Probabilistic Graphical Models : Homework 1

Raphael Avalos
raphael@avalos.fr

10/10/18

1 Formulas

1.1 Exercise 1

We have N samples (x_i, y_i) $\boldsymbol{\pi} = (\pi_1, ..., \pi_M), \boldsymbol{\theta} = (\theta_{1,1}, ...\theta_{M,K})$
$a_m =\mid \{i \mid z_i = m, \forall i \in [1, N]\} \mid$, $b_{m,k} =\mid \{i \mid z_i = m \text{ and } x_i = k, \forall i \in [1, N]\} \mid$

$$\tilde{\pi}_m = \frac{a_m}{N}$$
$$\tilde{\theta}_{m,k} = \frac{b_{m,k}}{N}$$

Moreover $p(y = 1 \mid x)$ have the same form of logistic regression.

1.2 Exercise 2

1.2.1 LDA

$$\tilde{\omega} = \frac{n}{N}$$
$$\tilde{\mu}_0 = \frac{1}{n} \sum_{\substack{i=1, \\ y_i=0}}^{N} x_i$$
$$\tilde{\mu}_1 = \frac{1}{N-n} \sum_{\substack{i=1, \\ y_i=1}}^{N} x_i$$
$$\tilde{\Sigma} = \frac{1}{N}(\sum_{\substack{i=1, \\ y_i=0}}^{N} (x_i - \mu_0)^T (x_i - \mu_0) + \sum_{\substack{i=1, \\ y_i=1}}^{N} (x_i - \mu_1)^T (x_i - \mu_1))$$

1.2.2 QDA

We have N samples and $n =\mid \{i, y_i = 0, \forall i \in [1, N]\} \mid$

$$\tilde{\omega} = \frac{n}{N}$$
$$\tilde{\mu}_0 = \frac{1}{n} \sum_{\substack{i=1, \\ y_i=0}}^{N} x_i$$
$$\tilde{\mu}_1 = \frac{1}{N-n} \sum_{\substack{i=1, \\ y_i=1}}^{N} x_i$$
$$\tilde{\Sigma}_0 = \frac{1}{n} \sum_{\substack{i=1, \\ y_i=0}}^{N} (x_i - \mu_0)^T (x_i - \mu_0)$$
$$\tilde{\Sigma}_1 = \frac{1}{N-n} \sum_{\substack{i=1, \\ y_i=1}}^{N} (x_i - \mu_1)^T (x_i - \mu_1)$$

2 Dataset A

Figure 1: LDA

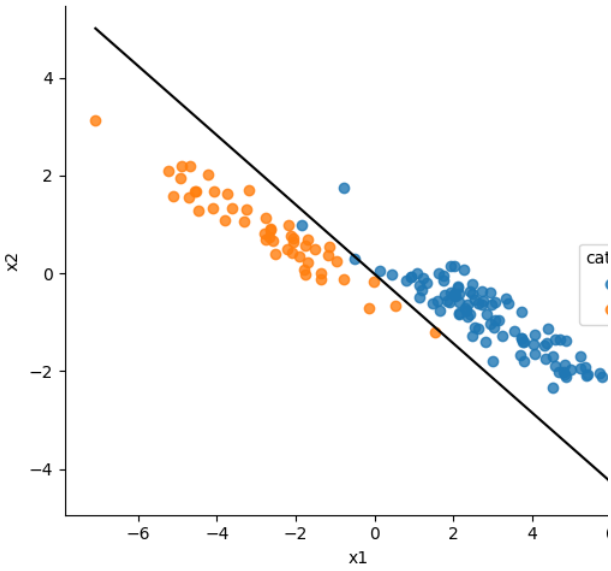


Figure 2: IRLS

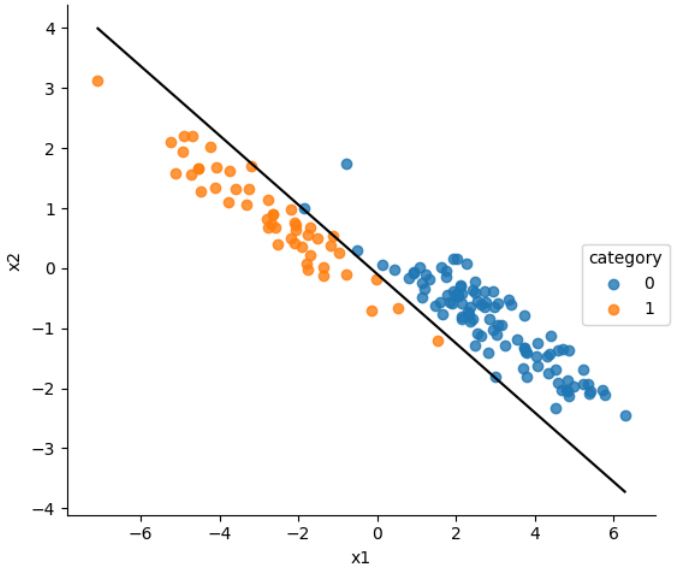


Figure 3: Linear Regression

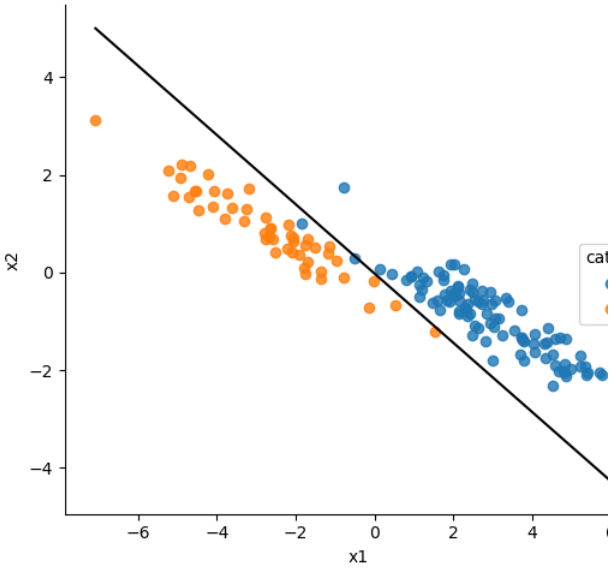
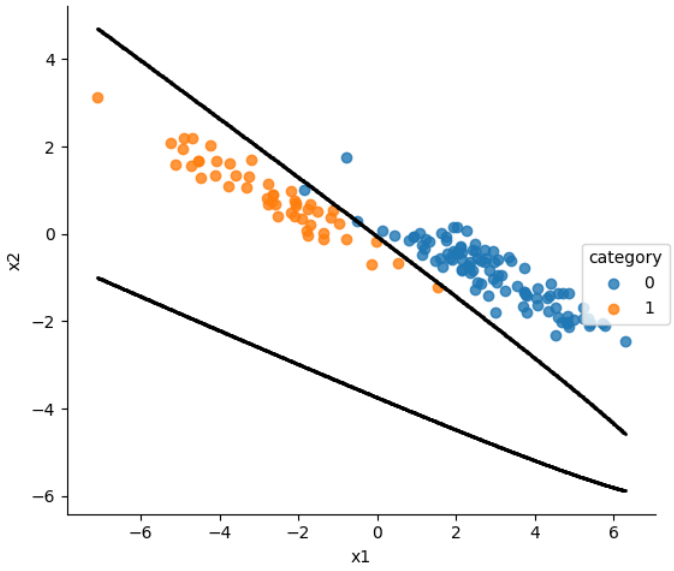


Figure 4: QDA



Results (% of success)

	Train	Test
LDA	98.7	98.0
IRLS	100	96.6
LR	98.7	97.9
QDA	99.3	98.0

In this dataset blue points and orange ones looks like they share the same covarience matrix Σ. This is confirmed by the great results of LDA which has the same success rate in the testing dataset of QDA which waives the equality of Σ_1 and Σ_2. IRLS acheived a 100% success rate in the trainig set but it clearly overfitted on the training sample (it makes the worst score on the testing set). Finally linear regression performs quite well (almost the same results as LDA) which is not a surprise since the two sets are almost linearly separable.

3 Dataset B

Figure 5: LDA

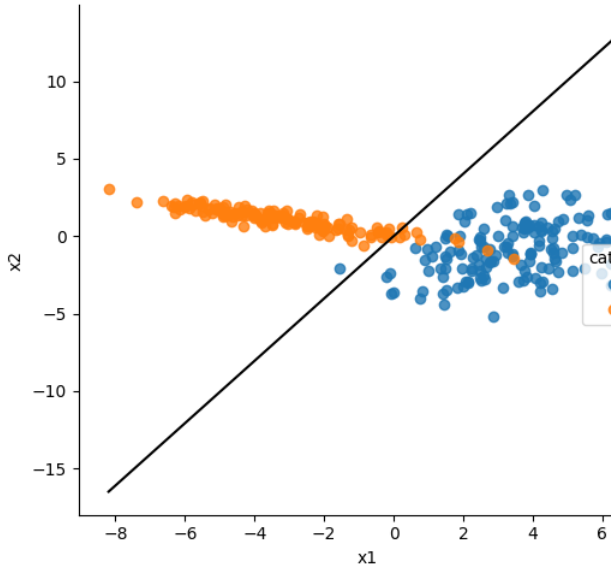


Figure 6: IRLS

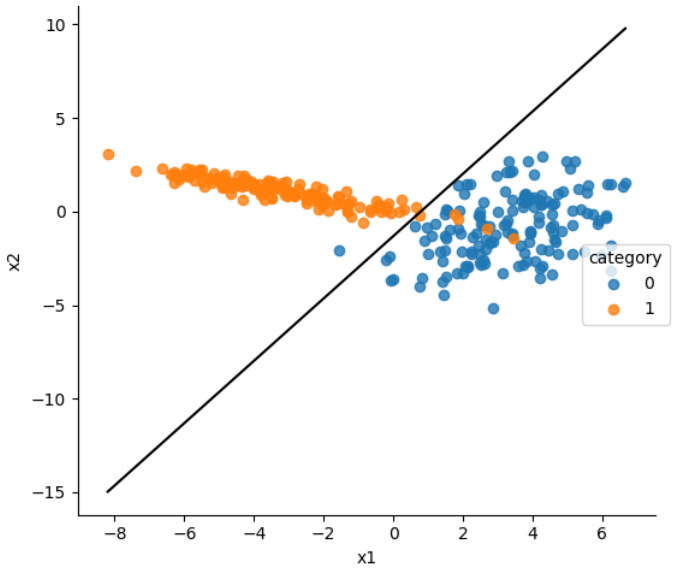


Figure 7: Linear Regression

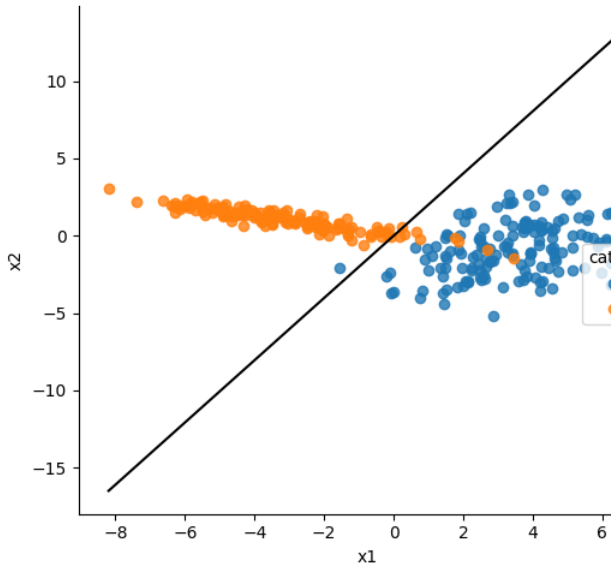
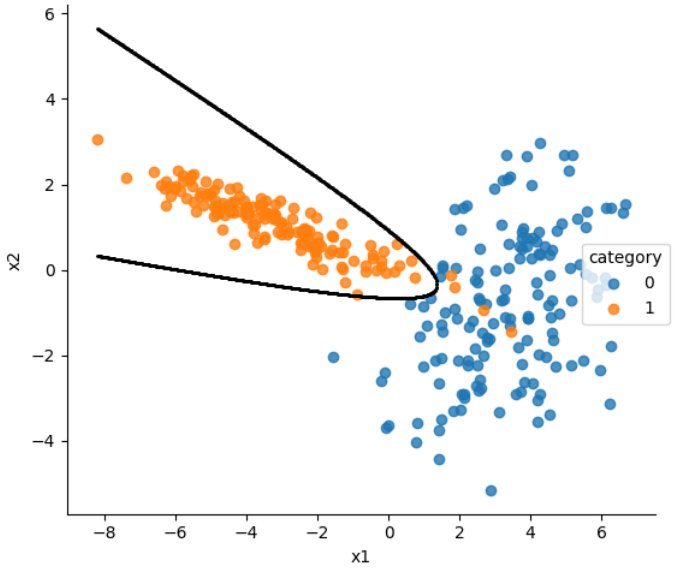


Figure 8: QDA



Results (% of success)

	Train	Test
LDA	97.0	95.9
IRLS	98.0	95.7
LR	97.0	95.9
QDA	98.7	98.0

In this dataset, the two distribution doesn't have the same covariance matrix therefore LDA is not the good algorithm to use. However if the two distribution follows indeed a normal distribution with different mean and covariance matrix QDA should perform really well. This assumption is verified by the preformance of QDA (it outperforms the other models). Linear regression and LDA gave the same boundary between both distribution; IRLS returns a shift of that same boundary which provide a better result in training but a worst results than the other two methods in testing.

4 Dataset C

Figure 9: LDA

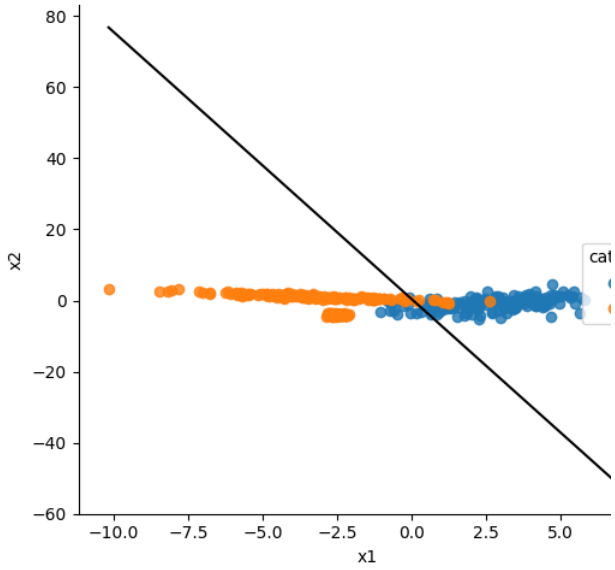


Figure 10: IRLS

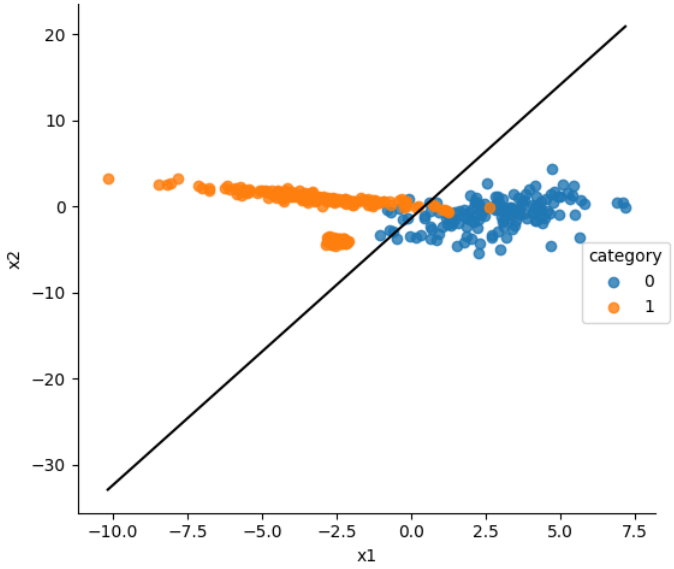


Figure 11: Linear Regression

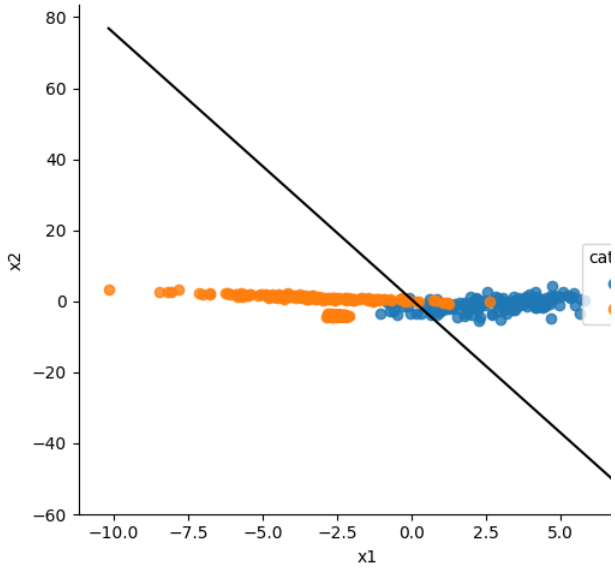
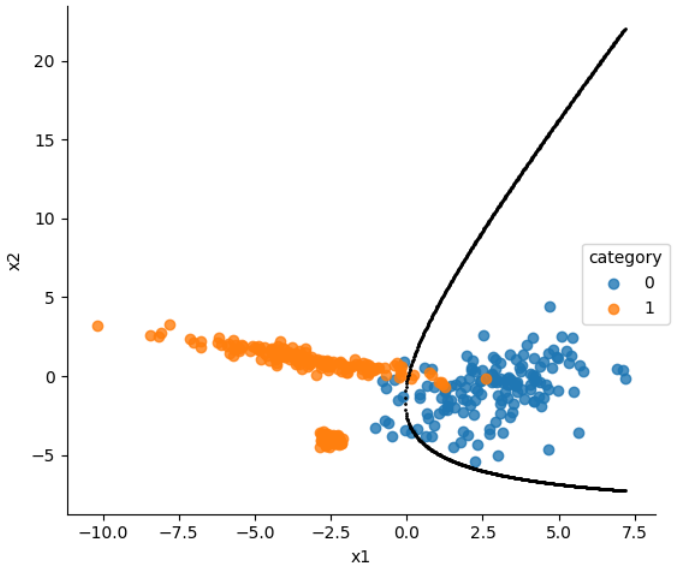


Figure 12: QDA



Results (% of success)

	Train	Test
LDA	94.5	95.8
IRLS	96.0	97.7
LR	94.5	95.8
QDA	94.8	96.2

In this dataset the orange distribution doesn't seem to follow a normal distribution because their are two modes. Therefore LDA and QDA eventhough they will provide a good approximation won't be the best option for that task. The two distribution aren't linearly separable; therefore the linear regression won't provide the best results. This leaves us with the logistic regression performed by the IRLS which as a matter of fact outperform the other models and manage to get the generalize surprisingly well the separation of the two distribution (its testing score is higher than its training one).

5 Proof

5.1 Exercise 1

We have N samples (x_i, y_i) $\boldsymbol{\pi} = (\pi_1, ..., \pi_M), \boldsymbol{\theta} = (\theta_{1,1}, ...\theta_{M,K})$
$a_m =\mid \{i \mid z_i = m, \forall i \in [1, N]\} \mid$, $b_{m,k} =\mid \{i \mid z_i = m \text{ and } x_i = k, \forall i \in [1, N]\} \mid$

$$\begin{aligned} l(\boldsymbol{\pi}, \boldsymbol{\theta}) &= \sum_{i=0}^{n} log(p(x_i, z_i)) \\ &= \sum_{i=0}^{n} log(p(x_i \mid z_i)p(z_i)) \\ &= \sum_{i=0}^{n} (log(\theta_{z_i,x_i}) + log(\pi_{x_i})) \end{aligned}$$

$l(\boldsymbol{\pi}, \boldsymbol{\theta})$ is concave. We want to minimize $-l(\boldsymbol{\pi}, \boldsymbol{\theta})$ subjected to $\sum_{k=1}^{K} \pi_k = 1$ and $\sum_{k=1}^{K} \sum_{m=1}^{M} \theta_{m,k} = 1$ Lets introduce the langrangian.

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda_1, \lambda_2) = -(\sum_{i=0}^{n} (log(\theta_{z_i,x_i}) + log(\pi_{x_i}))) + \lambda_1(\sum_{k=1}^{K} \pi_k - 1) + \lambda_2(\sum_{k=1}^{K} \sum_{m=1}^{M} \theta_{m,k} - 1)$$

The Slaters constraint qualification are trivialy verified and therefore the problem has strong duality property. Therefore we have

$$\min_{\boldsymbol{\pi}, \boldsymbol{\theta}} -l(\boldsymbol{\pi}, \boldsymbol{\theta}) = \max_{\lambda_1, \lambda_2} L(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda_1, \lambda_2)$$

Moreover the lagrangian is convex with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$

$$\frac{\partial L}{\partial \pi_m} = 0 \Rightarrow \tilde{\pi}_m = \frac{a_m}{\lambda_1}$$

$$\frac{\partial L}{\partial \theta_{m,k}} = 0 \Rightarrow \tilde{\theta}_{m,k} = \frac{b_{m,k}}{\lambda_2}$$

Using the constrains we can calculate λ_1, λ_2.

$$\tilde{\pi}_m = \frac{a_m}{N}$$
$$\tilde{\theta}_{m,k} = \frac{b_{m,k}}{N}$$

5.2 Exercise 2

5.2.1 Generative model LDA

We have N samples and $n =\mid \{i, y_i = 0, \forall i \in [1, N]\} \mid$

$$\begin{aligned} l(\omega, \Sigma, \mu_0, \mu_1) &= \sum_{i=1}^{N} log(p(x_i, y_i)) \\ &= \sum_{i=1}^{N} log(p(x_i \mid y_i)p(y_i)) \\ &= \sum_{\substack{i=1, \\ y_i=0}}^{N} log(p(x_i \mid y_i = 0)) + nlog(\omega) + \sum_{\substack{i=1, \\ y_i=0}}^{N} log(p(x_i \mid y_i = 1)) + (N - n)log(1 - \omega) \\ &= -\frac{Nd}{2} log(2\pi) + \frac{N}{2} log(\mid \Sigma^{-1} \mid) - \sum_{\substack{i=1, \\ y_i=0}}^{N} \frac{1}{2}(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) \\ &- \sum_{\substack{i=1, \\ y_i=1}}^{N} \frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + nlog(\omega) + (N - n)log(1 - \omega) \end{aligned}$$

This log likelyhood is not concave in $(\omega, \Sigma, \mu_0, \mu_1)$. It is concave in (ω, μ_0, μ_1) with Σ fixed.

$$\nabla_\omega l = \frac{n}{\omega} - \frac{N-n}{1-\omega}$$

$\nabla_\omega l = 0$ gives us :

$$\tilde{\omega} = \frac{n}{N}$$

Calculating the gradient in μ_0, μ_1 and equalating it to 0 gives us.

$$\tilde{\mu_0} = \frac{1}{n} \sum_{\substack{i=1, \\ y_i=0}}^{N} x_i$$

$$\tilde{\mu_1} = \frac{1}{N-n} \sum_{\substack{i=1, \\ y_i=1}}^{N} x_i$$

Let us now differentiate l w.r.t. Σ^{-1}.
Let $A = \Sigma^{-1}$, $\Sigma_0 = \frac{1}{n} \sum_{\substack{i=1, \\ y_i=0}}^{N} (x_i - \mu_0)^T (x_i - \mu_0)$,
$\Sigma_1 = \frac{1}{N-n} \sum_{\substack{i=1, \\ y_i=1}}^{N} (x_i - \mu_1)^T (x_i - \mu_1)$
We have :

$$\begin{aligned} l(\omega, \Sigma, \mu_0, \mu_1) = &- \frac{Nd}{2} log(2\pi) + \frac{N}{2} log(\mid \Sigma^{-1} \mid) - \frac{1}{2} \text{Trace}(A(n\Sigma_0 + (N-n)\Sigma_1) \\ &+ nlog(\omega) + (N-n)log(1-\omega) \\ \nabla_A l = &\frac{N}{2} A^{-1} - \frac{1}{2}(n\Sigma_0 + (N-n)\Sigma_1) \end{aligned}$$

Which leads to

$$\tilde{\Sigma} = \frac{n}{N}\Sigma_0 + \frac{N-n}{N}\Sigma_1$$

We have found a unique stationnary point for the likelyhood. To be sure it is a maximum we would have to calculate the Hessian.
Now we will calculate the $p(y = 1 \mid x)$.

$$p(y = 1 \mid x) = \frac{p(x \mid y = 1)p(y = 1)}{p(x)}$$

$$\begin{aligned} log(\frac{p(y=1 \mid x)}{p(y=0 \mid x)}) &= log(\frac{1-\omega}{\omega}) - \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)\Sigma^{-1}(x-\mu_0) \\ log(\frac{p(y=1 \mid x)}{p(y=0 \mid x)}) &= log(\frac{1-\omega}{\omega}) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) + x^T\Sigma^{-1}(\mu_1 - \mu_0) \\ p(y=1 \mid x) &= \frac{1}{1 + \frac{w}{1-w} \exp\left(\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)\right) \exp\left(-x^T\Sigma^{-1}(\mu_1 - \mu_0)\right)} \end{aligned}$$

It is of the form

$$p(y = 1 \mid x) = \frac{1}{1 + \exp\left(-(x^T a + \alpha)\right)}$$

It is the formula of logistic regression

5.2.2 QDA model

We have N samples and $n = \mid \{i, y_i = 0, \forall i \in [1, N]\} \mid$

$$\begin{aligned}
l(\omega, \Sigma_0, \Sigma_1, \mu_0, \mu_1) &= \sum_{i=1}^{N} log(p(x_i, y_i)) \\
&= \sum_{i=1}^{N} log(p(x_i \mid y_i)p(y_i)) \\
&= \sum_{\substack{i=1,\\ y_i=0}}^{N} log(p(x_i \mid y_i = 0)) + nlog(\omega) + \sum_{\substack{i=1,\\ y_i=0}}^{N} log(p(x_i \mid y_i = 1)) + (N-n)log(1-\omega) \\
&= nlog(\omega) - \frac{Nd}{2}log(2\pi) + \frac{n}{2}log(\mid \Sigma_0^{-1} \mid) - \sum_{\substack{i=1,\\ y_i=0}}^{N} \frac{1}{2}(x_i - \mu_0)^T \Sigma_0^{-1}(x_i - \mu_0) \\
&+ (N-n)log(1-\omega) + \frac{N-n}{2}log(\mid \Sigma_1^{-1} \mid) - \sum_{\substack{i=1,\\ y_i=1}}^{N} \frac{1}{2}(x_i - \mu_1)^T \Sigma_1^{-1}(x_i - \mu_1)
\end{aligned}$$

This log likelyhood is not concave in $(\omega, \Sigma_0, \Sigma_1, \mu_0, \mu_1)$. It is concave in (ω, μ_0, μ_1) with Σ_0 and Σ_1 fixed. We obtain like in the previous questions.

$$\begin{aligned}
\tilde{\omega} &= \frac{n}{N} \\
\tilde{\mu}_0 &= \frac{1}{n} \sum_{\substack{i=1,\\ y_i=0}}^{N} x_i \\
\tilde{\mu}_1 &= \frac{1}{N-n} \sum_{\substack{i=1,\\ y_i=1}}^{N} x_i
\end{aligned}$$

Differentiating l w.r.t. Σ_0^{-1} with the rest fixed and equalizing to 0 (and then doing the same with Σ_1^{-1}) gives us.

$$\begin{aligned}
\tilde{\Sigma}_0 &= \frac{1}{n} \sum_{\substack{i=1,\\ y_i=0}}^{N} (x_i - \mu_0)^T (x_i - \mu_0) \\
\tilde{\Sigma}_1 &= \frac{1}{N-n} \sum_{\substack{i=1,\\ y_i=1}}^{N} (x_i - \mu_1)^T (x_i - \mu_1)
\end{aligned}$$

We did not provide the calculations because it is amlost the same as above.