# Probabilistic Graphical Models

Homework 1

Jean-Christophe Corvisier
Mohammed Amine Kheldouni

22 October 2018

## 1 Learning in discrete graphical models

Let $X = (x_1, ..., x_n)$ and $Z = (z_1, ..., z_n)$ an i.i.d. sample of the defined variables. We aim to compute the maximum likelihood estimators $\widehat{\pi}$ and $\widehat{\theta}$.

$$log(p(X, Z; \pi, \theta)) = log\left(\prod_{i=1}^{n} = p_\theta(X|Z)p_\pi(z)\right) = \sum_{i=1}^{n} log(p_\theta(X|Z)) + log(p_\pi(Z))$$

Since the random variables are discrete and live in finite spaces, we can consider for $X_i = j$ (resp. $Z_i = j$) vectors in $[M]$ (resp. $[K]$) of zeros and one in the $j^{th}$ element such that $\{X_i = j\} = \{x_j^{(i)} = 1\}$. After this change of notation, we compute the maximum likelihood estimators by deriving the Lagrangian with respect to $\pi$ and $\theta$ as shown in the appendix.
After calculus (please refer to the appendix), the formulas for our estimators are as follows :

$$\widehat{\pi}_m = \frac{\sum\limits_i z_m^{(i)}}{n} \qquad \widehat{\theta}_{mk} = \frac{\sum\limits_i x_k^{(i)} z_m^{(i)}}{\sum\limits_i z_m^{(i)}}$$

## 2 Linear classification

### 2.1 Linear Discriminant Analysis formulas (LDA)

The computed form of the maximum likelihood estimator for the LDA model is :

$$\widehat{\pi} = \frac{1}{n}\sum_i y_i \qquad \widehat{\mu}_0 = \frac{\sum\limits_i (1-y_i)x_i}{\sum\limits_i (1-y_i)} \qquad \widehat{\mu}_1 = \frac{\sum\limits_i y_i x_i}{\sum\limits_i y_i} \qquad \widehat{\Sigma} = \frac{\sum\limits_i y_i(x_i-\mu_1)^T(x_i-\mu_1) + (1-y_i)(x_i-\mu_0)^T(x_i-\mu_0)}{N}$$

Writing down the equations $p(y = 1|x) = 0.5$ for LDA and Logistic Regression (cf. appendix), we find that in the LDA classifier, no assumptions are made regarding the distribution of the explanatory variables.

### 2.2 Quadratic Discriminant Analysis formulas (QDA)

The computed form of the maximum likelihood estimator for the QDA model is :

$$\widehat{\pi} = \frac{1}{n}\sum_i y_i \qquad \widehat{\mu}_0 = \frac{\sum\limits_i (1-y_i)x_i}{\sum\limits_i (1-y_i)} \qquad \widehat{\mu}_1 = \frac{\sum\limits_i y_i x_i}{\sum\limits_i y_i}$$

$$\widehat{\Sigma}_0 = \frac{\sum\limits_i (1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T}{\sum\limits_i (1-y_i)} \qquad \widehat{\Sigma}_1 = \frac{\sum\limits_i y_i(x_i-\mu_1)(x_i-\mu_1)^T}{\sum\limits_i y_i}$$

In the following, we display the traninng plots of classification but please refer to the appendix for the test datasets plots.

## 2.3 Results for training dataset A

### Data and classification boundaries for LDA

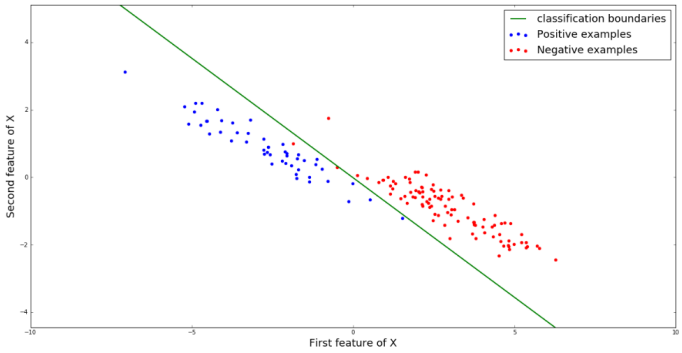


Figure showing the training set in data file A and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

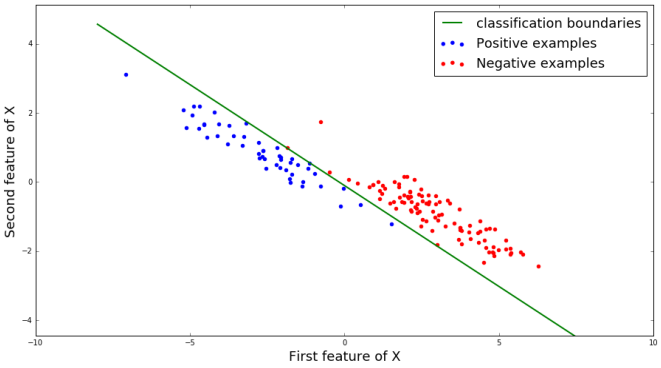


Figure showing the training set in data file A and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

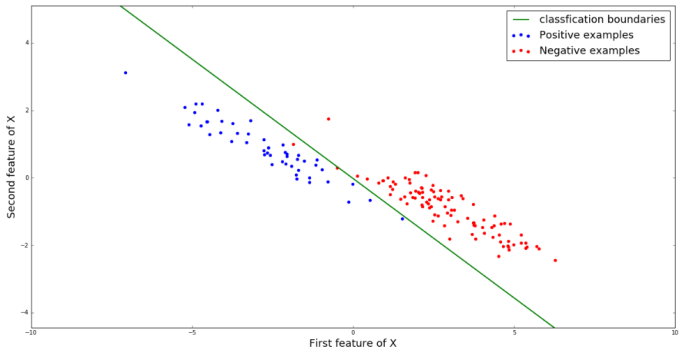


Figure showing the training set in data file A and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

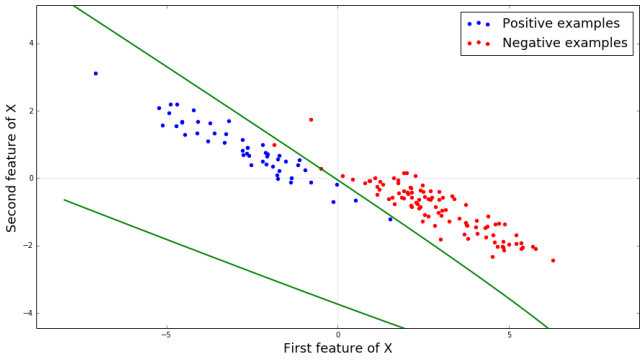


Figure showing the training set in data file A and its classification boundaries using the QDA classifier

### Table of errors

Defining the misclassification error as the ratio of *True Positive* and *True Negative* predictions over all predictions, we establish the following results of error :

| | Train dataset | Test dataset |
|---|---|---|
| LDA | 0.0133 | 0.02 |
| Logistic Regression | 0.0 | 0.0346 |
| Linear Regression | 0.0133 | 0.0206 |
| QDA | 0.0066 | 0.02 |

### Remarks and comments

Considering the table of misclassification errors for dataset A, we notice that the QDA classifier is doing better than the other classifiers except when the logistic regression commits no mistake on the training data. We explain these results by the fact that the hypothesis in the LDA that the two labels have the same covariance matrix ($\Sigma_0 = \Sigma_1$) is relaxed in QDA, providing better estimations of these parameters and better classification results. Comparing LDA with the two others, we acknowledge that this classifier has very similar results as the linear regression on both training and testing sets, and slightly better than the logistic regression on the test set but not on the training set.
The error is uniformly larger on the test data than on the training data. Indeed, the parameters have been computed and calibrated regarding the training data, and new examples from the test data may not perform well predicting the label given those parameters.

## 2.4 Results for dataset B

### Data and classification boundaries for LDA

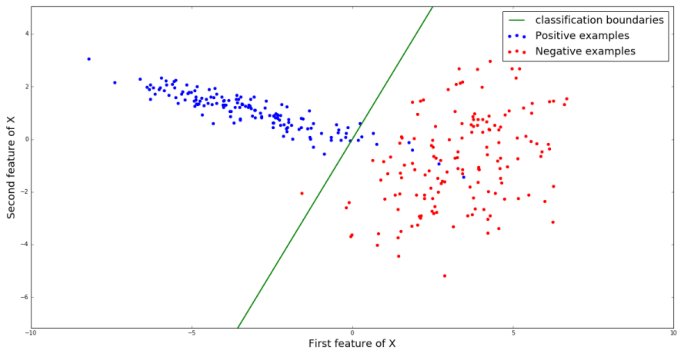


Figure showing the training set in data file B and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

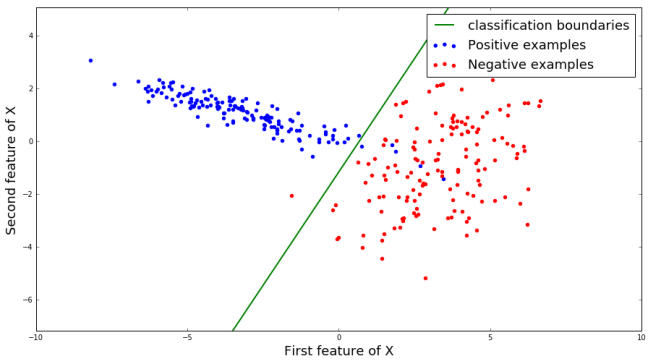


Figure showing the training set in data file B and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

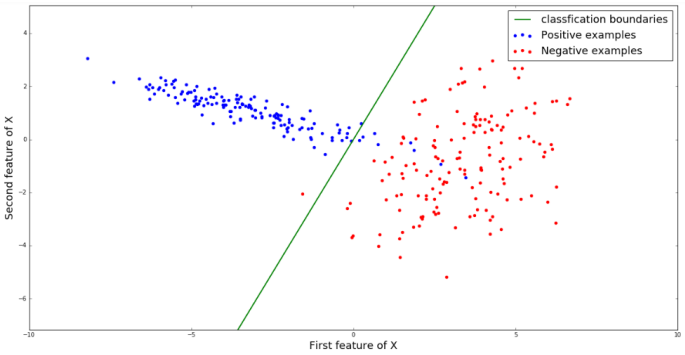


Figure showing the training set in data file B and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

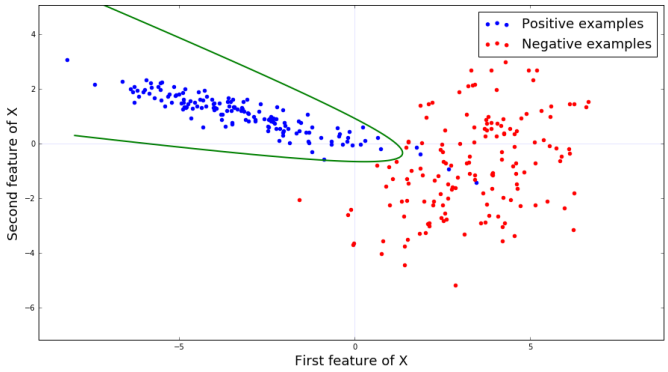


Figure showing the training set in data file B and its classification boundaries using the QDA classifier

### Table of errors

Defining the misclassification error as the ratio of *True Positive* and *True Negative* predictions over all predictions, we establish the following results of error :

| | Train dataset | Test dataset |
|---|---|---|
| LDA | 0.03 | 0.0415 |
| Logistic Regression | 0.02 | 0.0415 |
| Linear Regression | 0.03 | 0.415 |
| QDA | 0.0133 | 0.02 |

### Remarks and comments

Similarly to dataset A, in dataset B the QDA classifier still holds slightly better results than the other classifiers whereas the LDA, the linear regression and the logistic regression hold very similar errors in training and testing sets. In the dataset B, the positive labels are much more compact than the negative labels which brings a tightened conic for QDA compared with the one in dataset A.
The test errors are still and always bigger than the train errors due to new examples testing the predictability of our calibrated model.

## 2.5 Results for dataset C

### Data and classification boundaries for LDA

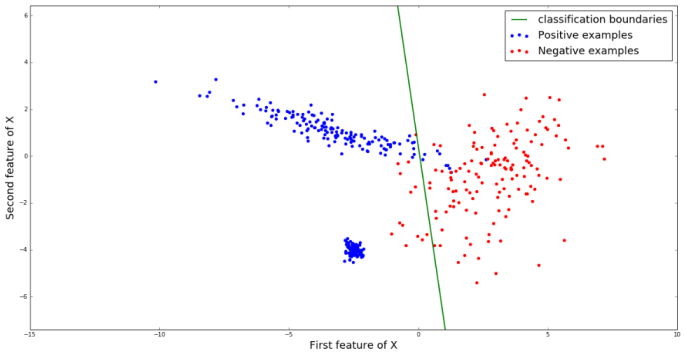


Figure showing the training set in data file C and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

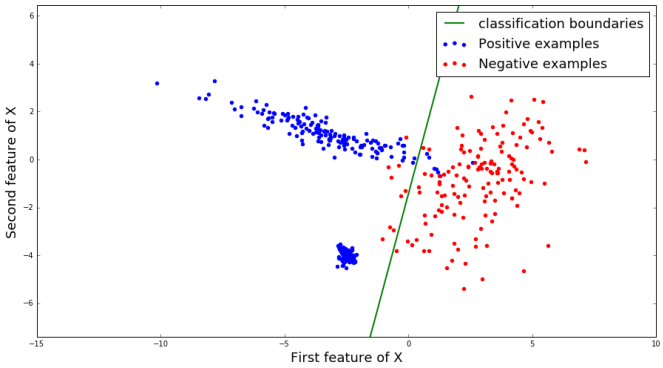


Figure showing the training set in data file C and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

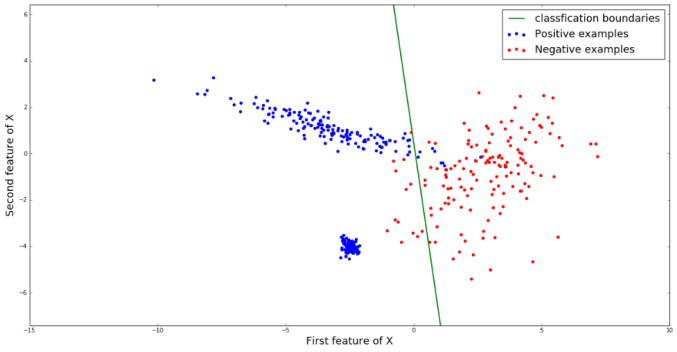


Figure showing the training set in data file C and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

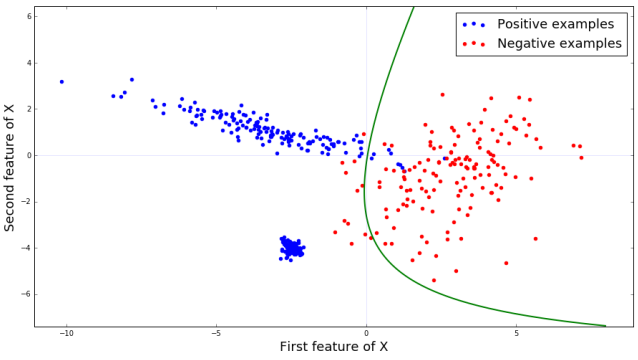


Figure showing the training set in data file C and its classification boundaries using the QDA classifier

### Table of errors

Defining the misclassification error as the ratio of *True Positive* and *True Negative* predictions over all predictions, we establish the following results of error :

| | Train dataset | Test dataset |
|---|---|---|
| LDA | 0.055 | 0.0423 |
| Logistic Regression | 0.04 | 0.0246 |
| Linear Regression | 0.055 | 0.0423 |
| QDA | 0.0525 | 0.0383 |

### Remarks and comments

Finally, in this new dataset (C) and unlike the previous datasets, the QDA classifier is performing poorly compared to the logistic regression which is the new leading classifier. Moreover, the LDA and the linear regression are still performing similarly and the train errors are unexpectedly higher than the test error. We assume that this comes from the data structure. Indeed, the positive labels form two compact clusters where the value of the second feature $x_2$ does not vary very much.
This reason may also explain the poor results in QDA.

Considering all the results, the logistic regression holds roughly good, stable and robust results whereas the LDA and the linear regression perform similarly in the three datasets.

# 3 Appendix

## 3.1 Proof for Exercise 1

Let $D_n = \{(x_1, z_1), ..., (x_n, z_n)\}$ an i.i.d. sample of observations, where $x$ and $z$ are discrete variables taking values respectively in $\{1, ..., K\}$ and $\{1, ..., M\}$. We consider the parameters $\pi$ and $\theta$ defined by the following probabilities $\mathbb{P}(z = m) = \pi_m$ and $\mathbb{P}(x = k|z = m) = \theta_{mk}$.
We can thus compute the log-likelihood, denoting $X = (x_1, ..., x_n)$ and $Z = (z_1, ..., z_n)$ an i.i.d. sample of the defined variables as follows

$$\begin{aligned} log(p(X, Z; \pi, \theta)) &= log\left(\prod_{i=1}^{n} p_\theta(X|Z)p_\pi(z)\right) \quad \text{(by independence and Bayes formula)} \\ &= \sum_{i=1}^{n} log(p_\theta(X|Z)) + log(p_\pi(Z)) \end{aligned}$$

Since the random variables are discrete and live in finite spaces, we can consider for $X_i = j$ (resp. $Z_i = j$) vectors in $[M]$ (resp. $[K]$) of zeros and one in the $j^{th}$ element such that $\{X_i = j\} = \{x_j^{(i)} = 1\}$

This implies the following formula for the log-likelihood

$$\begin{aligned} \mathbf{L}(\pi, \theta) = log(p(x, z; \pi, \theta)) &= \sum_{i=1}^{n} log\left(\prod_j \pi_j^{z_j^{(i)}} \prod_{m,k} \theta_{mk}^{x_k^{(i)} Z_m^{(i)}}\right) \\ &= \sum_{i}^{n} \left(\sum_m z_m^{(i)} log(\pi_m)\right) + \left(\sum_{m,k} x_k^{(i)} z_m^{(i)} log(\theta_{mk})\right) \end{aligned}$$

The log likelihood is a continuous concave function of the parameters $\pi$ and $\theta$ which both lie in a bounded compact sets. Then there exists maximal values of the parameters. Writting down the Lagrangian functional of this optimization program we have :

$$\mathcal{L}(\pi, \theta, \lambda, \mu) = \mathbf{L} - \lambda(\sum_m \pi_m - 1) - \sum_m \mu_m(\sum_k \theta_{mk} - 1)$$

Deriving the Lagrangian with respect to our parameters leads us to the maximum likelihood estimators :

$$\widehat{\pi}_m = \frac{\sum_i z_m^{(i)}}{n} \qquad \widehat{\theta}_{mk} = \frac{\sum_i x_k^{(i)} z_m^{(i)}}{\sum_i z_m^{(i)}}$$

## 3.2 Proof for Exercise 2

## 3.3 Linear Discriminant Analysis formulas (LDA)

Let $D_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ an i.i.d. sample of observations, where $x$ and $y$ are random variables taking values respectively in $\mathbb{R}^2$ and $\{0, 1\}$. We consider the parameters $\pi$, $\mu_0$, $\mu_1$ and $\Sigma$ and we define the following hypothesis : $\mathbb{P}(y = 1) = \pi$ and $(X|y = i) \sim \mathcal{N}(\mu_i, \Sigma)$ .
Therefore, computing the log-likelihood on the sample $D_n$ :

$$\begin{aligned} log(p(X, y; \pi, \mu_0, \mu 1, \Sigma)) &= log\left(\prod_{i=1}^{n} p_{\mu_0,\mu_1,\Sigma}(x_i|y_i)p_\pi(y_i)\right) \quad \text{(by independence and Bayes formula)} \\ &= \sum_{i=1}^{n} log(p_{\mu_0,\mu_1,\Sigma}(x_i|y_i)) + log(p_\pi(y_i)) \end{aligned}$$

Developping the formula for each of these probabilities, we have :

$$p_\pi(y_i) = (\pi)^{y_i} \times (1 - \pi)^{(1-y_i)}$$

$$p(x_i|y_i) = \left[\frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\right)\right]^{(1-y_i)} \times \left[\frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right)^{y_i}\right]$$

This implies the following formula for the log-likelihood

$$\begin{aligned}
\mathbf{L}(\pi,\mu_0,\mu_1,\Sigma) &= log(p(x,y;\pi,\mu_0,\mu_1,\Sigma)) \\
&= \sum_{i=1}^{n}(1-y_i)\left(\log\left(\frac{1}{2\pi|\Sigma|^{1/2}}\right) - \frac{1}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\right) \\
&\quad + y_i\left(\log\left(\frac{1}{2\pi|\Sigma|^{1/2}}\right) - \frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right) + y_i\log(\pi) + (1-y_i)log(1-\pi)
\end{aligned}$$

To estimate the maximum of the likelihood, we compute the gradient of the log-likelihood by computing the partial derivatives with respect $\pi$, $\mu_0$, $\mu_1$, $\Sigma$.
Deriving with respect to $\mu_0$ brings the following formula :

$$\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \mu_0} = \sum_{i=1}^{n}(1-y_i)(-\Sigma^{-1}\mu_0 + (\Sigma^{-1}x_i)^T\mu_0)$$

The equality $\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \mu_0} = 0$ gives the following result :

$$\widehat{\mu}_0 = \frac{\sum_{i=1}^{n} x_i(1-y_i)}{\sum_{i=1}^{n}(1-y_i)}$$

The same calculations hold for $\mu_1$ :

$$\widehat{\mu}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$$

Furthermore, we need to calculate $\widehat{\Sigma}$. We will use the following results (seen during the lectures) in the computation of such an estimator.

$$\frac{\partial log(|A|)}{\partial A} = A^{-1} \quad \text{for} \quad A \in \mathcal{S}_n^+(\mathbb{R})$$

$$\frac{\partial (x^T A x)}{\partial A} = xx^T \quad \text{for} \quad A \in \mathcal{M}_n(\mathbb{R})$$

With all this results, we have :

$$\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\left(\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0) + y_i(x_i-\mu_1)^T(x_i-\mu_1)\right)$$

From the equality $\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \Sigma^{-1}} = 0$, we deduce the formula for the estimator $\widehat{\Sigma}$ :

$$\widehat{\Sigma} = \frac{\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0) + y_i(x_i-\mu_1)^T(x_i-\mu_1)}{n}$$

Concerning the last estimator, the derivation with respect to $\pi$ provides the formula :

$$\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \pi} = \sum_{i=1}^{n}\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}$$

From the equation

$$\frac{\partial L(x,y,\mu_0,\mu_1,\pi,\Sigma)}{\partial \pi} = 0$$

, we conclude to the estimator :

$$\widehat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Now concerning the equation of the classification boundaries ($\mathbb{P}(y = 1|x) = 0.5$) we result to the following equation :

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

In the following is the proof of such equation :

$$\mathbb{P}(y = 1|x) = 0.5 = \mathbb{P}(y = 0|x)$$

Therefore we can use Bayes formula to obtain :

$$\mathbb{P}(x|y = 1)\mathbb{P}(y = 1) = \mathbb{P}(x|y = 0)\mathbb{P}(y = 0)$$

We can rewrite it :

$$F_{\mu_1,\Sigma}(x)\pi = F_{\mu_0,\Sigma}(x)(1 - \pi)$$

where the functions $F_{\mu,\Sigma}$ are the density of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$.

We then have :

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

Moving to the logarithm

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

We can simplify and obtain :

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

### 3.4 Quadratic Discriminant Analysis formulas (QDA)

The proof of the equations are similar to the QDA model, except that we consider two covariance matrices. The log-likelihood of the model can be written as the following :

$$\begin{aligned}
\mathbf{L}(\pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) &= log(p(x, y; \pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1)) \\
&= \sum_{i=1}^{n} (1 - y_i) \left( \log \left( \frac{1}{2\pi|\Sigma_0|^{1/2}} \right) - \frac{1}{2}(x_i - \mu_0)\Sigma_0^{-1}(x_i - \mu_0) \right) \\
&\quad + y_i \left( \log \left( \frac{1}{2\pi|\Sigma_1|^{1/2}} \right) - \frac{1}{2}(x_i - \mu_1)\Sigma_1^{-1}(x_i - \mu_1) \right) + y_i \log(\pi) + (1 - y_i) log(1 - \pi)
\end{aligned}$$

As we can see, the same reasoning holds for all estimators except $\Sigma_1 \neq \Sigma_0$ in this model. Therefore :

$$\widehat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\widehat{\mu}_0 = \frac{\sum_{i}^{n} x_i(1 - y_i)}{\sum_{i=1}^{n}(1 - y_i)}$$

$$\widehat{\mu}_1 = \frac{\sum_{i}^{n} x_i(y_i)}{\sum_{i=1}^{n} y_i}$$

Let us now consider the derivation with respect to $\Sigma_1$ and $\Sigma_0$ using the same arguments as before :

$$\frac{\partial L(x, y; \mu_0, \mu_1, \pi, \Sigma_0, \Sigma_1)}{\partial \Sigma_0^{-1}} = \frac{1}{2}\sum_{i=1}^{n}(1-y_i)\Sigma_0 - \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0)$$

Solving $\frac{\partial L(x,y;\mu_0,\mu_1,\pi,\Sigma_0,\Sigma_1)}{\partial \Sigma_0^{-1}} = 0$, we have :

$$\widehat{\Sigma}_0 = \frac{\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0)}{\sum_{i=1}^{n}(1-y_i)}$$

Same thing holds for $\Sigma_1$ :

$$\widehat{\Sigma}_1 = \frac{\sum_{i=1}^{n}(y_i)(x_i-\mu_1)^T(x_i-\mu_1)}{\sum_{i=1}^{n} y_i}$$

Furthermore, we provide the equation for the conic representing the classification boundaries in QDA characterized by $\mathbb{P}(y=1|x) = 0.5$.
As in the LDA proof, we use the fact that $\mathbb{P}(y=1|x) = 0.5 = \mathbb{P}(y=0|x)$
Bayes formula gives :

$$\mathbb{P}(x|y=1)\mathbb{P}(y=1) = \mathbb{P}(x|y=0)\mathbb{P}(y=0)$$

We can rewrite it with density functions :

$$F_{\mu_1,\Sigma_1}(x)\pi = F_{\mu_0,\Sigma_0}(x)(1-\pi)$$

where the functions $F_{\mu,\Sigma}$ are the density of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$
Pursuing the calculus, we then have :

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

Developing the expression gives the following result :

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

## 3.5 Classification boundaries for Test datasets

## Results for dataset A

### Data and classification boundaries for LDA

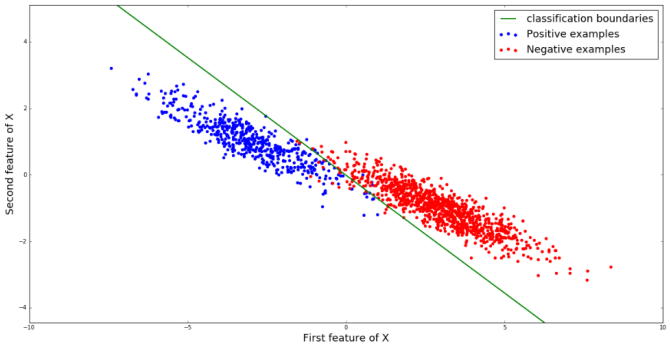


Figure showing the test set in data file A and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

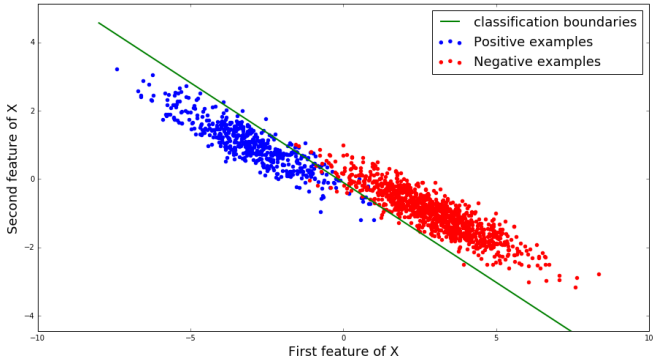


Figure showing the test set in data file A and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

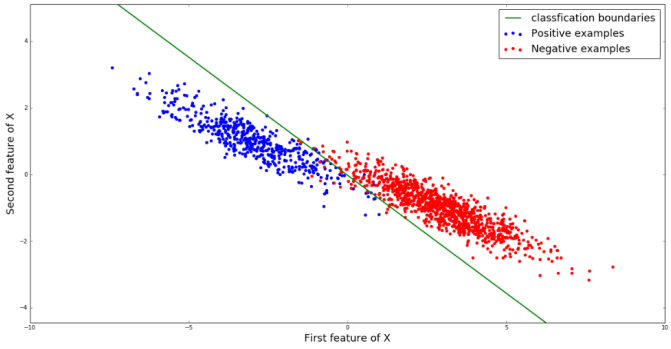


Figure showing the test set in data file A and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

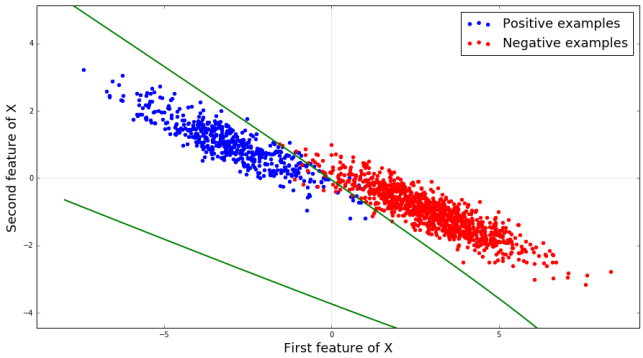


Figure showing the test data file A and its classification boundaries using the QDA classifier

## Results for dataset B

### Data and classification boundaries for LDA

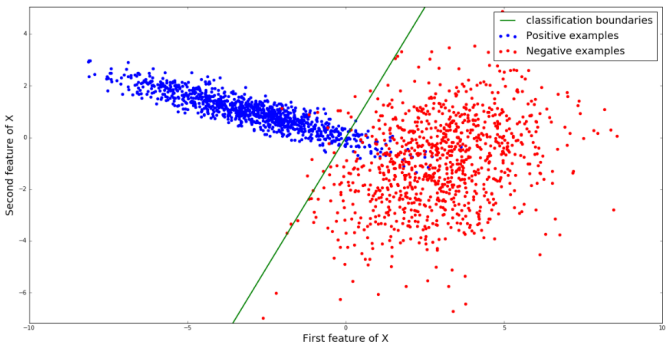


Figure showing the test set in data file B and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

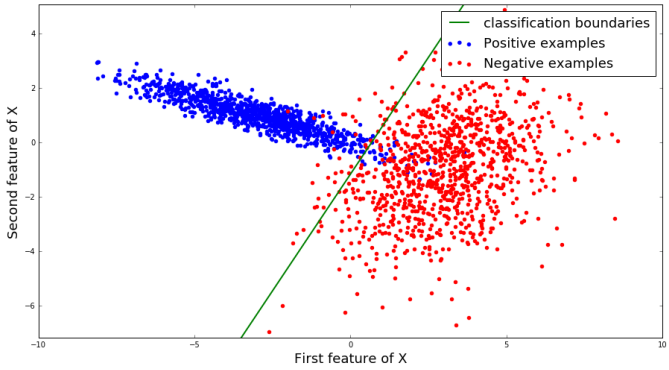


Figure showing the test set in data file B and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

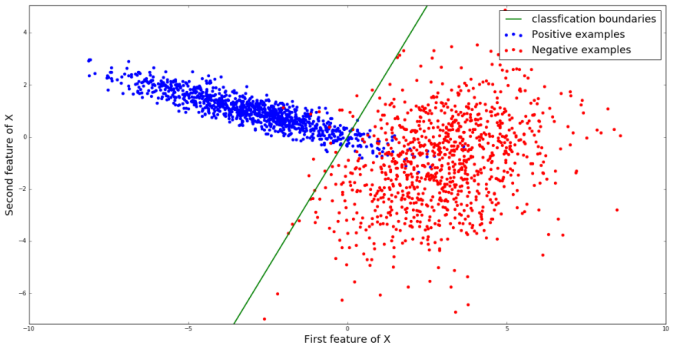


Figure showing the test set in data file B and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

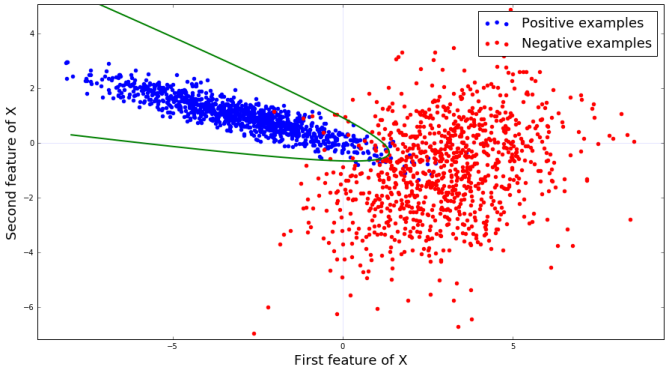


Figure showing the test set in data file B and its classification boundaries using the QDA classifier

## Results for dataset C

### Data and classification boundaries for LDA

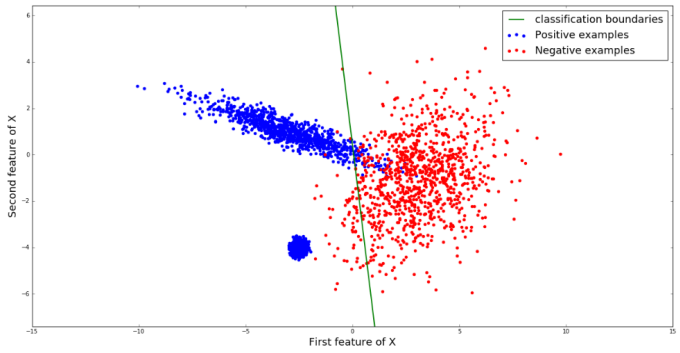


Figure showing the test set in data file C and its classification boundaries using the LDA classifier

### Data and classification boundaries for Logistic Regression

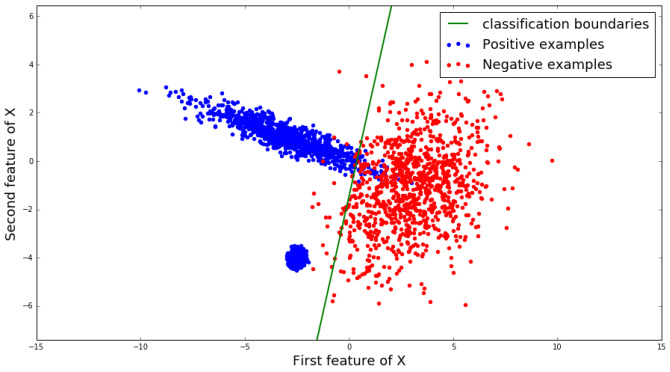


Figure showing the test set in data file C and its classification boundaries using the logistic regression classifier

### Data and classification boundaries for Linear Regression

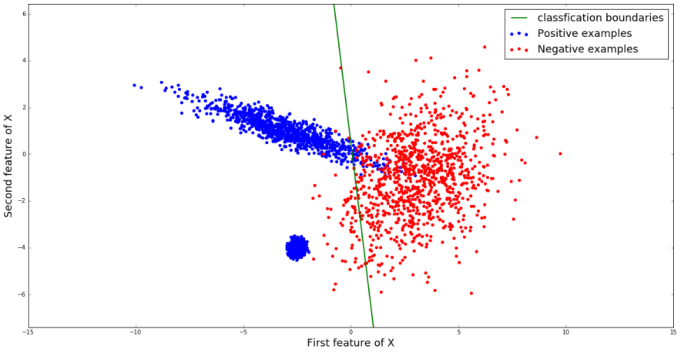


Figure showing the test set in data file C and its classification boundaries using the linear regression classifier

### Data and classification boundaries for QDA

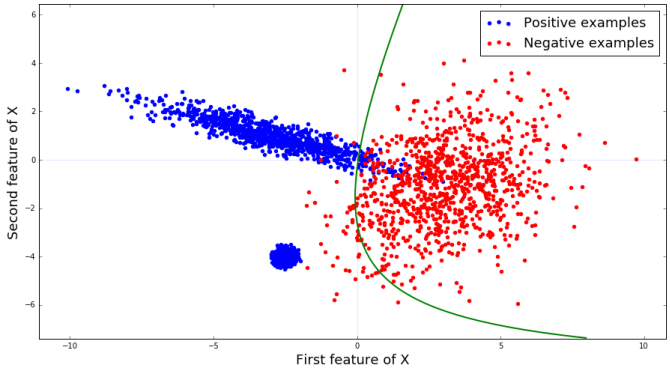


Figure showing the test data file C and its classification boundaries using the QDA classifier