Exercise 1: Learning in discrete graphical models

$\hat{\pi}_{m_i} = \frac{card\{z=m_i\}}{N}$ $\hat{\theta}_{m_i,k_i} = \frac{card\{z = m_i | x_i = k_i\}}{N}$
with N the size of the sample.

Exercise 2.1.(a): LDA formulas

$$\hat{\pi} = \frac{\sum_{i=1}^{N} y_i}{N}, \hat{\mu}_1 = \frac{\sum_{i=1}^{N} y_i x_i}{n}, \hat{\mu}_0 = \frac{\sum_{i=1}^{N} (1 - y_i) x_i}{N - n}, \hat{\Sigma} = \frac{1}{N} \sum_{i}^{N} \left(x_i - \mu_{y_i}\right)^T (x_i - \mu_{y_i})$$

with N the size of the sample and $n = card\{y_i = 1\}$

Conditional probability (it has the same form than logistic regression!):

$$\mathbb{P}(y = 1|x) = \frac{1}{1 + e^{a^T x + b}}$$

With:

$$b = \frac{1}{2}\left[\hat{\mu}_o^T \hat{\Sigma}^{-1} \hat{\mu}_0 - \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1\right] + \log \frac{\hat{\pi}}{1 - \hat{\pi}}$$
$$a = (\hat{\mu}_1^T - \hat{\mu}_o^T) \hat{\Sigma}^{-1}$$

Exercise 2.5.(a): QDA formulas

$$\hat{\pi} = \frac{\sum_{i=1}^{N} y_i}{N}, \hat{\mu}_1 = \frac{\sum_{i=1}^{N} y_i x_i}{n}, \hat{\mu}_0 = \frac{\sum_{i=1}^{N} (1 - y_i) x_i}{N - n}$$
$$\hat{\Sigma}_1 = \frac{1}{n} \sum_{i}^{N} y_i \left(x_i - \mu_{y_i}\right)^T (x_i - \mu_{y_i}), \hat{\Sigma}_0 = \frac{1}{N - n} \sum_{i}^{N} (1 - y_i) \left(x_i - \mu_{y_i}\right)^T (x_i - \mu_{y_i})$$

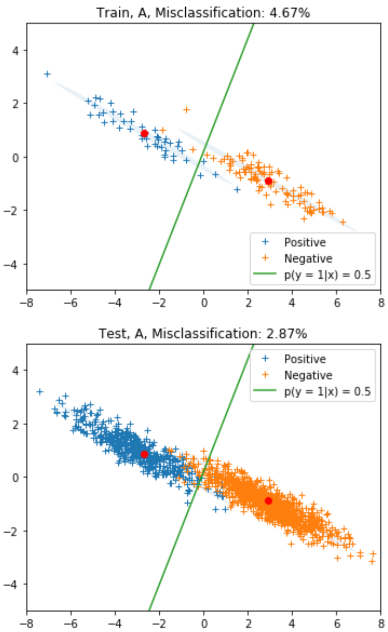
Conditional probability:

$$\mathbb{P}(y = 1|x) = \frac{1}{1 + e^{x^T A x + B^T x + C}}$$

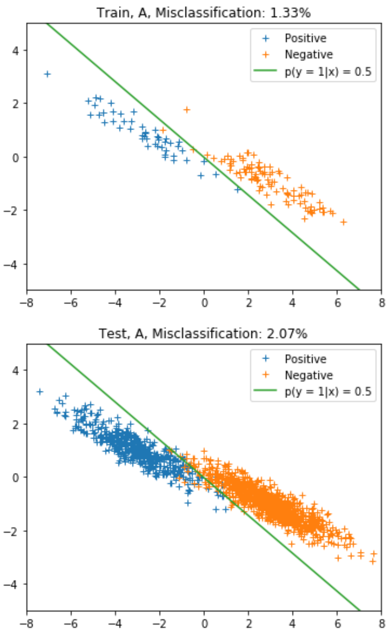
With:

$$C = \frac{1}{2}\left[\hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 - \hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0\right] + \log \frac{\hat{\pi}}{1 - \hat{\pi}} + \frac{1}{2} \log \det \Sigma_1 - \frac{1}{2} \log \det \Sigma_2$$
$$B = (\hat{\mu}_0^T - \hat{\mu}_1^T) \hat{\Sigma}^{-1}$$
$$A = \left(\hat{\Sigma}_1^{-1} - \hat{\Sigma}_o^{-1}\right)/2$$

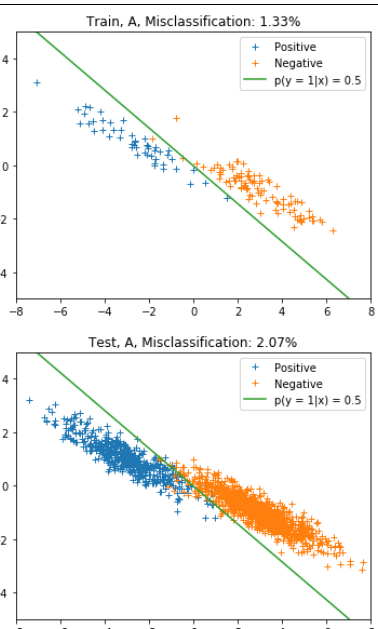
Dataset A



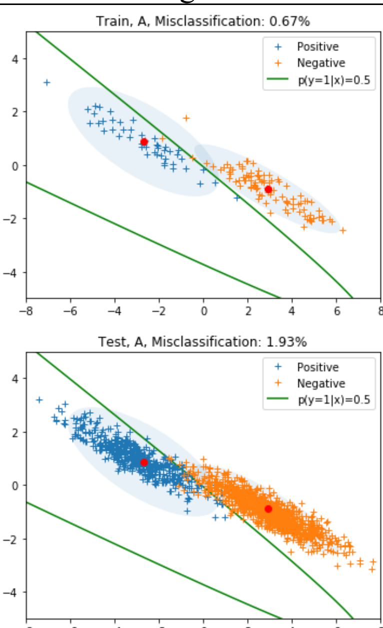
LDA



Logistic



Linear regression



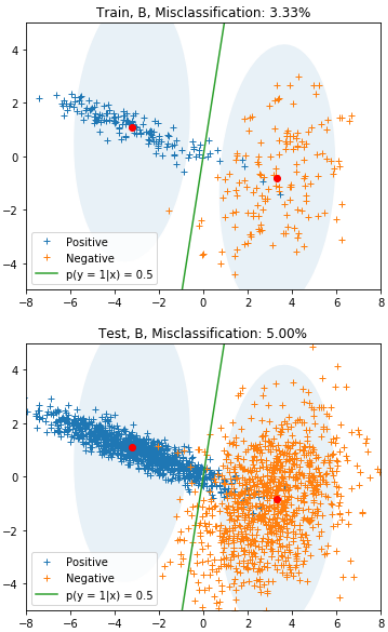
QDA

Table of misclassifications for test set

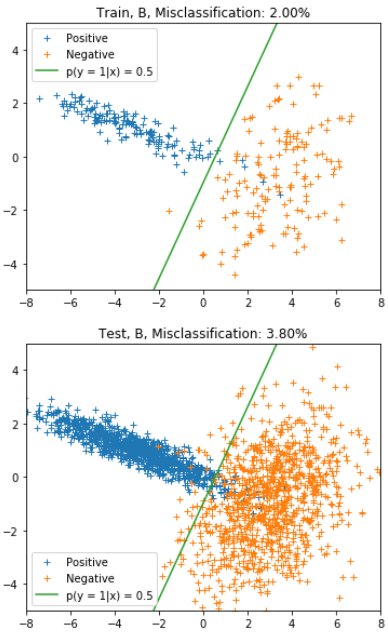
LDA	Logistic	Linear	QDA
2.87 %	2.07 %	2.07%	1.93%

LDA appears being highly unproductive as the dataset is ill-conditionned. All of the other ones provide good results with correct generalization (test set's performances are close to train set's), as the points are linearly separable.

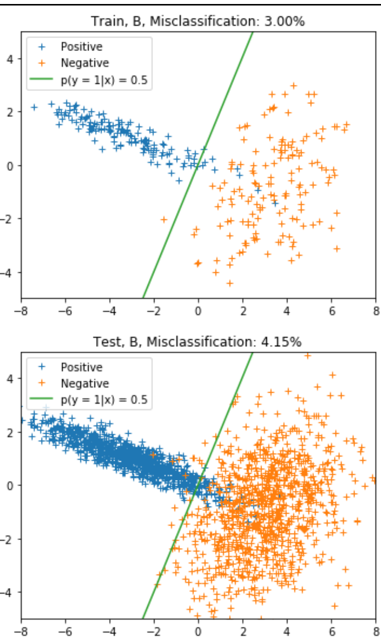
Dataset B



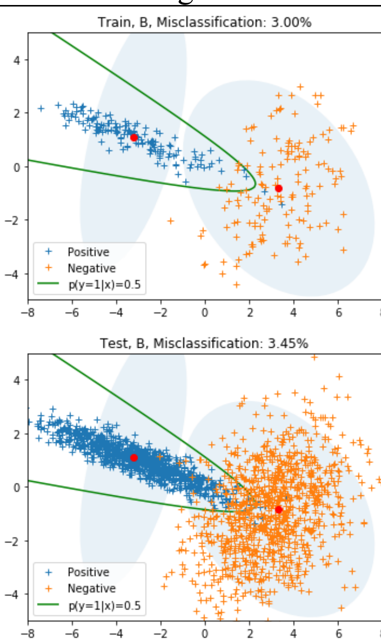
LDA



Logistic



Linear



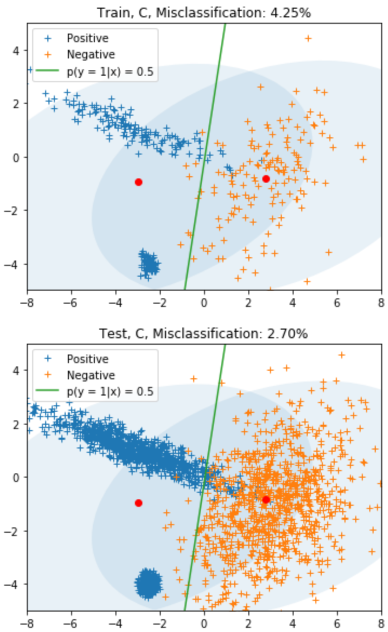
QDA

Table of misclassifications for test set

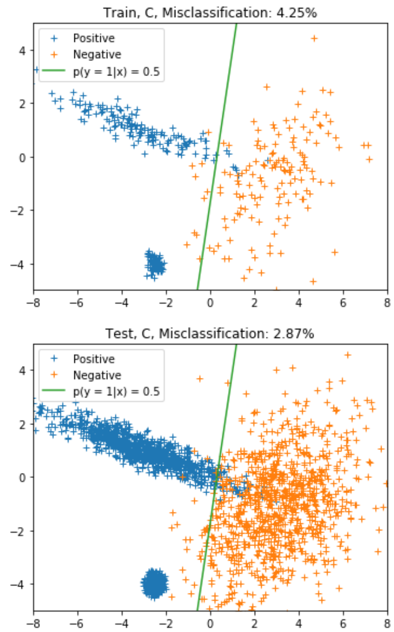
LDA	Logistic	Linear	QDA
5.00 %	3.80 %	4.15%	3.45%

LDA provides bad results, because the assumption $\Sigma_0 = \Sigma_1$ is not respected. appears being highly unproductive as the dataset is ill-conditionned. The logistic and linear regression provide modest results, because the data positive and negative overlap. QDA provides best results, as it can exclude points finer.

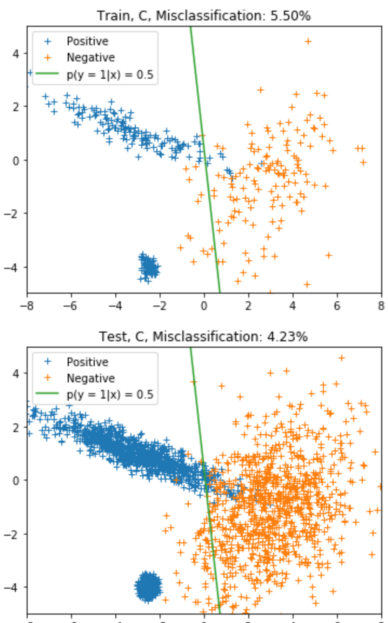
Dataset C



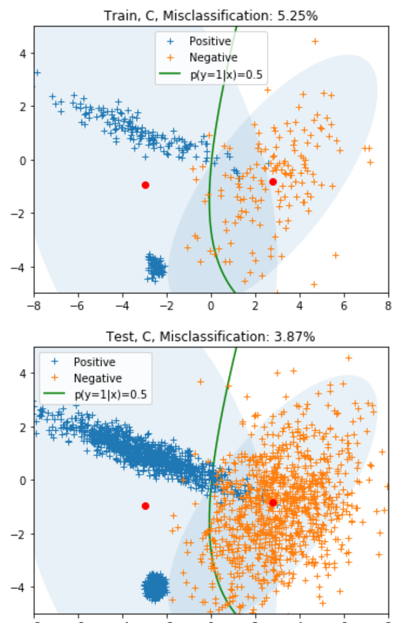
LDA



Logistic



Linear



QDA

Table of misclassifications for test set

LDA	Logistic	Linear	QDA
2.7 %	2.87 %	4.23%	3.87%

Prior models are wrong, because the positive samples are acting like a mixture of 2 Gaussians. Hence, LDA and QDA provide wrong estimate of variance (blue ellipses). The least square of linear regression provide a too important weight to the bottom left blob. Only the logistic is robust to this perturbation.

Proof

We consider only the case of the QDA. To get the proof of LDA, we consider the case where $\Sigma_1 = \Sigma_2$.
The log-likelihood is defined by:

$$l(\pi, \mu, \Sigma) = \log \Pi_{i=1}^{N} p(x_i, y_i | \pi, \mu, \Sigma) = \sum_{i=1}^{N} \log p(y_i | \pi) + \log p(x_i | y_i, \pi, \mu, \Sigma)$$

By definition of the probability density, we can separate the log likelihood into two different likelihood and maximize both of them:

$$l_{Ber}(\pi) = \sum_{i=1}^{N} y_i \log \pi + (1 - y_i) \log 1 - \pi$$

$$l_N(\mu, \Sigma) = -\sum_{i=1}^{N} \log \det \Sigma_{y_i} + \left(x_i - \mu_{y_i}\right)^T \Sigma_{y_i}^{-1} \left(x_i - \mu_{y_i}\right)$$

The two likelihood are concave with respect to their parameters. Consequently, their maximum values are achieved when the gradient is zero.

$$\frac{\partial l_{Ber}(\pi)}{\partial \pi} = 0 \Leftrightarrow \pi = \frac{n}{N}$$

$$\frac{\partial l_N(\mu, \Sigma)}{\partial \mu_1} = 0 \Leftrightarrow \mu_1 = \frac{1}{n} \sum_{i=1}^{N} x_i y_i$$

$$\frac{\partial l_N(\mu, \Sigma)}{\partial \mu_0} = 0 \Leftrightarrow \mu_0 = \frac{1}{n} \sum_{i=1}^{N} x_i (1 - y_i)$$

Using the gradient of $\log \det \Sigma_{y_i}$ and $\left(x_i - \mu_{y_i}\right)^T \Sigma_{y_i}^{-1} \left(x_i - \mu_{y_i}\right)$ as explained in lessons, we have:

$$\frac{\partial l_N(\mu, \Sigma)}{\partial \Sigma_1} = 0 \Leftrightarrow \Sigma_1 = \frac{1}{n} \sum_{i=1}^{N} y_i (x_i - \hat{\mu}_1)^T (x_i - \hat{\mu}_1)$$

$$\frac{\partial l_N(\mu, \Sigma)}{\partial \Sigma_0} = 0 \Leftrightarrow \Sigma_0 = \frac{1}{n} \sum_{i=1}^{N} (1 - y_i)(x_i - \hat{\mu}_0)^T (x_i - \hat{\mu}_0)$$

Finally, to get the boundary, we use the Bayes rule:

$$\mathbb{P}(y = 1 | x) = 0.5 \Leftrightarrow \frac{p(x | y = 1) p(y = 1)}{p(x | y = 0) p(y = 0) + p(x | y = 1) p(y = 1)} = 0.5$$

Replacing the definition of the density function, we have the following equivalence:

$$\frac{1}{1 + e^{x^T A x + B^T x + C}} = 0.5$$

With:

$$A = \frac{1}{2} (\Sigma_1^{-1} - \Sigma_0^{-1})$$

$$B = (\widehat{\mu}_0^T - \widehat{\mu}_1^T) \hat{\Sigma}^{-1}$$

$$C = \frac{1}{2} \widehat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 - \frac{1}{2} \widehat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0 + \log \frac{1 - \pi}{\pi} + \frac{1}{2} \log \det \Sigma_1 - \frac{1}{2} \log \det \Sigma_0$$