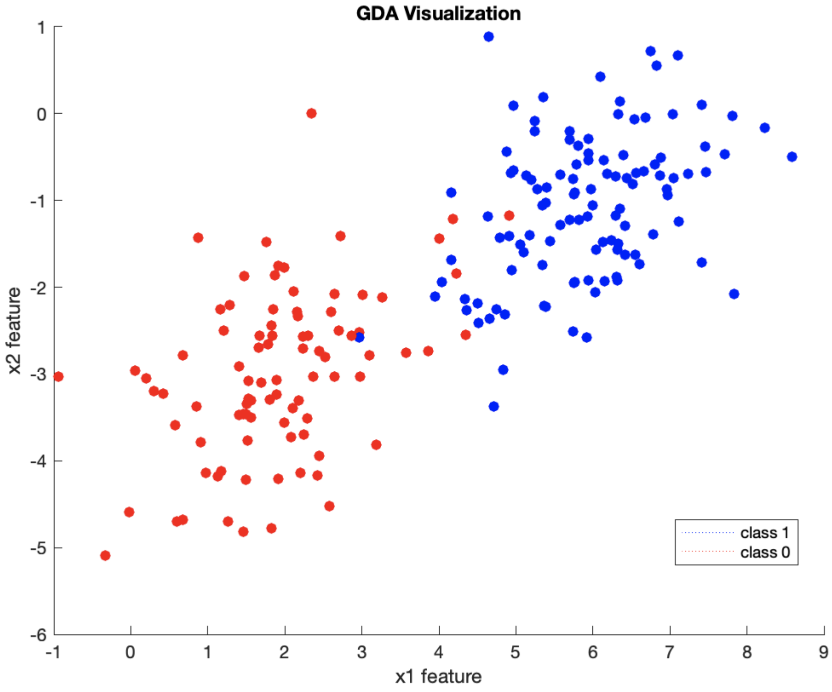
This shows that $\frac{\partial^2 L}{\partial \mu_0^2} = -\Sigma^{-1} \sum_{i=1}^{m} (1-y^i)$ is negative definite, so were done.

Zack Berger

Question 5

Part A - Visualization

By inspection of the following plot, the data is NOT linearly separable…



Part B - Maximum Likelihood Estimates

Using the GDA model, the following estimates were learned from the data:

P(y = 0) = 0.4450

μ0 = $[1.9195, -2.9972]^T$

μ1 = $[5.8982, -1.0793]^T$

Σ =

$$\begin{matrix} 1.0181 & 0.3887 \\ 0.3887 & 0.8036 \end{matrix}$$

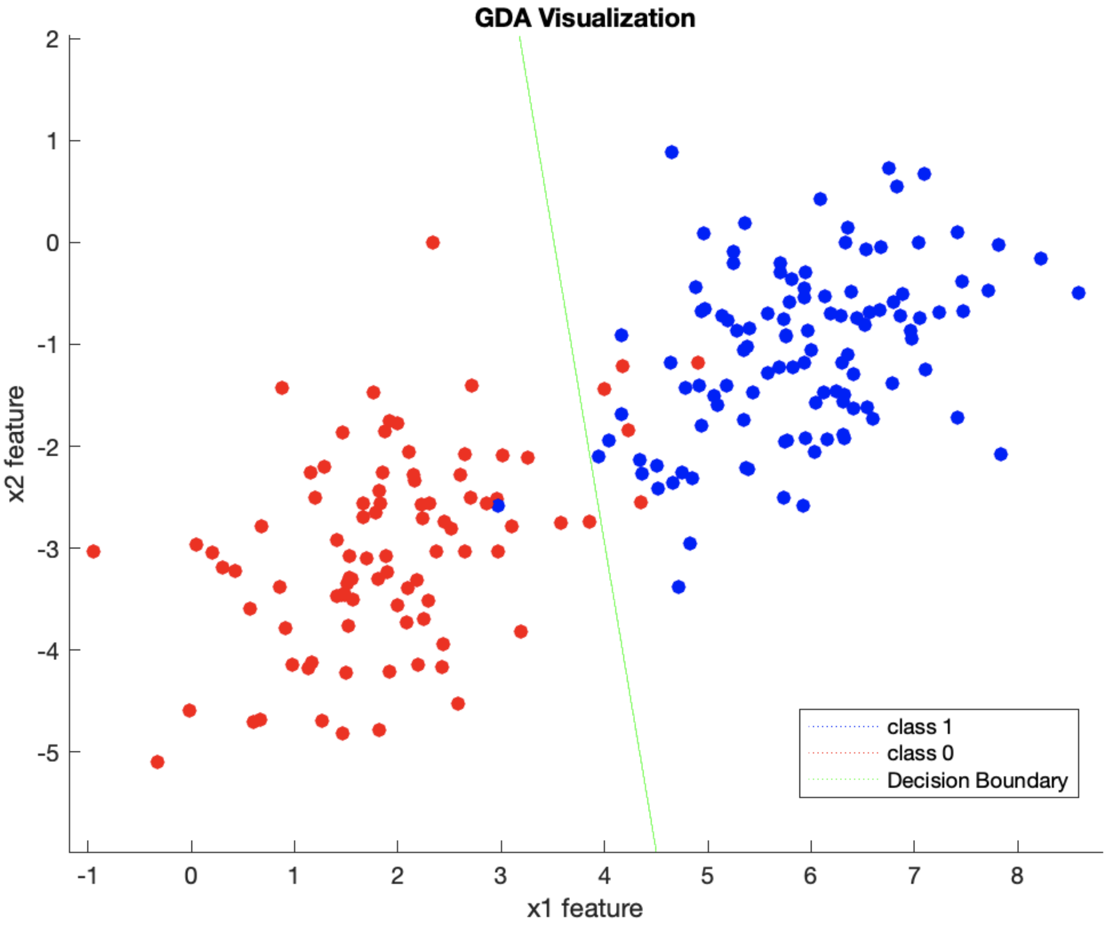
Part C - Decision Boundary

Decision boundary $w^T x + b = 0$ parameterized by

$w = [-3.6755, -0.6090]^T$

$b = 12.9050$

Plot of decision boundary…

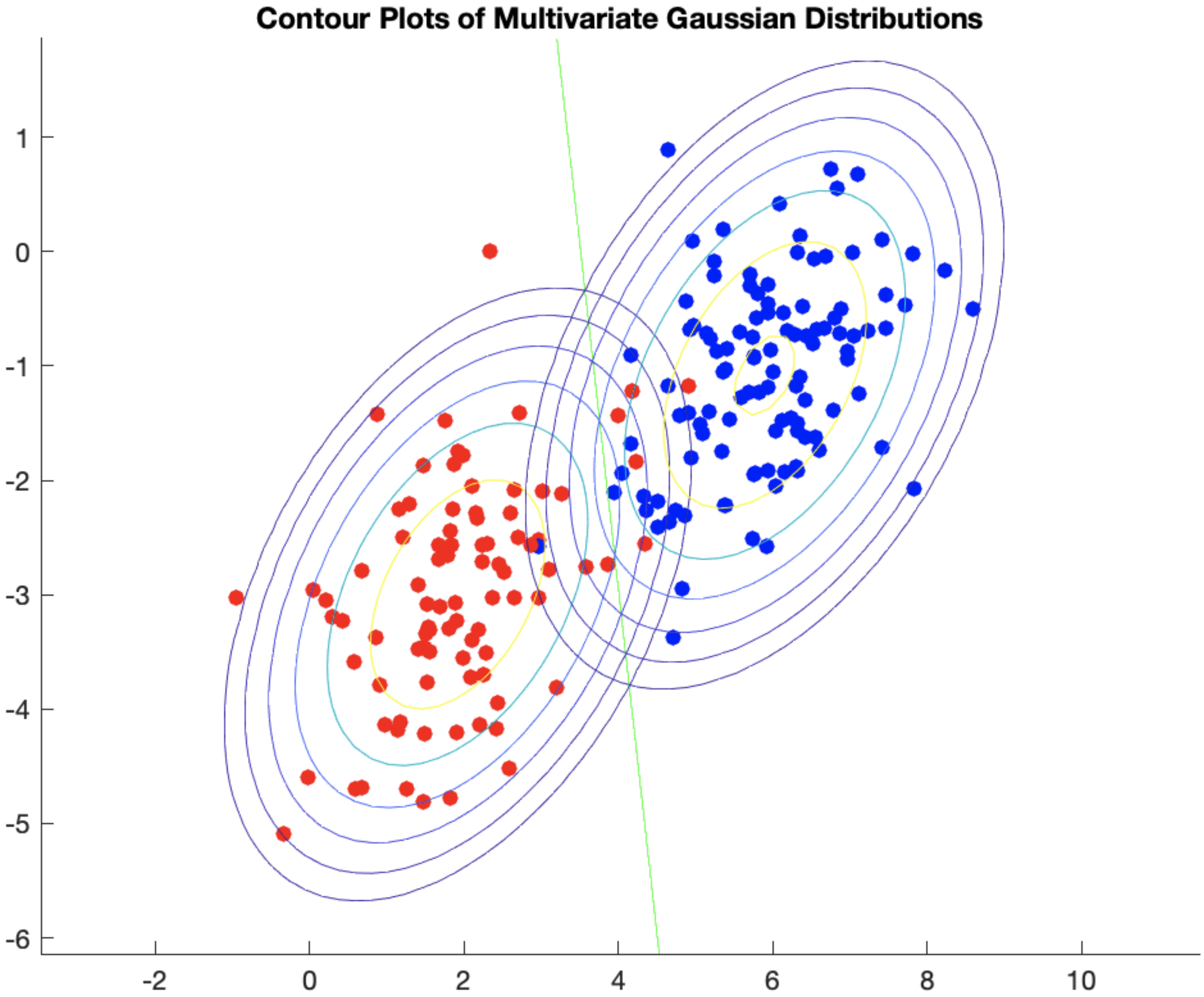


Part D - Contour Plots

The red points represent class 0, whereas the blue points represent class 1. The decision boundary is in green. The P(X, Y = 0) contour plot is graphed to the left of the boundary, and the P(X, Y = 1) contour plot is graphed to the right of the boundary.

The decision boundary passes through the points where the two distributions have equal probabilities. This is because of the following logic:

> The decision boundary is defined $w^T x + b = a$
> It has its solutions when a = ln(P(X, Y = 0) / P(X, Y = 1)) = 0
> This occurs when P(X, Y = 0) / P(X, Y = 1) = 1
> In such a case, P(X, Y = 0) = P(X, Y = 1)



Code Attachment

```
% Import data
data = readtable('/Users/zackberger/Desktop/ML/HW/HW_6/data.csv');
data = data{:,:};

% Separate data into column vectors
x = data(:,[1,2]);
x1 = data(:,1);
x2 = data(:,2);
y = data(:,3);
num_data = numel(y);


% Learn the mu parameters and P(y=0)
mu_zero = zeros([2 1]);
mu_one = zeros([2 1]);
num_class_one = 0;

for i = 1 : num_data
    if y(i) == 1
        mu_one = mu_one + x(i, :).';
        num_class_one = num_class_one + 1;
    else
        mu_zero = mu_zero + x(i, :).';
    end
end

num_class_zero = num_data - num_class_one;

mu_one = (1 / num_class_one) * mu_one;
mu_zero = (1 / num_class_zero) * mu_zero;

prob_y_zero = num_class_zero / (num_class_zero + num_class_one);


% Learn the sigma parameter
sigma = zeros(2);
for i = 1 : num_data
    if y(i) == 1
        sigma = sigma + (x(i, :).' - mu_one)*(x(i, :).' - mu_one).';
    else
        sigma = sigma + (x(i, :).' - mu_zero)*(x(i, :).' - mu_zero).';
    end
end

sigma = (1/num_data) * sigma;


% Find linear decision boundary
inv_sigma = inv(sigma);

w = inv_sigma*(mu_zero - mu_one);
b = log(prob_y_zero / (1 - prob_y_zero)) + 0.5*(mu_one.'*inv_sigma*mu_one - mu_zero.'*inv_sigma*mu_zero);

hold on
```

```
% Scatter plot feature vectors
for i = 1: numel(x1)
    if y(i) == 1
        scatter(x1(i), x2(i), 'filled', 'b')
    else
        scatter(x1(i), x2(i), 'filled', 'r')
    end
end

% Plot decision boundary
x = -1:1/10000:10;
y = (-1*b - w(1)*x) / w(2);
plot(x,y, 'g');

title("GDA Visualization");
ylim([-6 2])
% xlabel("x1 feature");
% ylabel("x2 feature");

% Create legend
L(1) = plot(nan, nan, 'b:');
L(2) = plot(nan, nan, 'r:');
L(3) = plot(nan, nan, 'g:');
% legend(L, {'class 1', 'class 0', 'Decision Boundary'})


% Plot contours of the two multivariate Gaussian distributions
f = @(x,y) prob_y_zero * (1 / (2*pi)) * (1 / sqrt(det(sigma))) * exp(-0.5 * ([x;y] - mu_zero).' * inv_sigma * ([x;y] - mu_zero));
g = @(x,y) (1 - prob_y_zero) * (1 / (2*pi)) * (1 / sqrt(det(sigma))) * exp(-0.5 * ([x;y] - mu_one).' * inv_sigma * ([x;y] - mu_one));

fcontour(f, 'LevelList', logspace(-3,-1,7));
fcontour(g, 'LevelList', logspace(-3,-1,7));

hold off

title("Contour Plots of Multivariate Gaussian Distributions");
```