Probabilistic Graphical Models

Ayman Chaouki - Otmane Sakhi
First Homework

Exercice 1 : Learning in Discrete Graphical Models

MLE computation on (Z, X) to estimate $\mathbf{\Pi} = \{\pi_1, ..., \pi_M\}$ and $\mathbf{\Theta} = \{\theta_{m,k}\}_{1 \leq m \leq M, 1 \leq k \leq K}$

$$\begin{cases} \hat{\theta}_{m,k} = \frac{1}{\sum_{i=1}^{N} z_{i,m}} \sum_{i=1}^{N} z_{i,m} x_{i,k} \quad \forall k \in \{1, ..., K\}, \forall m \in \{1, ..., M\} \\ \\ \hat{\pi}_m = \frac{1}{N} \sum_{i=1}^{N} z_{i,m} \quad \forall m \in \{1, ..., M\} \end{cases}$$

Exercice 2.1 (a) : LDA Formulas

MLE computation on (X, Y) to estimate μ_0, μ_1, Σ_0, Σ_1 and π

$$\begin{cases} \hat{\pi} &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ \hat{\mu_0} &= \frac{1}{\sum_{i=1}^{N} (1-y_i)} \sum_{i=1}^{N} (1 - y_i) x_i \\ \hat{\mu_1} &= \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i x_i \\ \hat{\Sigma} &= \frac{\sum_{i=1}^{N} y_i}{N} \tilde{\Sigma_1} + \frac{\sum_{i=1}^{N} (1-y_i)}{N} \tilde{\Sigma_0} \end{cases}$$

With $\tilde{\Sigma_0} = \frac{1}{\sum_{i=1}^{N} (1-y_i)} \sum_{i=1}^{N} (1 - y_i) \left(x_i - \mu_1\right) \left(x_i - \mu_1\right)^T$, $\tilde{\Sigma_1} = \frac{1}{\sum_{i=1}^{N} (y_i)} \sum_{i=1}^{N} y_i \left(x_i - \mu_1\right) \left(x_i - \mu_1\right)^T$

Inference $p\left(y = 1 | x\right) = \sigma\left(a + b^T x\right)$ Where

$$\begin{cases} a &= \frac{1}{2} \left(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1\right) + log\left(\frac{\pi}{1-\pi}\right)) \\ b &= \Sigma^{-1} \left(\mu_1 - \mu_0\right) \end{cases}$$

Exercice 2.5 (a) : QDA Formulas

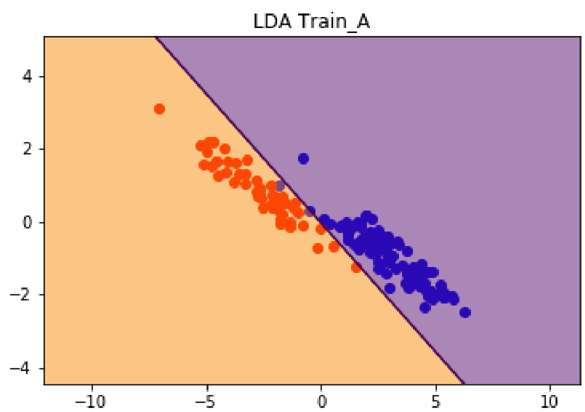
MLE computation on (X, Y) to estimate μ_0, μ_1, Σ_0, Σ_1 and π

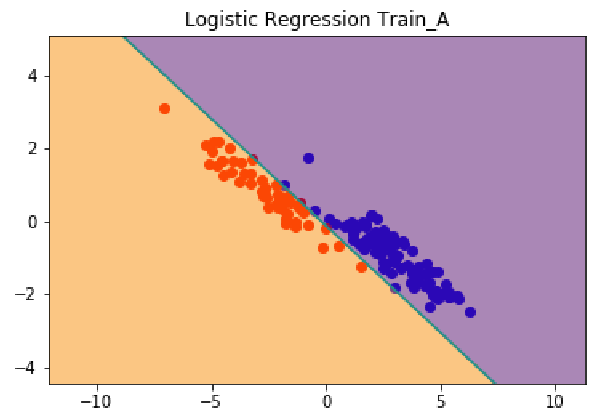
$$\begin{cases} \hat{\pi} &= \frac{1}{N} \sum_{i=1}^{N} y_i \\ \hat{\mu_0} &= \frac{1}{\sum_{i=1}^{N} (1-y_i)} \sum_{i=1}^{N} (1 - y_i) x_i \\ \hat{\mu_1} &= \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i x_i \\ \hat{\Sigma_0} &= \tilde{\Sigma_0} = \frac{1}{\sum_{i=1}^{N} (1-y_i)} \sum_{i=1}^{N} (1 - y_i) \left(x_i - \mu_1\right) \left(x_i - \mu_1\right)^T \\ \hat{\Sigma_1} &= \tilde{\Sigma_1} = \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i \left(x_i - \mu_1\right) \left(x_i - \mu_1\right)^T \end{cases}$$

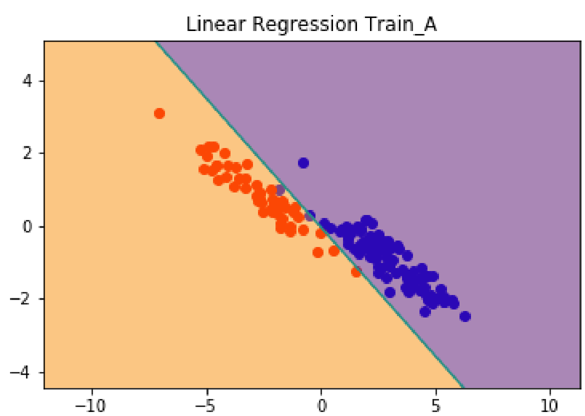
Inference $p\left(y = 1 | x\right) = \sigma\left(a + b^T x + x^T c x\right)$ Where

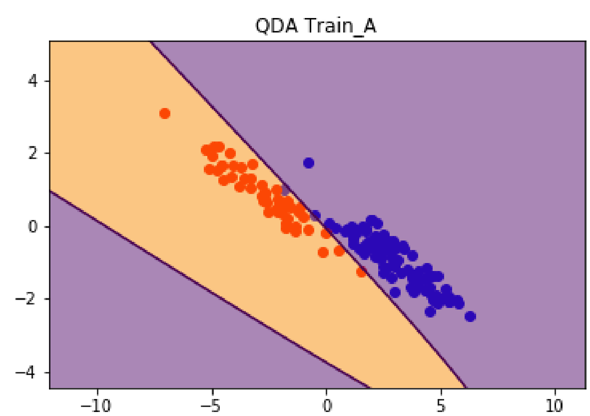
$$\begin{cases} a &= \frac{1}{2} \left(\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1\right) + log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2} log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ b &= \Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0 \\ c &= \frac{1}{2} \left(\Sigma_0^{-1} - \Sigma_1^{-1}\right) \end{cases}$$

Dataset A







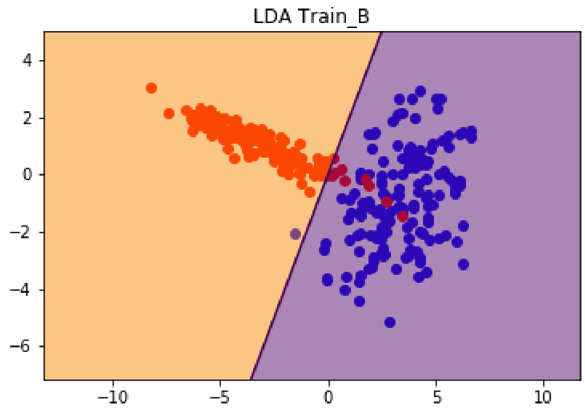


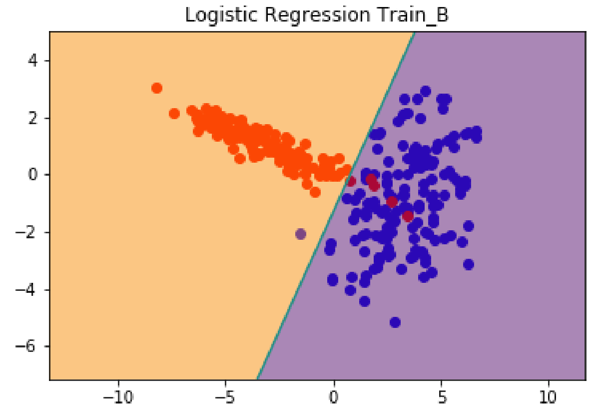
Miss-classification Error

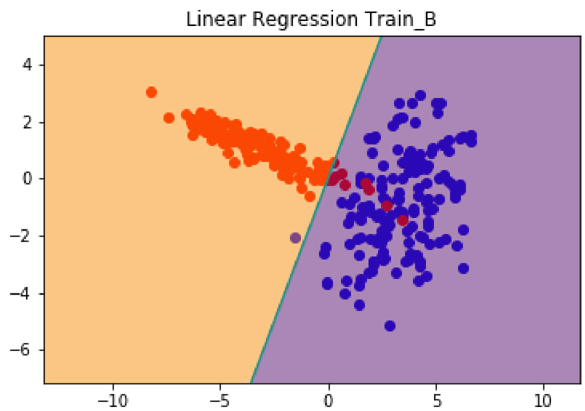
Model	Training	Test
LDA	1.33%	**2.0%**
Logistic Regression	**0%**	3.53%
Linear Regression	1.33%	2.07%
QDA	0.67%	**2.0%**

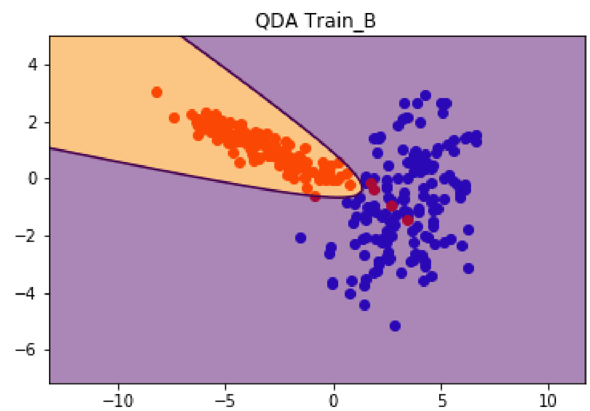
- The Dataset A is drawn, conditionned on the class, from two multivariate gaussians with different means and pretty close covariance matrices. The LDA assumption holds and we expect it to have the best results along with QDA.
- Few training observations give the lead to models with good assumptions about the data distribution.
- The Logistic Regression overfits the 150 points of the training data and doesn't generalize well. The LDA/QDA have the best generalization results as expected, and the linear regression has decent performances on both splits.

Dataset B







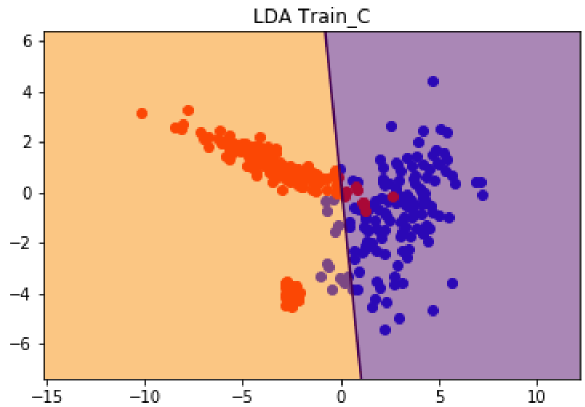


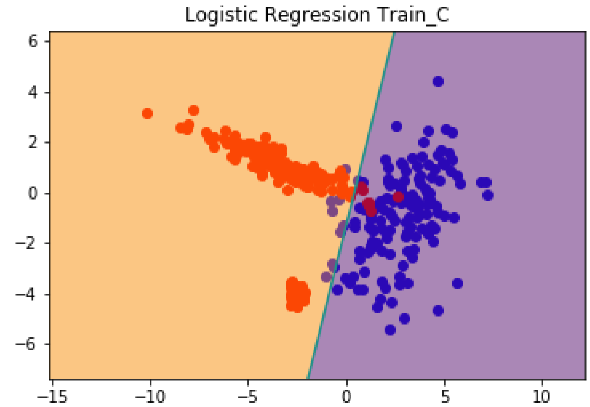
Miss-classification Error

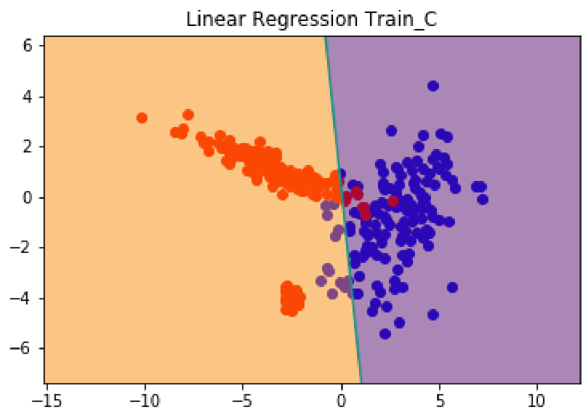
Model	Training	Test
LDA	3.0%	4.15%
Logistic Regression	2.0%	4.35%
Linear Regression	3.0%	4.15%
QDA	**1.33%**	**2.0%**

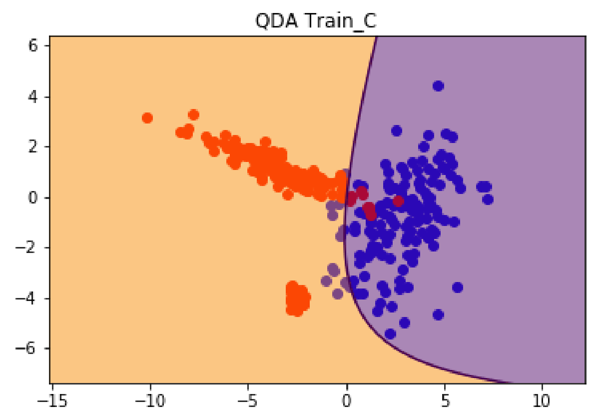
- The Dataset B is drawn, conditionned on the class, from two multivariate gaussians with different means and different covariance matrices. The QDA assumption holds and we expect it to have the best results.
- The Dataset is not linearly separable, so we expect models with higher polynomial degrees to fit better.
- The QDA is a quadratic model, with good assumptions about the data distribution, thus, it outperforms all the other models on the train and test split as expected.

Dataset C









Miss-classification Error

Model	Training	Test
LDA	5.5%	4.23%
Logistic Regression	**4.0%**	**2.27%**
Linear Regression	5.5%	4.23%
QDA	4.23%	3.83%

- The Dataset C is drawn, from three multivariate gaussians to build two classes. The use of the LDA/QDA models can't be justified for this dataset. In addition, we have 400 training observations, so we expect models with relaxed assumptions to generalize better.
- The least square estimator is not suited to this kind of tasks (its Likelihood computations is based on gaussian residuals).
- As a result, The Logistic Regression is the model that suits this task. We can see that it, indeed, outmatches all the other models.

Detailed proofs of Page 1

Exercice 1 : Learning in Discrete Graphical Models

In this exercice, we try to tackle the problem of learning in discrete graphical models with two nodes. let X and Z two discrete random variables with respectively M and K classes with probabilities $P(X=m)=\pi_m$ and $P(Z=k|X=m)=\theta_{m,k}$.

Let $\mathbf{\Pi}=\{\pi_1,...,\pi_M\}$ and $\mathbf{\Theta}=\{\theta_{m,k}\}_{1\leq m\leq M, 1\leq k\leq K}$ be the parameters of our model.

Based on an i.i.d sample of size N, we need to compute the maximum of the complete loglikelihood $l(\mathbf{\Pi},\mathbf{\Theta})=log(P(X,Z))=\sum_{i=1}^N log(P(X_i,Z_i))$.

the problem can be written as :

$$\begin{cases} \max_{\mathbf{\Pi},\mathbf{\Theta}} l(\mathbf{\Pi},\mathbf{\Theta}) \\ s.t. \sum_{m=1}^M \pi_m = 1 \\ and \sum_{k=1}^K \theta_{m,k}=1, \forall k\in\{1,\dots,K\} \end{cases}$$

Let's rewrite $Z_i=(Z_{i,1},...,Z_{i,K})$ and $X_i=(X_{i,1},...,X_{i,M})$ as one hot vectors to ease the computations.

Thus, we want to maximize

$$l(\mathbf{\Pi},\mathbf{\Theta})=\sum_{i=1}^N\sum_{k=1}^K\sum_{m=1}^M z_{i,m}x_{i,k}log(\theta_{m,k})+\sum_{i=1}^N\sum_{m=1}^M z_{i,m}log(\pi_m)$$

which is clearly concave on the parameters under affine equality constraints. The strong duality holds and we can solve the dual problem to get the estimators of our parameters.

The Lagrangian function can be written as :

$$L(\mathbf{\Pi},\mathbf{\Theta},\mathbf{\Lambda})=\sum_{i=1}^N\sum_{k=1}^K\sum_{m=1}^M z_{i,m}x_{i,k}log(\theta_{m,k})+\sum_{i=1}^N\sum_{m=1}^M z_{i,m}log(\pi_m)+\sum_{m=1}^M\lambda_m(1-\sum_{k=1}^K\theta_{m,k})+\lambda_0(1-\sum_{m=1}^M\pi_m)$$

Setting the gradient to zero yields to the following system:

$$\begin{cases} \frac{1}{\lambda_m}\sum_{i=1}^N z_{i,m}x_{i,k}=\hat{\theta}_{m,k} \quad \forall k\in\{1,...,K\},\forall m\in\{1,...,M\} \\ \frac{1}{\lambda_0}\sum_{i=1}^N z_{i,m}=\hat{\pi}_m \quad \forall m\in\{1,...,M\} \\ s.t. \sum_{m=1}^M\hat{\pi}_m=1, \quad \sum_{k=1}^K\hat{\theta}_{m,k}=1,\forall k\in\{1,\dots,K\} \end{cases}$$

which leads to the following final estimator :

$$\begin{cases} \hat{\theta}_{m,k}=\dfrac{1}{\sum_{i=1}^N z_{i,m}}\sum_{i=1}^N z_{i,m}x_{i,k} \quad \forall k\in\{1,...,K\},\forall m\in\{1,...,M\} \\ \\ \hat{\pi}_m=\dfrac{1}{N}\sum_{i=1}^N z_{i,m} \quad \forall m\in\{1,...,M\} \end{cases}$$

Exercice 2.1 (a) : LDA

We start with inference computation to get the form of $p(y|x)$

$$\begin{aligned}
p(y|x) &= \frac{p(x|y)p(y)}{p(x)} \\
&\propto p(x|y)p(y) \\
&\propto \pi^y (1-\pi)^{1-y} \mathcal{N}(x|\mu_y, \Sigma_y) \\
&\propto \pi^y (1-\pi)^{1-y} \left[\frac{1}{|\Sigma_y|^{1/2}} exp\left(-\frac{1}{2} (x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y) \right) \right]
\end{aligned}$$

According to the Fisher assumption $\Sigma_0 = \Sigma_1 = \Sigma$

$$\begin{aligned}
p(y|x) &\propto \pi^y (1-\pi)^{1-y} exp\left(-\frac{1}{2} (x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y) \right) \\
&\propto \pi^y (1-\pi)^{1-y} exp\left(-\frac{y}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T \Sigma^T (x-\mu_0) \right) \\
&\propto \pi^y (1-\pi)^{1-y} exp\Bigg(-\frac{y}{2} x^T \Sigma^{-1} x + \frac{y}{2} x^T \Sigma^{-1} \mu_1 + \frac{y}{2} \mu_1^T \Sigma^{-1} x - \frac{y}{2} \mu_1^T \Sigma^{-1} \mu_1 \\
&+ \frac{y}{2} x^T \Sigma^{-1} x - \frac{y}{2} x^T \Sigma^{-1} \mu_0 - \frac{y}{2} \mu_0^T \Sigma^{-1} x + \frac{y}{2} \mu_0^T \Sigma^{-1} \mu_0 \Bigg) \\
&\propto \pi^y (1-\pi)^{1-y} exp\left(-\frac{y}{2} \left(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0 \right) - y \left(-\mu_1^T \Sigma^{-1} + \mu_0^T \Sigma^{-1} \right) x \right) \\
&\propto exp\left(ya + yb^T x \right)
\end{aligned}$$

Where

$$\begin{cases} a &= \frac{1}{2} \left(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1 \right) + log\left(\frac{\pi}{1-\pi} \right)) \\ b &= \Sigma^{-1} (\mu_1 - \mu_0) \end{cases}$$

Now that We have the closed forms of a and b with respect to the parameters μ_0, μ_1, Σ_0, Σ_1 and π, We can use MLE to compute these parameters.

$$\begin{aligned}
\mathcal{L}\left(\pi,\mu_0,\mu_1,\Sigma_0,\Sigma_1\right) &= \sum_{i=1}^{N} log\left(p(x_i,y_i)\right) \\
&= \sum_{i=1}^{N} log\left(p(x_i|y_i)p(y_i)\right) \\
&= \sum_{i=1}^{N} log\left(\pi^{y_i}(1-\pi)^{1-y_i}\right) + \sum_{i=1}^{N} \mathcal{N}\left(x_i|\mu_{y_i},\Sigma_{y_i}\right) \\
&= \sum_{i=1}^{N} \left(y_i log(\pi) + (1-y_i) log(1-\pi)\right) + \sum_{i=1}^{N} \mathcal{N}\left(x_i|\mu_{y_i},\Sigma_{y_i}\right)
\end{aligned}$$

$\mathcal{L}$ is concave with respect to π:

$$\nabla_\pi \mathcal{L} = \sum_{i=1}^{N} \left(\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}\right)$$

So by setting $\nabla_\pi \mathcal{L} = 0$

$$\begin{aligned}
\sum_{i=1}^{N} \left(\frac{y_i}{\hat{\pi}} - \frac{1-y_i}{1-\hat{\pi}}\right) &= 0 \\
\sum_{i=1}^{N} \left((1-\hat{\pi})y_i - \hat{\pi}(1-y_i)\right) &= 0 \\
\sum_{i=1}^{N} y_i - \hat{\pi}\sum_{i=1}^{N} y_i - N\hat{\pi} + \hat{\pi}\sum_{i=1}^{N} y_i &= 0
\end{aligned}$$

We get $\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N} y_i$. Let's move to the second term *sum* of $\mathcal{L}$, We know that:

$$\begin{aligned}
\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i},\Sigma_{y_i}\right)\right) = \sum_{i=1}^{N} \Bigg(& -\frac{y_i}{2}\left(x_i-\mu_1\right)^T \Sigma^{-1}\left(x_i-\mu_1\right) \\
& - \frac{1-y_i}{2}\left(x_i-\mu_0\right)^T \Sigma^{-1}\left(x_i-\mu_0\right) \\
& + \frac{1}{2} log\left(|\Sigma^{-1}|\right) - log\left(2\pi\right)\Bigg)
\end{aligned}$$

Gradient with respect to μ_1:

$$\nabla_{\mu_1}\mathcal{L} = \sum_{i=1}^{N}\left(-y_i\Sigma^{-1}\left(x_i-\mu_1\right)\right)$$

By setting $\nabla_{\mu_1}\mathcal{L} = 0$, We get $\hat{\mu_1} = \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i x_i$

Gradient with respect to μ_0:

$$\nabla_{\mu_0}\mathcal{L} = \sum_{i=1}^{N} \left(-(1-y_i)\Sigma^{-1}\left(x_i - \mu_0\right)\right)$$

By setting $\nabla_{\mu_0}\mathcal{L} = 0$, We get $\hat{\mu_0} = \frac{1}{\sum_{i=1}^{N}(1-y_i)} \sum_{i=1}^{N}(1-y_i) x_i$

Note: $x^T A x = Tr(x^T A x) = Tr(A x x^T)$

We can rewrite

$$\begin{aligned}\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right)\right) = \sum_{i=1}^{N}\Bigg(&-\frac{y_i}{2} Tr\left(\Sigma^{-1}\left(x_i - \mu_1\right)\left(x_i - \mu_1\right)^T\right) \\ &- \frac{1-y_i}{2} Tr\left(\Sigma^{-1}\left(x_i - \mu_0\right)\left(x_i - \mu_0\right)^T\right) \\ &+ \frac{1}{2} log\left(|\Sigma^{-1}|\right) - log\left(2\pi\right)\Bigg)\end{aligned}$$

Let's set the sample convariance matrices

$$\begin{cases} \tilde{\Sigma_1} &= \frac{1}{\sum_{i=1}^{N} y_i} \sum_{i=1}^{N} y_i\left(x_i - \mu_1\right)\left(x_i - \mu_1\right)^T \\ \tilde{\Sigma_0} &= \frac{1}{\sum_{i=1}^{N}(1-y_i)} \sum_{i=1}^{N}(1-y_i)\left(x_i - \mu_1\right)\left(x_i - \mu_1\right)^T \end{cases}$$

We have then

$$\begin{aligned}\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right)\right) = &-\frac{\sum_{i=1}^{N} y_i}{2} Tr\left(\Sigma^{-1}\tilde{\Sigma_1}\right) - \frac{\sum_{i=1}^{N}(1-y_i)}{2} Tr\left(\Sigma^{-1}\tilde{\Sigma_0}\right) \\ &+ \frac{N}{2} log\left(|\Sigma^{-1}|\right) - N log\left(2\pi\right)\end{aligned}$$

Gradient with respect to Σ:

$$\nabla_{\Sigma_1}\mathcal{L} = -\frac{\sum_{i=1}^{N} y_i}{2}\tilde{\Sigma_1} - \frac{\sum_{i=1}^{N}(1-y_i)}{2}\tilde{\Sigma_0} + \frac{N}{2}\hat{\Sigma}$$

By setting $\nabla_{\Sigma}\mathcal{L} = 0$ We get $\hat{\Sigma} = \frac{\sum_{i=1}^{N} y_i}{N}\tilde{\Sigma_1} + \frac{\sum_{i=1}^{N}(1-y_i)}{N}\tilde{\Sigma_0}$

Exercice 2.5 (a) : QDA

We start with inference computation to get the form of $p(y|x)$

$$\begin{aligned}
p(y|x) &= \frac{p(x|y)p(y)}{p(x)} \\
&\propto p(x|y)p(y) \\
&\propto \pi^y (1-\pi)^{1-y} \mathcal{N}(x|\mu_y, \Sigma_y) \\
&\propto \pi^y (1-\pi)^{1-y} \left[\frac{1}{|\Sigma_y|^{1/2}} exp\left(-\frac{1}{2} (x-\mu_y)^T \Sigma_y^{-1} (x-\mu_y) \right) \right]
\end{aligned}$$

Given a 2-D multivariate normal distribution with mean μ and covariance matrix Σ, We can write:

$$\begin{aligned}
\mathcal{N}(x|\mu, \Sigma) &= \frac{1}{2\pi|\Sigma|^{1/2}} exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right) \\
&= exp\left((-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) + \frac{1}{2} log\left(|\Sigma^{-1}| \right) - log(2\pi) \right)
\end{aligned}$$

This way We can write:

$$\begin{aligned}
\mathcal{N}(x_i|\mu_{y_i}, \Sigma_{y_i}) = exp\Bigg(&-\frac{y_i}{2} (x_i-\mu_1)^T \Sigma_1^{-1} (x_i-\mu_1) + \frac{y_i}{2} log\left(|\Sigma_1^{-1}| \right) \\
&- \frac{1-y_i}{2} (x_i-\mu_0)^T \Sigma_0^{-1} (x_i-\mu_0) + \frac{1-y_i}{2} \left(|\Sigma_0^{-1}| \right) - log(2\pi) \Bigg)
\end{aligned}$$

which leads to

$$\begin{aligned}
p(y|x) &\propto \pi^y (1-\pi)^{1-y} exp\Bigg(-\frac{y}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) + \frac{y}{2} log\left(|\Sigma_1^{-1}| \right) \\
&- \frac{1-y}{2} (x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0) + \frac{1-y}{2} \left(|\Sigma_0^{-1}| \right) \Bigg) \\
&\propto \pi^y (1-\pi)^{1-y} exp\Bigg(-\frac{y}{2} x^T \Sigma_1^{-1} x + \frac{y}{2} x^T \Sigma_1^{-1} \mu_1 + \frac{y}{2} \mu_1^T \Sigma_1^{-1} x \\
&- \frac{y}{2} \mu_1^T \Sigma_1^{-1} \mu_1 + \frac{y}{2} x^T \Sigma_0^{-1} x - \frac{y}{2} x^T \Sigma_0^{-1} \mu_0 \\
&- \frac{y}{2} \mu_0^T \Sigma_0^{-1} x + \frac{y}{2} \mu_0^T \Sigma_0^{-1} \mu_0 + \frac{y}{2} \left(|\Sigma_1^{-1}| \right) + \frac{1-y}{2} \left(|\Sigma_0^{-1}| \right) \Bigg) \\
&\propto \pi^y (1-\pi)^{1-y} exp\Bigg(-\frac{y}{2} \left(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0 + log\left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) \right) \\
&- y \left(-\mu_1^T \Sigma_1^{-1} + \mu_0^T \Sigma_0^{-1} \right) x - \frac{y}{2} x^T \left(\Sigma_1^{-1} - \Sigma_0^{-1} \right) x \Bigg) \\
&\propto exp\left(ya + yb^T x + yx^T cx \right)
\end{aligned}$$

Where

$$\begin{cases} a &= \frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1\right) + log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2}log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) \\ b &= \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0 \\ c &= \frac{1}{2}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right) \end{cases}$$

And by normalizing $p(y=0|x) + p(y=1|x) = 1$, We get:

$$\begin{aligned} p(y=1|x) &= \frac{exp\left(a + b^T + x^Tcx\right)}{1 + exp\left(a + b^T + x^Tcx\right)} \\ &= \sigma\left(a + b^Tx + x^Tcx\right) \end{aligned}$$

Now that We have $p(y=1|x)$ and the closed forms of a, b and c with respect to the paramaters μ_0, μ_1, Σ_0, Σ_1 and π, We can compute the MLE to estimate these parameters:

$$\begin{aligned} \mathcal{L}\left(\pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1\right) &= \sum_{i=1}^{N} log\left(p(x_i, y_i)\right) \\ &= \sum_{i=1}^{N} log\left(p(x_i|y_i)p(y_i)\right) \\ &= \sum_{i=1}^{N} log\left(\pi^{y_i}(1-\pi)^{1-y_i}\right) + \sum_{i=1}^{N} \mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right) \\ &= \sum_{i=1}^{N} \left(y_i log(\pi) + (1-y_i)log(1-\pi)\right) + \sum_{i=1}^{N} \mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right) \end{aligned}$$

$\mathcal{L}$ is concave with respect to π:

$$\nabla_\pi \mathcal{L} = \sum_{i=1}^{N}\left(\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}\right)$$

So by setting $\nabla_\pi \mathcal{L} = 0$

$$\begin{aligned} \sum_{i=1}^{N}\left(\frac{y_i}{\hat{\pi}} - \frac{1-y_i}{1-\hat{\pi}}\right) &= 0 \\ \sum_{i=1}^{N}\left((1-\hat{\pi})y_i - \hat{\pi}(1-y_i)\right) &= 0 \\ \sum_{i=1}^{N} y_i - \hat{\pi}\sum_{i=1}^{N} y_i - N\hat{\pi} + \hat{\pi}\sum_{i=1}^{N} y_i &= 0 \end{aligned}$$

We get $\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N} y_i$
Let's move to the second term *sum* of $\mathcal{L}$, We know that:

$$\begin{aligned}\mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right) = exp\Bigg(&-\frac{y_i}{2}\left(x_i-\mu_1\right)^T \Sigma_1^{-1}\left(x_i-\mu_1\right) + \frac{y_i}{2}log\left(|\Sigma_1^{-1}|\right)\\ &- \frac{1-y_i}{2}\left(x_i-\mu_0\right)^T \Sigma_0^{-1}\left(x_i-\mu_0\right) + \frac{1-y_i}{2}\left(|\Sigma_0^{-1}|\right)\\ &- log\left(2\pi\right)\Bigg)\end{aligned}$$

And so the second term of $\mathcal{L}$ can be written

$$\begin{aligned}\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right)\right) = \sum_{i=1}^{N}\Bigg(&-\frac{y_i}{2}\left(x_i-\mu_1\right)^T \Sigma_1^{-1}\left(x_i-\mu_1\right) + \frac{y_i}{2}log\left(|\Sigma_1^{-1}|\right)\\ &- \frac{1-y_i}{2}\left(x_i-\mu_0\right)^T \Sigma_0^{-1}\left(x_i-\mu_0\right) + \frac{1-y_i}{2}log\left(|\Sigma_0^{-1}|\right)\\ &- log\left(2\pi\right)\Bigg)\end{aligned}$$

Gradient with respect to μ_1:

$$\nabla_{\mu_1}\mathcal{L} = \sum_{i=1}^{N}\left(-y_i\Sigma_1^{-1}\left(x_i-\mu_1\right)\right)$$

By setting $\nabla_{\mu_1}\mathcal{L} = 0$, We get $\hat{\mu_1} = \frac{1}{\sum_{i=1}^{N} y_i}\sum_{i=1}^{N} y_i x_i$
Gradient with respect to μ_0:

$$\nabla_{\mu_0}\mathcal{L} = \sum_{i=1}^{N}\left(-(1-y_i)\Sigma_0^{-1}\left(x_i-\mu_0\right)\right)$$

By setting $\nabla_{\mu_0}\mathcal{L} = 0$, We get $\hat{\mu_0} = \frac{1}{\sum_{i=1}^{N}(1-y_i)}\sum_{i=1}^{N}(1-y_i)x_i$

Note: $x^TAx = Tr(x^TAx) = Tr(Axx^T)$
So We can rewrite

$$\begin{aligned}\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i}, \Sigma_{y_i}\right)\right) = \sum_{i=1}^{N}\Bigg(&-\frac{y_i}{2}Tr\left(\Sigma_1^{-1}\left(x_i-\mu_1\right)\left(x_i-\mu_1\right)^T\right) + \frac{y_i}{2}log\left(|\Sigma_1^{-1}|\right)\\ &- \frac{1-y_i}{2}Tr\left(\Sigma_0^{-1}\left(x_i-\mu_0\right)\left(x_i-\mu_0\right)^T\right) + \frac{1-y_i}{2}log\left(|\Sigma_0^{-1}|\right)\\ &- log\left(2\pi\right)\Bigg)\end{aligned}$$

Let's set

$$\begin{cases}\tilde{\Sigma_1} &= \frac{1}{\sum_{i=1}^{N} y_i}\sum_{i=1}^{N} y_i\left(x_i-\mu_1\right)\left(x_i-\mu_1\right)^T\\ \tilde{\Sigma_0} &= \frac{1}{\sum_{i=1}^{N}(1-y_i)}\sum_{i=1}^{N}(1-y_i)\left(x_i-\mu_1\right)\left(x_i-\mu_1\right)^T\end{cases}$$

We get

$$\sum_{i=1}^{N} log\left(\mathcal{N}\left(x_i|\mu_{y_i},\Sigma_{y_i}\right)\right) = -\frac{\sum_{i=1}^{N} y_i}{2} Tr\left(\Sigma_1^{-1}\tilde{\Sigma_1}\right) - \frac{\sum_{i=1}^{N}(1-y_i)}{2} Tr\left(\Sigma_0^{-1}\tilde{\Sigma_0}\right)$$
$$+ \frac{\sum_{i=1}^{N} y_i}{2} log\left(|\Sigma_1^{-1}|\right) + \frac{\sum_{i=1}^{N}(1-y_i)}{2} log\left(|\Sigma_0^{-1}|\right)$$
$$- N log\left(2\pi\right)$$

Gradient with respect to Σ_1:

$$\nabla_{\Sigma_1}\mathcal{L} = -\frac{\sum_{i=1}^{N} y_i}{2}\tilde{\Sigma_1} + \frac{\sum_{i=1}^{N} y_i}{2}\hat{\Sigma_1}$$

By setting $\nabla_{\Sigma_1}\mathcal{L} = 0$ We get $\hat{\Sigma_1} = \tilde{\Sigma_1}$ **Gradient with respect to** Σ_0:

$$\nabla_{\Sigma_0}\mathcal{L} = -\frac{\sum_{i=1}^{N}(1-y_i)}{2}\tilde{\Sigma_0} + \frac{\sum_{i=1}^{N}(1-y_i)}{2}\hat{\Sigma_0}$$

By setting $\nabla_{\Sigma_0}\mathcal{L} = 0$ We get $\hat{\Sigma_0} = \tilde{\Sigma_0}$