Homework 1

1 Learning in discrete graphical models

Z and X are discrete random variables with respectively M and K different values, such that $p(Z = m) = \pi_m$ and $p(X = k|Z = m) = \theta_{mk}$.

Consider $\Pi = \{\pi_m\}_{1 \leq m \leq M}$ and $\Theta = \{\theta_{mk}\}_{1 \leq m \leq M, 1 \leq k \leq K}$

Consider $D = \{(x_i, z_i), i \in \{1, ..n\}\}$ an i.i.d sample of n observations. We compute the complete log-likelihood of the discrete model as

$$
\begin{aligned}
l(\Pi, \Theta) &= log(p(X, Z)) = log(\prod_{i=1}^{n} p(x_i, z_i)) \\
l(\Pi, \Theta) &= \sum_{i=1}^{n} \log(p(z_i)p(x_i|z_i)) \\
&= \sum_{i=1}^{n} \log(p(z_i)) + \sum_{i=1}^{n} \log(p(x_i|z_i)) \\
&= \sum_{i=1}^{n} \log(\prod_{m=1}^{M} \pi_m^{z_i^m}) + \sum_{i=1}^{n} \log(\prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{z_i^m x_i^k}) \\
&= \sum_{i=1}^{n} \sum_{m=1}^{M} z_i^m \log(\pi_m) + \sum_{i=1}^{n} \sum_{m=1}^{M} \sum_{k=1}^{K} z_i^m x_i^k \log(\theta_{mk})
\end{aligned}
$$

where $x_i^k \in \{0, 1\}$ with $x_i^k = 1$ iff $x_i = k$ and $z_i^m \in \{0, 1\}$ with $z_i^m = 1$ iff $z_i = m$, for $k \in \{1, ..., K\}$ and $m \in \{1, ..., M\}$.

In order to find the MLE estimators, we need to solve this constrained problem:

$$
\max_{\Pi, \Theta} l(\Pi, \Theta)
$$

$$
s.t. \begin{cases} \sum_{m=1}^{M} \pi_m = 1, \forall k \in \{1, ..., K\} \\ \sum_{k=1}^{K} \theta_{mk} = 1, \forall m \in \{1, ..., M\} \\ \pi_m \geq 0, \forall m \in \{1, M\} \\ \theta_{mk} \geq 0, \forall m \in \{1, M\}, \forall k \in \{1, K\} \end{cases} \tag{1}
$$

We consider Lagrangian multipliers $\lambda^{(1)} \in \mathbb{R}^M$ positive, $\lambda^{(2)} \in \mathbb{R}^{MxK}$ positive, $\lambda^{(3)} \in \mathbb{R}$ and $\lambda^{(4)} \in \mathbb{R}^M$.

The Lagrangian of our problem is the following:

$$\begin{aligned}L(\Pi,\Theta,\Lambda) &= -l(\Pi,\Theta) - \sum_{m=1}^{M}\lambda_m^{(1)}\pi_m - \sum_{m=1}^{M}\sum_{k=1}^{K}\lambda_{mk}^{(2)}\theta_{mk} + \lambda^{(3)}(\sum_{m=1}^{M}\pi_m - 1) + \sum_{m=1}^{M}\lambda_m^{(4)}(\sum_{k=1}^{K}\theta_{mk} - 1) \\ &= -\sum_{m=1}^{M}(\sum_{i=1}^{n} z_i^m)\log(\pi_m) - \sum_{k=1}^{K}\sum_{m=1}^{M}(\sum_{i=1}^{n} z_i^m x_i^k)\log(\theta_{mk}) - \sum_{m=1}^{M}\lambda_m^{(1)}\pi_m - \sum_{m=1}^{M}\sum_{k=1}^{K}\lambda_{mk}^{(2)}\theta_{mk} \\ &+ \lambda^{(3)}(\sum_{m=1}^{M}\pi_m - 1) + \sum_{m=1}^{M}\lambda_m^{(4)}(\sum_{k=1}^{K}\theta_{mk} - 1)\end{aligned}$$

Since $\sum_{i=1}^{n} z_i^m \geq 0$, $\forall$m$\in \{1,...M\}$ and $\sum_{i=1}^{n} z_i^m x_i^k \geq 0$, $\forall (m,k) \in \{1,...M\}x\{1,...K\}$, -l is convex. It is therefore a convex optimization problem.

Besides, it is trivial that there exist Π and Θ such that $\pi_m > 0,\ \forall m \in \{1,...M\}$, $\theta_{mk} > 0,\ \forall (m,k) \in \{1,...M\}x\{1,...K\}$, $\sum_{m=1}^{M}\pi_m = 1$ and $\sum_{k=1}^{K}\theta_{mk} = 1,\ \forall m \in \{1,...M\}$.

Hence, by Slater's constraint qualification, the problem has strong duality propriety. We can write

$$\min_{\Pi,\Theta} -l(\Pi,\Theta) = \max_{\Lambda}\min_{\Pi,\Theta} L(\Pi,\Theta,\Lambda)$$

By setting the gradient of the Lagrangian with respect to π_m and θ_{mk} to 0, and by using the KKT conditions, we obtain:

$$\mathrm{P}_1 \begin{cases} \frac{\partial L}{\partial \pi_m} = -\frac{\sum_{i=1}^{n} z_i^m}{\hat{\pi}_m} - \lambda_m^{(1)} + \lambda^{(3)} = 0, \forall m \in \{1, M\} \\ \frac{\partial L}{\partial \theta_{mk}} = -\frac{\sum_{i=1}^{n} z_i^m x_i^k}{\hat{\theta}_{mk}} - \lambda_{mk}^{(2)} + \lambda_m^{(4)} = 0, \forall m \in \{1, M\}, \forall k \in \{1, K\} \\ \lambda_{mk}^{(2)}\hat{\theta}_{mk} = 0, \forall m \in \{1, M\}, \forall k \in \{1, K\} \\ \lambda_m^{(1)}\hat{\pi}_m = 0, \forall m \in \{1, M\} \\ \sum_{m=1}^{M}\hat{\pi}_m = 1 \\ \sum_{k=1}^{K}\hat{\theta}_{mk} = 1 \end{cases}$$

We solve the system and obtain

$$\mathrm{P}_2 \begin{cases} \lambda_m^{(1)} = 0, \forall m \in \{1, M\} \\ \lambda_{mk}^{(2)} = 0, \forall m \in \{1, M\}, \forall k \in \{1, K\} \\ \hat{\pi}_m = \frac{\sum_{i=1}^n z_i^m}{\lambda^{(3)}} \\ \hat{\theta}_{mk} = \frac{\sum_{i=1}^n z_i^m x_i^k}{\lambda_m^{(4)}} \\ \sum_{m=1}^M \hat{\pi}_m = 1 \\ \sum_{k=1}^K \hat{\theta}_{mk} = 1 \end{cases}$$

As a result, we obtain

$$\lambda_m^{(1)} = 0, \forall m \in \{1, M\}$$

$$\lambda_{mk}^{(2)} = 0, \forall m \in \{1, M\}, \forall k \in \{1, K\}$$

$$\lambda^{(3)} = n$$

$$\lambda_m^{(4)} = \sum_{i=1}^n z_i^m, \forall m \in \{1, M\}$$

The maximum likelihood estimator for Π and Θ is

$$\hat{\pi}_m = \frac{\sum_{i=1}^n z_i^m}{n}, \forall m \in \{1, M\}$$

$$\hat{\theta}_{mk} = \frac{\sum_{i=1}^n z_i^m x_i^k}{\sum_{i=1}^n z_i^m}, \forall k \in \{1, K\}$$

2 Linear classification

Let $D = \{(x_i, y_i)\}_{1 \leq in}$ be an i.i.d sample of observations, where $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$.

2.1 Generative model LDA

2.1.1 Maximum likelihood estimators

We assume the following

$$y \sim \mathcal{B}(\pi) \quad and \quad x|y = i \sim \mathcal{N}(\mu_i, \Sigma).$$

We start by computing the complete log-likelihood

$$\begin{aligned}
l(\pi, \mu_0, \mu_1, \Sigma) &= \sum_{i=1}^{n} \log(p(x_i, y_i)) \\
&= \sum_{i=1}^{n} \log(p(x_i|y_i)p(y_i)) \\
&= \sum_{i=1}^{n} \log(p(x_i|y_i)) + \sum_{i=1}^{n} \log(p(y_i)) \\
&= \sum_{i=1}^{n} y_i \left[\log(p(x_i|y_i = 1)) + \log(\pi)\right] + (1 - y_i) \left[\log(p(x_i|y_i = 0)) + \log(1 - \pi)\right] \\
&= \sum_{i=1}^{n} y_i \left[\log(\mathcal{N}(\mu_1, \Sigma)) + \log(\pi)\right] + (1 - y_i) \left[\log(\mathcal{N}(\mu_0, \Sigma)) + \log(1 - \pi)\right] \\
&= \sum_{i=1}^{n} y_i \left[-\frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \log(\pi)\right] \\
&+ \sum_{i=1}^{N} (1 - y_i) \left[-\frac{1}{2}(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) + \log(1 - \pi)\right] - \frac{n}{2} \log(2\pi \det(\Sigma))
\end{aligned}$$

Now, we compute the gradient of the log likelihood with respect to π, μ_0, μ_1, Σ and set it to zero.

$$\begin{cases}
\nabla_\pi l(\pi, \mu_0, \mu_1, \Sigma) & = \frac{\sum_{i=1}^{n} y_i}{\pi} + \frac{\sum_{i=1}^{n} (1 - y_i)}{1 - \pi} \\
\nabla_{\mu_0} l(\pi, \mu_0, \mu_1, \Sigma) & = \sum_{i=1}^{n} (1 - y_i) \Sigma^{-1}(x_i - \mu_0) \\
\nabla_{\mu_1} l(\pi, \mu_0, \mu_1, \Sigma) & = \sum_{i=1}^{n} y_i \Sigma^{-1}(x_i - \mu_1) \\
\nabla_\Sigma l(\pi, \mu_0, \mu_1, \Sigma) & = -\nabla_\Sigma \left[\sum_{i=1}^{n} \frac{1}{2} y_i (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \frac{1}{2}(1 - y_i)(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) \right. \\
& \left. + \frac{n}{2} \log(\det(\Sigma)) \right.
\end{cases} \tag{2}$$

Consider the following notation

$$\begin{cases}
A = \Sigma^{-1} \\
\tilde{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i)(x_i - \mu_0)(x_i - \mu_0)^T \\
\tilde{\Sigma}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i (x_i - \mu_1)(x_i - \mu_1)^T
\end{cases}$$

We can rewrite the gradient with respect to Σ as

$$\nabla_\Sigma l(\pi, \mu_0, \mu_1, \Sigma) = -\nabla_\Sigma \left[\frac{n}{2} \operatorname{Tr}(A(\tilde{\Sigma}_0 + \tilde{\Sigma}_1)) - \frac{n}{2} \log(\det(A))\right] \tag{3}$$

Note that

$$\nabla_A \operatorname{Tr}(A\Sigma) = \Sigma$$
$$\nabla_A \log(\det A) = A^{-1}$$

Therefore (3) becomes

$$\nabla_\Sigma l(\pi, \mu_0, \mu_1, \Sigma) = -\frac{n}{2}(\tilde{\Sigma}_0 + \tilde{\Sigma}_1) + \frac{n}{2}A^{-1}$$

Hence, we obtain

$$\begin{cases} \hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i \\ \hat{\mu}_0 = \frac{\sum_{i=1}^{n} x_i(1-y_i)}{\sum_{i=1}^{n}(1-y_i)} \\ \hat{\mu}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} \\ \hat{\Sigma} = \tilde{\Sigma}_0 + \tilde{\Sigma}_1 \end{cases}$$

2.1.2 Conditional Distribution Form

We apply Bayes rule

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x)} \\ &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \\ &= \frac{\frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2})\pi}{\frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(\frac{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2})(1-\pi) + \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2})\pi} \\ &= \frac{\frac{\pi}{1-\pi} \exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2})}{1 + \frac{\pi}{1-\pi} \exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2})} \\ &= \frac{\exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2} + \log(\frac{\pi}{1-\pi}))}{1 + \exp(\frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2} + \log(\frac{\pi}{1-\pi}))} \end{aligned}$$

As Σ is symmetric, we can rewrite the argument of exp as

$$\begin{aligned} C &= \frac{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)}{2} + \log(\frac{\pi}{1-\pi}) \\ &= (\mu_0^T - \mu_1^T)\Sigma^{-1}x - \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} + \log(\frac{\pi}{1-\pi}) \end{aligned}$$

Therefore, we obtain

$$p(y=1|x) = \frac{\exp((\mu_0^T - \mu_1^T)\Sigma^{-1}x - \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} + \log(\frac{\pi}{1-\pi}))}{1 + \exp((\mu_0^T - \mu_1^T)\Sigma^{-1}x - \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} + \log(\frac{\pi}{1-\pi}))}$$

which is of the same form as in logistic regression

$$p(y=1|x) = \frac{\exp(a^Tx + b)}{1 + \exp(a^Tx + b)}$$

where

$$a = \Sigma^{-1}(\mu_0 - \mu_1)$$
$$b = -\frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} + \log(\frac{\pi}{1-\pi})$$

2.1.3 p(y=1|x)= 0.5 Line Representation

$p(y=1|x) = 0.5$ is equivalent to writing $a^Tx + b = 0$, and therefore

$$\left(\Sigma^{-1}(\mu_1 - \mu_0)\right)^T x - \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} + \log(\frac{\pi}{1-\pi}) = 0$$

2.1.4 Data Classification Graphical Representation

The following figure represents LDA model results obtained on both train and test set of the three datasets. The line corresponding to p(y=1|x)=0.5 is presented aswell in the figures below.

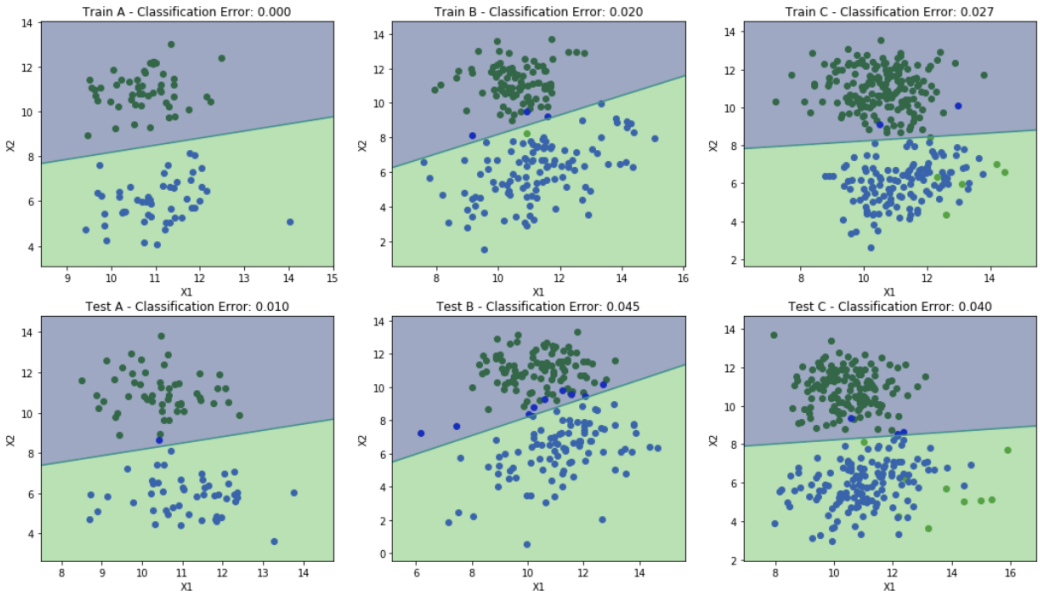

Figure 1: LDA results

2.2 Logistic regression

2.2.1 Numerical values

Model	Intercept	Weight 1	Weight 2
A	208.303	9.235	-35.935
B	13.43	1.842	-3.714
C	18.807	-0.277	-1.914

Table 1: Logistic Regression weights

2.2.2 Data Classification Graphical Representation

The following figure represents Logistic Regression model results obtained on both train and test set of the three datasets. The line corresponding to p(y=1|x)=0.5 is presented aswell in the figures below.

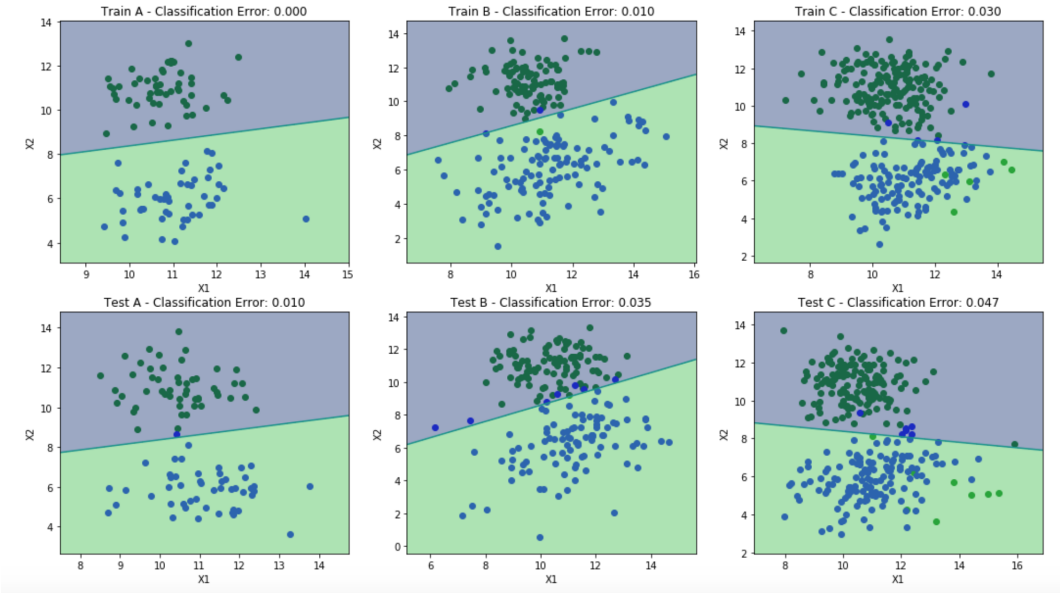

Figure 2: Logistic regression results

2.3 Linear regression

2.3.1 Numerical values

Model	Intercept	Weight 1	Weight 2
A	1.383	0.0558	-0.176
B	0.882	0.0826	-0.147
C	1.64	0.0167	-0.159

Table 2: Linear Regression weights

2.3.2 Data Classification Graphical Representation

The following figure represents Linear Regression model results obtained on both train and test set of the three datasets. The line corresponding to p(y=1|x)=0.5 is presented aswell in the figures below.

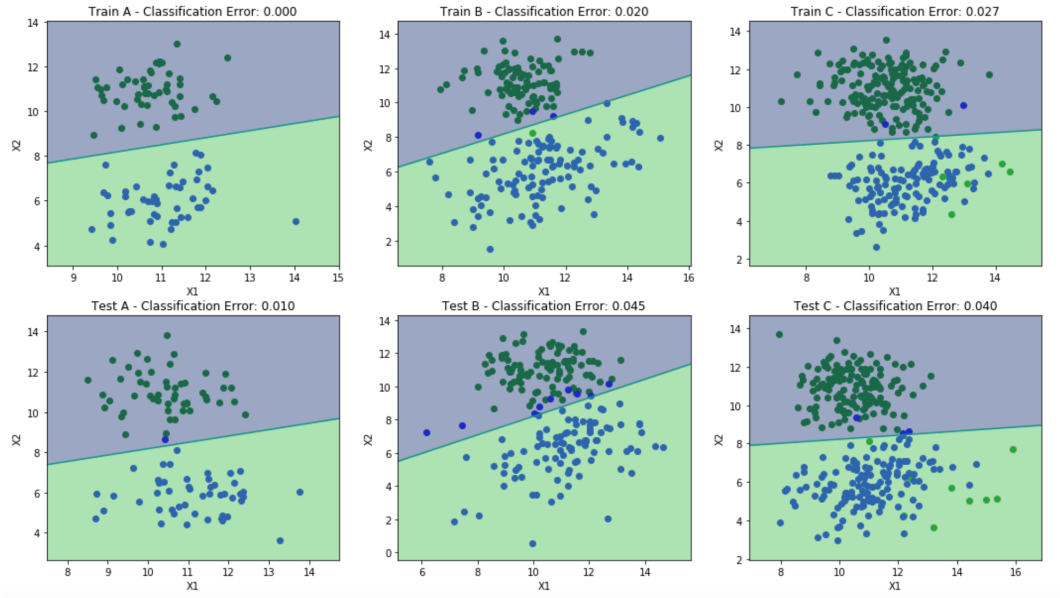

Figure 3: Linear regression results

2.4 Application

2.4.1 Missclassification error

Dataset	Train error	Test error
A	0.00	0.01
B	0.02	0.045
C	0.027	0.04

Table 3: LDA model

Dataset	Train error	Test error
A	0.00	0.01
B	0.01	0.035
C	0.03	0.047

Table 4: Logistic Regression model

Dataset	Train error	Test error
A	0.00	0.01
B	0.02	0.045
C	0.027	0.04

Table 5: Linear Regression model

2.4.2 Error analysis

Dataset A -

- The dataset A is drawn from two multivariate gaussian distributions, with different means and similar covariance matrices.
- The assumptions made on the LDA model hold, and therefore is supposed to have better results than the rest of the models.
- All three models along with Logistic and Linear Regression have similar and good results, as it turns out that the train and test data are separable, which explains why the training error is equal to 0.

Dataset B -

- The dataset B is drawn from two multivariate gaussian distributions, with different means and different covariance matrices.
- The assumptions made on the LDA model do not hold this time.
- LDA and both linear and logistic regressions lead to linear classifiers. They have a lower precision compared to their performances in the first dataset, simply because the data is not linearly separable.
- Logistic Regression have the best results on the dataset.

Dataset C -

- The dataset C is drawn aswell from two multivariate gaussian distributions, with different means and different covariance matrices, with some noisy data points that fall into the opposite class.
- The three models have similar results. It is not possible to separate the noisy data points that sneaked into one of the classes with a linear model.

2.5 QDA

Now, we assume the following

$$y \sim \mathcal{B}(\pi) \quad and \quad x|y = i \sim \mathcal{N}(\mu_i, \Sigma_i).$$

We compute again the complete log likelihood $l(\pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1)$ and obtain the following parameters estimator

$$\begin{aligned}
\hat{\pi} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
\hat{\mu}_0 &= \frac{\sum_{i=1}^{n} x_i(1-y_i)}{\sum_{i=1}^{n}(1-y_i)} \\
\hat{\mu}_1 &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i} \\
\hat{\Sigma}_0 &= \frac{\sum_{i=1}^{n}(1-y_i)(x_i-\hat{\mu}_0)(x_i-\hat{\mu}_0)^T}{\sum_{i=1}^{n}(1-y_i)} \\
\hat{\Sigma}_1 &= \frac{\sum_{i=1}^{n} y_i(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^T}{\sum_{i=1}^{n} y_i}
\end{aligned}$$

The conditional distribution of $y = 1|x$ can be rewritten as:

$$p(y=1|x) = \sigma\left(\frac{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}{2} + \log\left(\frac{\pi}{1-\pi}\right)\right)$$

Therefore, $p(y=1|x) = 0.5$ is equivalent to:

$$\frac{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)}{2} + \log\left(\frac{\pi}{1-\pi}\right) = 0$$

which has a conic form

$$x^T A x + b^T x + c = 0$$

where

$$\begin{aligned}
A &= \frac{1}{2}(\hat{\Sigma}_0^{-1} - \hat{\Sigma}_1^{-1}) \\
b &= \hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\Sigma}_0^{-1}\hat{\mu}_0 \\
c &= \frac{1}{2}(\hat{\mu}_0^T\hat{\Sigma}_0^{-1}\hat{\mu}_0 - \hat{\mu}_1^T\hat{\Sigma}_1^{-1}\hat{\mu}_1) + \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)
\end{aligned}$$

2.5.1 Numerical Values

1. Dataset A -

   $\hat{\pi} = 0.48$

   $\hat{\mu}_0 = (10.73, 10.94)$ $\hat{\mu}_1 = (11.03, 5.99)$

   $\hat{\Sigma}_0 = \begin{pmatrix} 0.46 & 0.098 \\ 0.098 & 0.71 \end{pmatrix}$ $\hat{\Sigma}_1 = \begin{pmatrix} 0.72 & 0.18 \\ 0.18 & 0.934 \end{pmatrix}$

2. Dataset B -

   $\hat{\pi} = 0.55$

   $\hat{\mu}_0 = (10.58, 11.17)$ $\hat{\mu}_1 = (11.25, 6.095)$

   $\hat{\Sigma}_0 = \begin{pmatrix} 0.76 & 0.053 \\ 0.053 & 1.107 \end{pmatrix}$ $\hat{\Sigma}_1 = \begin{pmatrix} 2.36 & 1.23 \\ 1.23 & 2.84 \end{pmatrix}$

3. Dataset C -

   $\hat{\pi} = 0.417$

   $\hat{\mu}_0 = (10.62, 10.84)$ $\hat{\mu}_1 = (11.18, 6.04)$

   $\hat{\Sigma}_0 = \begin{pmatrix} 1.285 & -0.433 \\ -0.433 & 1.826 \end{pmatrix}$ $\hat{\Sigma}_1 = \begin{pmatrix} 1.267 & 0.457 \\ 0.457 & 1.44 \end{pmatrix}$

2.5.2 Data Classification Graphical Representation

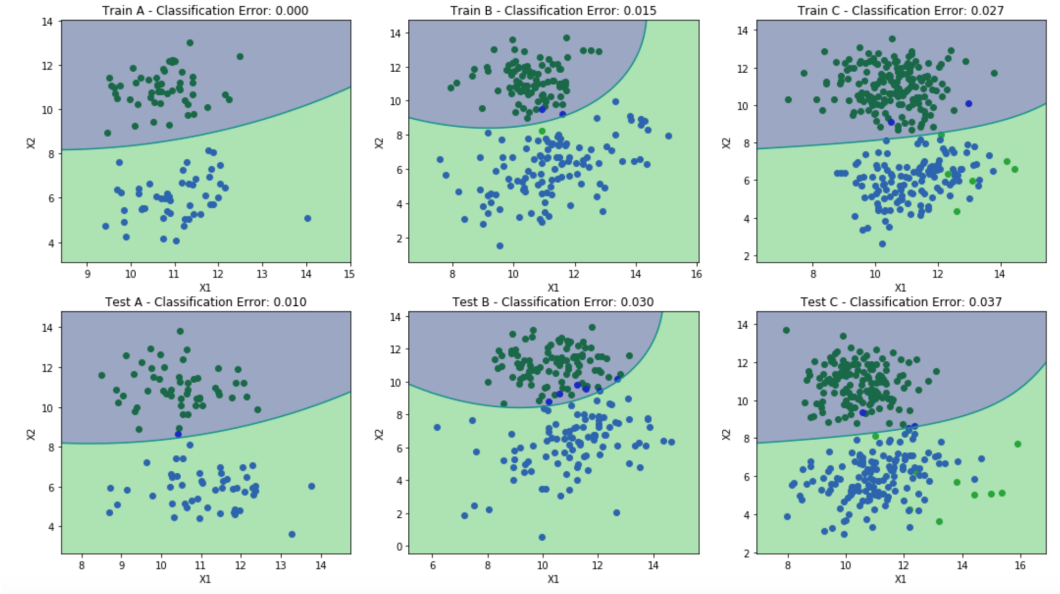

Figure 4: QDA results

2.5.3 Missclassification error

Dataset	Train error	Test error
A	0.00	0.01
B	0.015	0.03
C	0.027	0.037

Table 6: QDA model

2.5.4 QDA Error Analysis

- Dataset A - QDA model has the same performance as the latter three models on this dataset. As we have mentioned before, data points are linearly separable, so using a quadratic model do not make a difference in the classification task.
- Dataset B - QDA has the same performance as Logistic Regression on the train set of the dataset B, except it has the lowest error on the test set. QDA generalizes better than the other competing models.
- Dataset C - QDA has a slightly better performance among the tested models. The model do not seem to be able to separate the noisy data points from the class 0 that fall near to data points from the class 1.