Probabilistic Graphical Models: Homework 1

Petrovich Mathis, Bricout Raphaël

October 23, 2018

See the full exercises in Annex (page 5).

Exercise 1: Learning in Discrete Graphical Models

We would like to maximize the log of $L(\pi, \theta)$.

Variables are i.i.d. so: $l(\pi, \theta) = \log(L(\pi, \theta)) = \sum_{i=1}^{n} \left(\sum_{m=1}^{M} \left(z_m^i \log(\pi_m) + \sum_{k=1}^{K} x_k^i z_m^i \log(\theta_{m,k}) \right) \right)$

With constraints: $\sum_{m=1}^{M} \pi_m = 1$ and $\sum_{k=1}^{K} \sum_{m=1}^{M} \theta_{m,k} = 1$.

We compute the Laplacian, which is convex. We are under Slater's conditions so we can obtain the solution by computing the gradient with relation to π_m, $\theta_{m,k}$, and λ_1 and λ_2 to determine the unknowns.

We get : $\pi_m = \frac{1}{\lambda_1} \sum_{i=1}^{m} z_m^i$, $\lambda_1 = n$ and thus: $\hat{\pi}_m = \frac{w_m}{n}$ with w_m the number of times where $z_m^i = 1$.

Finally: $\hat{\theta}_{m,k} = \frac{w_{m,k}}{n}$ with $w_{m,k}$ the number of times where $x_i = k$ and $z_m^i = 1$.

Exercice 2.1(a): LDA Formulas

With: $x_1^0, ..., x_n^0$ data where $y = 0$ and $x_1^1, ..., x_m^1$ data where $y = 1$, we search $\hat{\theta} = (\hat{\pi}, \hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}) = \underset{\pi, \mu_0, \mu_1, \Sigma}{\arg\max}\, L(\pi, \mu_0, \mu_1, \Sigma)$.

Because variables are i.i.d., law of total probability and the form of the normal law:

$l(\theta) = \sum_{i=1}^{n} \left(-\frac{1}{2}(x_i^0 - \mu_0)^T \Sigma^{-1} (x_i^0 - \mu_0) \right) + \sum_{i=1}^{m} \left(-\frac{1}{2}(x_i^1 - \mu_1)^T \Sigma^{-1} (x_i^1 - \mu_1) \right)$
$+ n \log(1 - \pi) + m \log(\pi) + \frac{n}{2} \log\det \Sigma^{-1} + \frac{m}{2} \log\det \Sigma^{-1}$

Computing the gradient with relation to μ_0, μ_1 and π: $\hat{\mu_0} = \frac{1}{n} \sum_{i=1}^{n} x_i^0$, $\hat{\mu_1} = \frac{1}{n} \sum_{i=1}^{n} x_i^1$, $\hat{\pi} = \frac{m}{n+m}$

And for $\nabla = \Sigma^{-1}$: $\hat{\Sigma} = \frac{1}{n+m} \sum_{i=1}^{n} (x_i^0 - \mu_0)(x_i^0 - \mu_0)^T + \sum_{i=1}^{m} (x_i^1 - \mu_1)(x_i^1 - \mu_1)^T$

Then we use Bayes rule, the law of total probability to get close to the form of the logistic regression ($\Pi = P(Y = 1)$):

$$P(Y = 1 \mid X = x) = \frac{\exp(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1} (x - \mu_1))\Pi}{\exp(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1} (x - \mu_1))\Pi + \exp(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1} (x - \mu_0))(1 - \Pi)}$$

And as: $P(Y = 0 \mid X = x) = 1 - P(Y = 1 \mid X = x)$ we get $P(Y = 0 \mid X = x)$ as well.

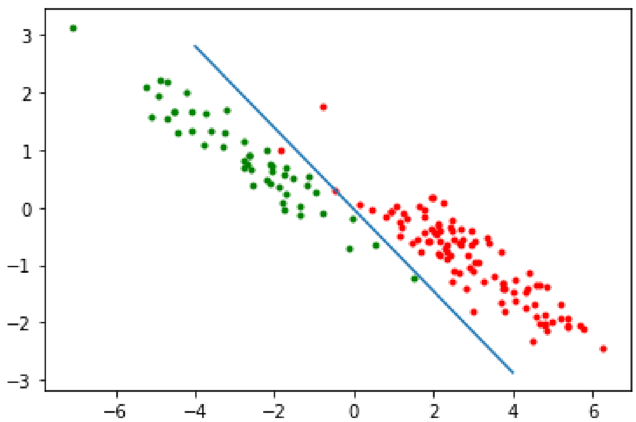
Exercise 2.5(a): QDA Formulas

With the same notations than previously, we get the same estimators except for sigmas: $\hat{\mu_0} = \frac{1}{n} \sum_{i=1}^{n} x_i^0$, $\hat{\mu_1} = \frac{1}{n} \sum_{i=1}^{n} x_i^1$,

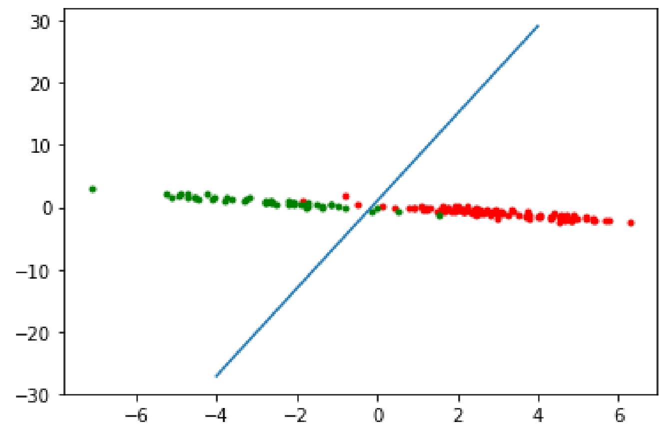
$\hat{\pi} = \frac{m}{n+m}$ and finally: $\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i^0 - \mu_0)(x_i^0 - \mu_0)^T$ and $\hat{\Sigma}_1 = \frac{1}{m} \sum_{i=1}^{m} (x_i^1 - \mu_1)(x_i^1 - \mu_1)^T$

Classification dataset A

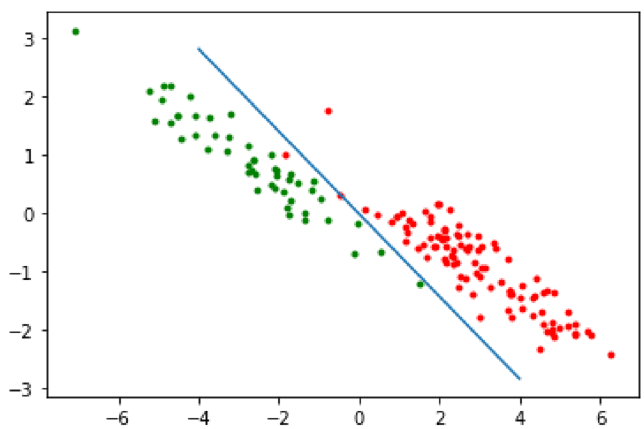
LDA



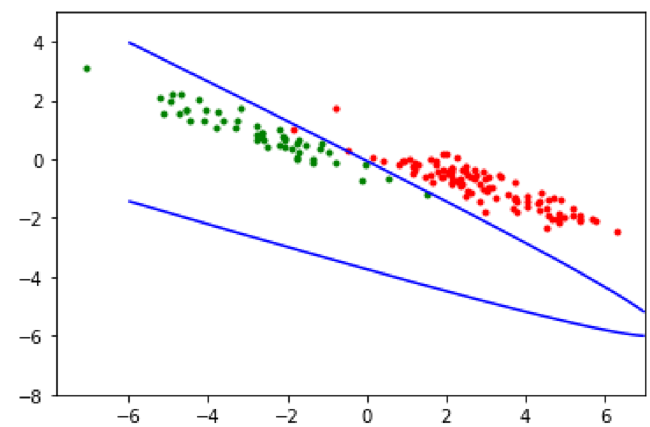
Logistic Regression



Least Squares



QDA



Comments dataset A

Error

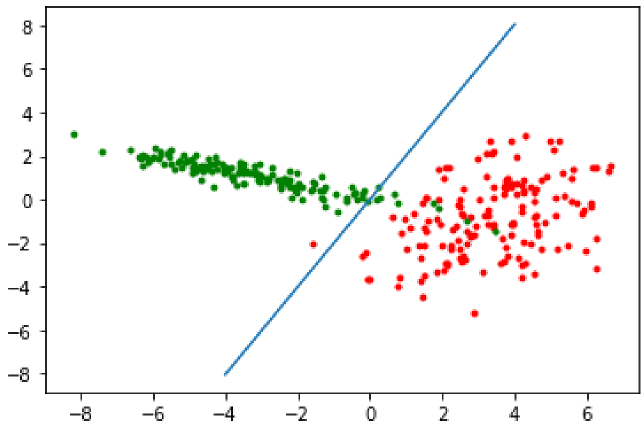
	train (%)	test (%)
LDA	1.33	2.06
Log	4.66	2.60
LS	1.33	2.06
QDA	0.66	2.00

Other Comments

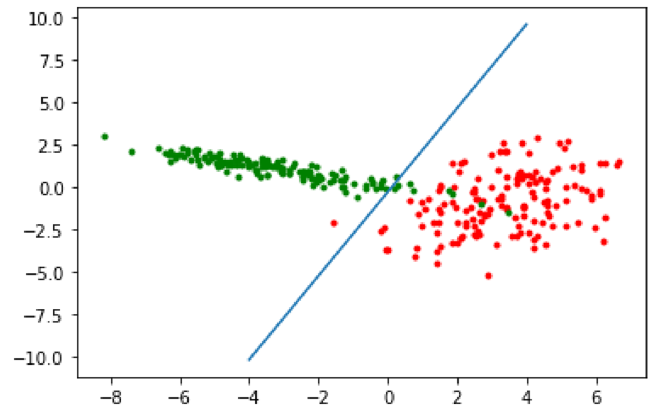
- LDA and Least Square are very similar
- QDA is the best but Least Square and LDA are very good to.
- Logistic are the only one to do better on test data

Classification dataset B

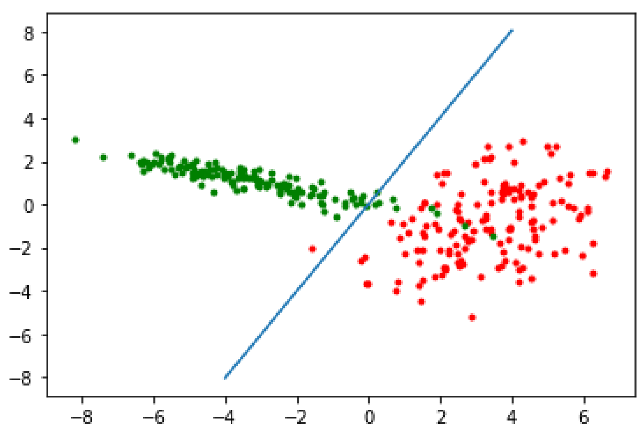
LDA



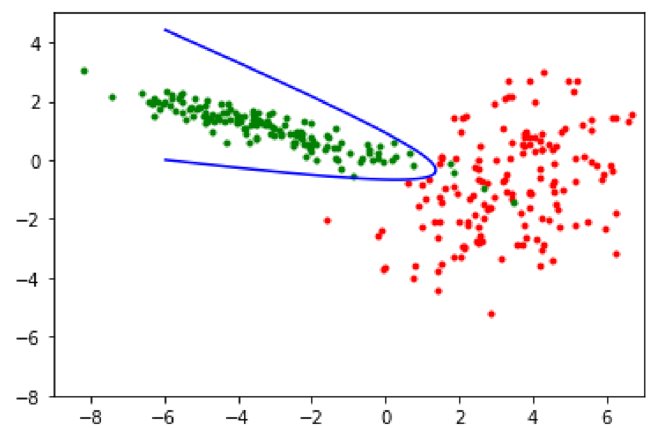
Logistic Regression



Least Squares



QDA



Comments dataset B

Error

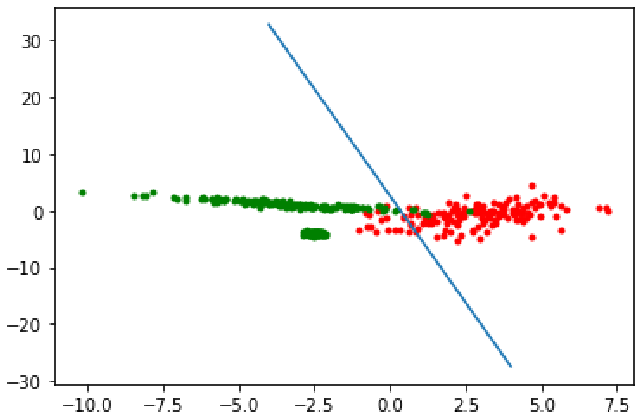
	train (%)	test (%)
LDA	3.00	4.15
Log	3.00	4.25
LS	3.00	4.15
QDA	1.33	2.00

Other Comments

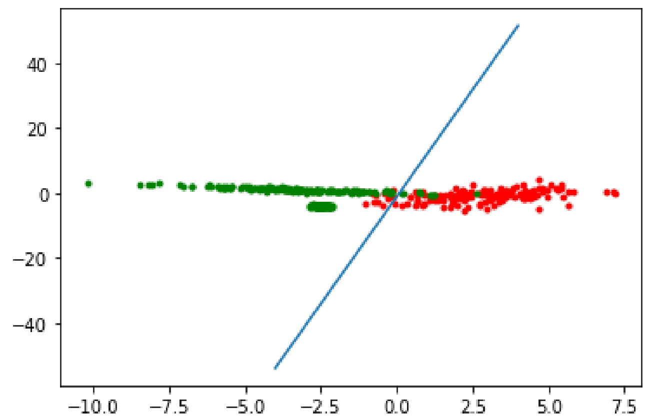
- LDA, Least Square and Logistic are very similar
- QDA works very well here

Classification dataset C

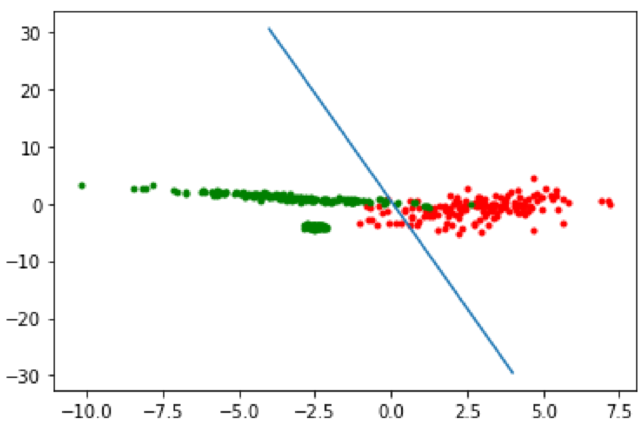
LDA



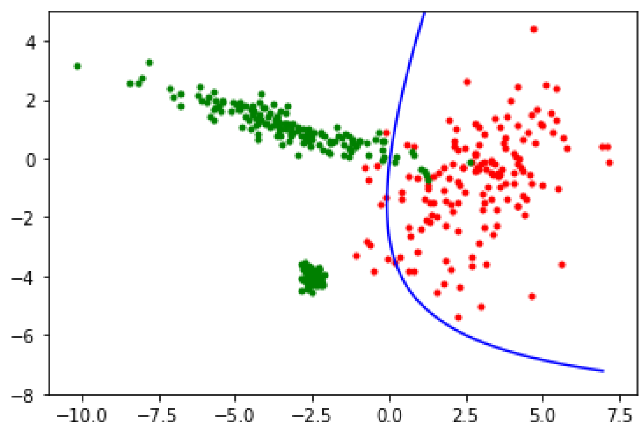
Logistic Regression



Least Squares



QDA



Comments dataset C

Error

	train (%)	test (%)
LDA	6.50	4.33
Log	4.75	3.10
LS	5.50	4.23
QDA	5.25	3.83

Other Comments

- LDA and Least Square are very similar
- QDA and Logistic are very good and can they generalize well on test data

Annex

Exercise 1: Learning in Discrete Graphical Models

For a sample of n observations:

$$\hat{\theta}, \hat{\pi} = \arg\max_{\pi,\theta} L(\pi, \theta) = \arg\max_{\pi,\theta} l(\pi, \theta)$$

With, because variables are i.i.d.:

$$L(\pi, \theta) = \prod_{i=1}^{n} p(x_i, z_i|\pi, \theta) = \prod_{i=1}^{n} \left(\prod_{m=1}^{M} \left(\pi_m^{z_m^i} \prod_{k=1}^{K} \theta_{m,k}{}^{x_k^i z_m^i} \right) \right)$$

(with z^i the odd vector that is 0 everywhere except 1 in it's i^{th} coordinate)
And:

$$l(\pi, \theta) = \log(L(\pi, \theta)) = \sum_{i=1}^{n} \left(\sum_{m=1}^{M} \left(z_m^i \log(\pi_m) + \sum_{k=1}^{K} x_k^i z_m^i \log(\theta_{m,k}) \right) \right)$$

The goal is thus to minimize $-l(\pi, \theta)$ subject to :

- $\sum_{m=1}^{M} \pi_m = 1$
- $\sum_{k=1}^{K} \sum_{m=1}^{M} \theta_{m,k} = 1$

We compute the Laplacian:

$$\mathcal{L}(\pi, \theta, \lambda_1, \lambda_2) = -l(\pi, \theta) + \lambda_1 \left(\sum_{m=1}^{M} \pi_m - 1 \right) + \lambda_2 \left(\sum_{k=1}^{K} \sum_{m=1}^{M} \theta_{m,k} - 1 \right)$$

The aforementioned function is convex, and there exists π and θ strictly feasible (uniform case, where $\pi_m = \frac{1}{M}$ and $\theta_{m,k} = \frac{1}{MK}$).
We are under Slater's conditions so we can obtain the solution by computing the gradient.

$$\nabla_{\pi_m} \mathcal{L}(\pi, \theta, \lambda_1, \lambda_2) = \sum_{i=1}^{m} -\frac{z_m^i}{\pi_m} + \lambda_1$$

Thus:

$$\nabla_{\pi_m} \mathcal{L}(\pi, \theta, \lambda_1, \lambda_2) = 0 \Leftrightarrow \pi_m = \frac{1}{\lambda_1} \sum_{i=1}^{m} z_m^i = \frac{w_m}{\lambda_1}$$

With w_m the number of times where $z_m^i = 1$.

With the additional gradient

$$\nabla_{\lambda_1} \mathcal{L}(\pi, \theta, \lambda_1, \lambda_2) = \sum_{m=1}^{M} \pi_m - 1 = 0$$

We get : $\lambda_1 = n$ and thus:

$$\hat{\pi}_m = \frac{w_m}{n}$$

To find the MLE for θ, we compute the gradient with respect to $\theta_{m,k}$ and then λ_2 to find the last unknown:

$$\nabla_{\theta_{m,k}}\mathcal{L}(\pi,\theta,\lambda_1,\lambda_2)=0 \Leftrightarrow -\sum_{i=1}^{m}\frac{x_k^i z_i^m}{\theta_{m,k}}+\lambda_2=0$$

$$\nabla_{\lambda_2}\mathcal{L}(\pi,\theta,\lambda_1,\lambda_2)=0 \Leftrightarrow \sum_{m=1}^{M}\sum_{k_1}^{K}\theta_{m,k}=1$$

Finally:

$$\hat{\theta}_{m,k}=\frac{w_{m,k}}{n}$$

with $w_{m,k}$ the number of times where $x_i=k$ and $z_m^i=1$.

Exercice 2.1(a): LDA Formulas

Let:

- $x_1^0,...,x_n^0$ data where $y=0$
- $x_1^1,...,x_m^1$ data where $y=1$

Then, $\hat{\theta}=(\hat{\pi},\hat{\mu}_0,\hat{\mu}_1,\hat{\Sigma})=\underset{\pi,\mu_0,\mu_1,\Sigma}{\arg\max}\ L(\pi,\mu_0,\mu_1,\Sigma)$
Where

$$L(\pi,\mu_0,\mu_1,\Sigma)=p(x_1^0,...,x_n^0,x_1^1,...,x_m^1|\pi,\mu_0,\mu_1,\Sigma)$$

And because variables are i.i.d. we get:

$$L(\theta)=\prod_{i=1}^{n}p(x_i^0|\theta)\prod_{i=1}^{m}p(x_i^1|\theta)$$

$$\hat{\theta}=\arg\max_{\theta}L(\theta)=\arg\max_{\theta}\log(L(\theta))=\arg\max_{\theta}l(\theta)$$

We use the law of total probability and the form of the normal law to deduce:

$$l(\theta)=\sum_{i=1}^{n}\left(-\frac{1}{2}(x_i^0-\mu_0)^T\Sigma^{-1}(x_i^0-\mu_0)\right)+\sum_{i=1}^{m}\left(-\frac{1}{2}(x_i^1-\mu_1)^T\Sigma^{-1}(x_i^1-\mu_1)\right)$$
$$+n\log(1-\pi)+m\log(\pi)+\frac{n}{2}\log\det\Sigma^{-1}+\frac{m}{2}\log\det\Sigma^{-1}$$

Computing the gradient with relation to μ_0, μ_1 and π:

$$\nabla_{\mu_0}l(\theta)=0\Leftrightarrow\sum_{i=1}^{n}\Sigma^{-1}(x_i^0-\mu_0)=0\Leftrightarrow\hat{\mu_0}=\frac{1}{n}\sum_{i=1}^{n}x_i^0$$

Same way:

$$\hat{\mu_1}=\frac{1}{n}\sum_{i=1}^{n}x_i^1$$

$$\nabla_{\pi}l(\theta)=\frac{m}{\pi}-\frac{n}{1-\pi}=0\Rightarrow\hat{\pi}=\frac{m}{n+m}$$

The log-likelihood is concave in $\nabla=\Sigma^{-1}$:

$$\nabla_{\Delta}l(\theta)=0\Rightarrow\hat{\Sigma}=\frac{1}{n+m}\sum_{i=1}^{n}(x_i^0-\mu_0)(x_i^0-\mu_0)^T+\sum_{i=1}^{m}(x_i^1-\mu_1)(x_i^1-\mu_1)^T$$

By Bayes Rule:

$$P(Y=1 \mid X=x) = \frac{P(X=x \mid Y=1)P(Y=1)}{P(X=x)}$$

By hypothesis:

$$P(X=x \mid Y=1) = \frac{1}{(2\pi)^{d/2}\sqrt{|det\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)$$

And:

$$P(Y=1) = \Pi$$

Law of total probability:

$$P(X=x) = P(X=x \mid Y=0)P(Y=0) + P(X=x \mid Y=1)P(Y=1)$$

So we get: (with simplifying the constant term)

$$P(Y=1 \mid X=x) = \frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\Pi}{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\Pi + \exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))(1-\Pi)}$$

And as: $P(Y=0 \mid X=x) = 1 - P(Y=1 \mid X=x)$

$$P(Y=0 \mid X=x) = \frac{\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))(1-\Pi)}{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))\Pi + \exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))(1-\Pi)}$$

If we continue the computation for $P(Y=1 \mid X=x)$ we get:

$$P(Y=1 \mid X=x) = \frac{1}{1+\exp(-\frac{1}{2}((x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1)))\frac{(1-\Pi)}{\Pi}}$$

Computation of the interior:
$$\begin{aligned}
Int &= (x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\
Int &= x^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}\mu_0 - x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}\mu_1 \\
Int &= 2(\mu_1-\mu_0)^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1
\end{aligned}$$

So we get:

$$P(Y=1 \mid X=x) = \frac{1}{1+\exp(-(\mu_1-\mu_0)^T\Sigma^{-1}x - \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) + \log(\frac{1-\Pi}{\Pi}))}$$

With:

- $w = \Sigma^{-1}(\mu_1-\mu_0)$
- $b = \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \log(\frac{1-\Pi}{\Pi})$

$$P(Y=1 \mid X=x) = \frac{1}{1+\exp(-(w^Tx+b))} = \sigma(w^Tx+b)$$

Exercise 2.5(a): QDA Formulas

With the same notations than previously:

$$l(\theta) = \sum_{i=1}^{n}(-\frac{1}{2}(x_i^0-\mu_0)^T\Sigma_0^{-1}(x_i^0-\mu_0)) + \sum_{i=1}^{m}(-\frac{1}{2}(x_i^1-\mu_1)^T\Sigma_1^{-1}(x_i^1-\mu_1))$$
$$+ n\log(1-\pi) + m\log(\pi) + \frac{n}{2}\log\det\Sigma_0^{-1} + \frac{m}{2}\log\det\Sigma_1^{-1}$$

We get also:

$$\hat{\mu_0} = \frac{1}{n}\sum_{i=1}^{n} x_i^0$$

and

$$\hat{\mu_1} = \frac{1}{n}\sum_{i=1}^{n} x_i^1$$

as well as:

$$\hat{\pi} = \frac{m}{n+m}$$

And eventually computing the gradient with relation to $\Delta_0 = \Sigma_0^{-1}$ and $\Delta_1 = \Sigma_1^{-1}$ we get:

$$\nabla_{\Delta_0} l(\theta) = \frac{n}{2}\Delta_0^{-1} - \frac{1}{2}\sum_{i=1}^{n}(x_i^0 - \mu_0)(x_i^0 - \mu_0)^T$$

$$\nabla_{\Delta_0} l(\theta) = 0 \Leftrightarrow \hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i^0 - \mu_0)(x_i^0 - \mu_0)^T$$

And:

$$\nabla_{\Delta_1} l(\theta) = 0 \Leftrightarrow \hat{\Sigma}_1 = \frac{1}{m}\sum_{i=1}^{m}(x_i^1 - \mu_1)(x_i^1 - \mu_1)^T$$

Exercise 2.5(b): Conic formula

$$P(Y = 1 \mid X = x) = \frac{1}{2}$$

$$\Rightarrow -\frac{1}{2}\left(x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - 2\mu_0^T\Sigma_0^{-1}x + 2\mu_1^T\Sigma_1^{-1}x + \mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1\right) + \log(\frac{1-\Pi}{\Pi}) + \log(\sqrt{\frac{|det\Sigma_1|}{|det\Sigma_0|}}) = 0$$

It's a conic.