Master M2 MVA 2016/2017 – Graphical models

Exercices due October 31st, 2016.

SOLUTIONS

1 Learning in discrete models

Consider the following model : z and x are discrete variables taking respectively M and K different values with $p(z = m) = \pi_m$ and $p(x = k|z = m) = \theta_{mk}$. Compute the maximum likelihood estimator for π and θ based on an i.i.d. sample of observations.

Using the same notations as in class, we consider an i.i.d. sample of size N of observatins of the form (z_m^n, x_k^n) with $n \in \{1, \dots, N\}$, $k \in \{1, \dots, K\}$ and $m \in \{1, \dots, M\}$, with $z_m^n = 1$ if $z = m$ for the nth exemple (and 0 else), $x_k^n = 1$ if $x = k$ for the nth exemple (and 0 else).

Given the i.i.d. assumption, the log-likelihood is of the form :

$$\begin{aligned}
\ell(\pi, \theta) &= \log \prod_{n=1}^{N} \prod_{m=1}^{M} \pi_m^{z_m^n} \prod_{n=1}^{N} \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{z_m^n x_k^n} \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} z_m^n \log \pi_m + \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} z_m^n x_k^n \log \theta_{mk} \\
&= \sum_{m=1}^{M} \left(\sum_{n=1}^{N} z_m^n \right) \log \pi_m + \sum_{k=1}^{K} \sum_{m=1}^{M} \left(\sum_{n=1}^{N} z_m^n x_k^n \right) \log \theta_{mk} \\
&= \sum_{m=1}^{M} N_m \log \pi_m + \sum_{k=1}^{K} \sum_{m=1}^{M} N_{mk} \log \theta_{mk}
\end{aligned}$$

where N_m is the number of examples such that $z = m$ and N_{mk} is the number of examples such that $z = m$ and $x = k$.

The function to maximize is a sum of two independent terms, that can be maximized separately. The first term was dealt with in class and we showed that the solution is $\boxed{\pi_m = N_m/N}$.

For the second term, we need to maximize $\sum_{k=1}^{K} \sum_{m=1}^{M} N_{mk} \log \theta_{mk}$ under the constraints that $\forall k, m,\ \theta_{mk} \geqslant 0$ and $\forall m,\ \sum_{k=1}^{K} \theta_{mk} = 1$. We can ignore the positivity constraints given that the likelihood decreases to $-\infty$ when the parameters approach 0. The

Lagrangian takes the form :

$$\mathcal{L}(\theta, \lambda) = \sum_{k=1}^{K} \sum_{m=1}^{M} N_{mk} \log \theta_{mk} - \sum_{m=1}^{M} \lambda_m \left(\sum_{k=1}^{K} \theta_{mk} - 1 \right).$$

The maximisation of $\mathcal{L}(\theta, \lambda)$ with respect to θ decouples, since :

$$\frac{\partial \mathcal{L}}{\partial \theta_{mk}} = \frac{N_{mk}}{\theta_{mk}} - \lambda_m$$

The Lagrangian is therefore maximized when $\frac{\partial \mathcal{L}}{\partial \theta_{mk}} = \frac{N_{mk}}{\theta_{mk}} - \lambda_m = 0$, or equivalently when $\theta_{mk} = N_{mk}/\lambda_m$ (which is positive). The dual objective is therefore equal to :

$$\begin{aligned} g(\lambda) &= \max_{\theta} \mathcal{L}(\theta, \lambda) \\ &= \sum_{k=1}^{K} \sum_{m=1}^{M} N_{mk} \log N_{mk}/\lambda_m - \sum_{m=1}^{M} \lambda_m \left(\sum_{k=1}^{K} N_{mk}/\lambda_m - 1 \right) \\ &= \sum_{k=1}^{K} \sum_{m=1}^{M} N_{mk} \log N_{mk} - \sum_{m=1}^{M} N_m \log \lambda_m - \sum_{m=1}^{M} (N_m - \lambda_m) \end{aligned}$$

When minimizing with respect to λ_m we obtain $\lambda_m = N_m$ (which can also be obtained directly via the constraint $\sum_k \theta_{mk} = 1$).

The solution is thus $\boxed{\theta_{mk} = N_{km}/N_m}$

2 Linear classification

Given an i.i.d. sample of size N composed of observations of the form (y_n, x_n) where $n \in \{1, \ldots, N\}$, $y_n \in \{0, 1\}$ and $x_n \in \mathbb{R}^d$ (here, $d = 2$).

2.1 Linear Discriminant Analysis

The maximization of the likelihood leads to

$$\begin{aligned}
\pi &= \frac{n_1}{N} \\
\mu_1 &= \frac{1}{n_1}\sum_n y_n x_n \\
\mu_0 &= \frac{1}{n_0}\sum_n (1-y_n) x_n \\
\Sigma &= \frac{1}{N}\sum_n (x_n - y_n\mu_1 - (1-y_n)\mu_0)(x_n - y_n\mu_1 - (1-y_n)\mu_0)^\top \\
\Sigma &= \frac{n_1}{N}\Sigma_1 + \frac{n_0}{N}\Sigma_0 \text{ (equivalently...)}
\end{aligned}$$

with $n_1 = \sum_n y_n$, $n_0 = N - n_1 = \sum_n (1-y_n)$, the empirical averages in each class $\mu_1 = \sum_n y_n x_n$ and $\mu_0 = \sum_n (1-y_n) x_n$, and the class specific covariance matrices $\Sigma_1 = \frac{1}{n_1}\sum_n y_n (x_n - \mu_1)(x_n - \mu_1)^\top$ and $\Sigma_0 = \frac{1}{n_0}\sum_n y_n (x_n - \mu_0)(x_n - \mu_0)^\top$.

Following the calculations done in class, we obtain $p(y=1|x) = \sigma(\beta^\top x + \gamma)$ with

$$\begin{aligned}
\beta &= \Sigma^{-1}(\mu_1 - \mu_0) \\
\gamma &= -\frac{1}{2}(\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 + \mu_0) + \log\frac{\pi}{1-\pi}
\end{aligned}$$

2.2 Quadratic Discriminant Analysis

Upon maximizing the likelihood, we obtain the same estimators as for LDA, except that the matrices Σ_0 et Σ_1 are not combined to form Σ.

With the same calculations as in class, we obtain $p(y=1|x) = \sigma(\frac{1}{2}x^\top Q x + \beta^\top x + \gamma)$ with

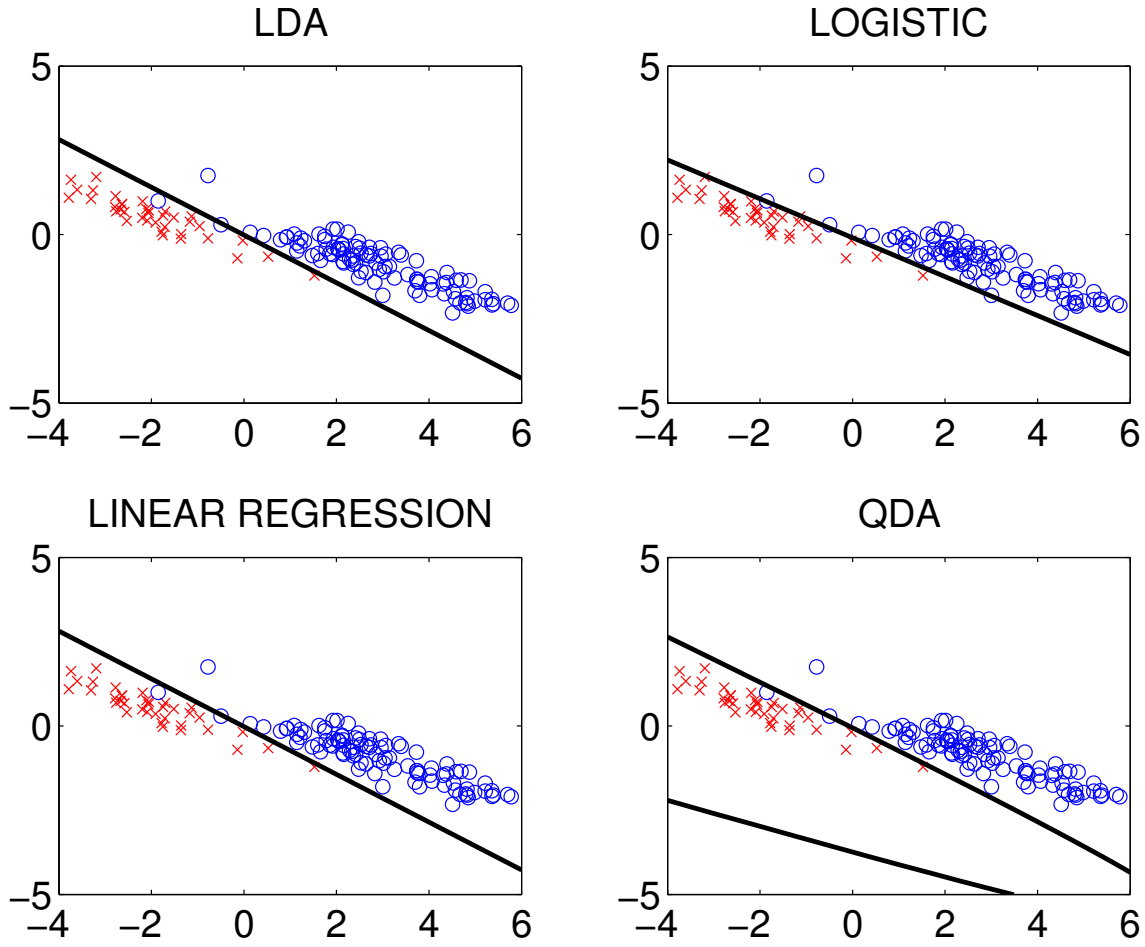
$$\begin{aligned}
Q &= \Sigma_2^{-1} - \Sigma_1^{-1} \\
\beta &= \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0 \\
\gamma &= \frac{1}{2}\mu_0^\top \Sigma_0^{-1}\mu_0 - \frac{1}{2}\mu_1^\top \Sigma_1^{-1}\mu_1 + \log\frac{\pi}{1-\pi} + \frac{1}{2}\log\det\Sigma_0 - \frac{1}{2}\log\det\Sigma_1
\end{aligned}$$

The conic is either an ellipse or an hyperbole, which can be plotted using the canonical representation :

$$\frac{1}{2}(x - Q^{-1}\beta)^\top Q (x - Q^{-1}\beta) = \frac{1}{2}\beta^\top Q^{-1}\beta - \gamma$$

and a diagonalization of Q (see the code for the details).

2.3 Model comparisons - Dataset A



Misclassification error in %

	Train	Test
LDA	1.3	2.0
Linear reg.	1.3	2.1
Logistic reg.	0.0	3.4
QDA	0.6	2.0

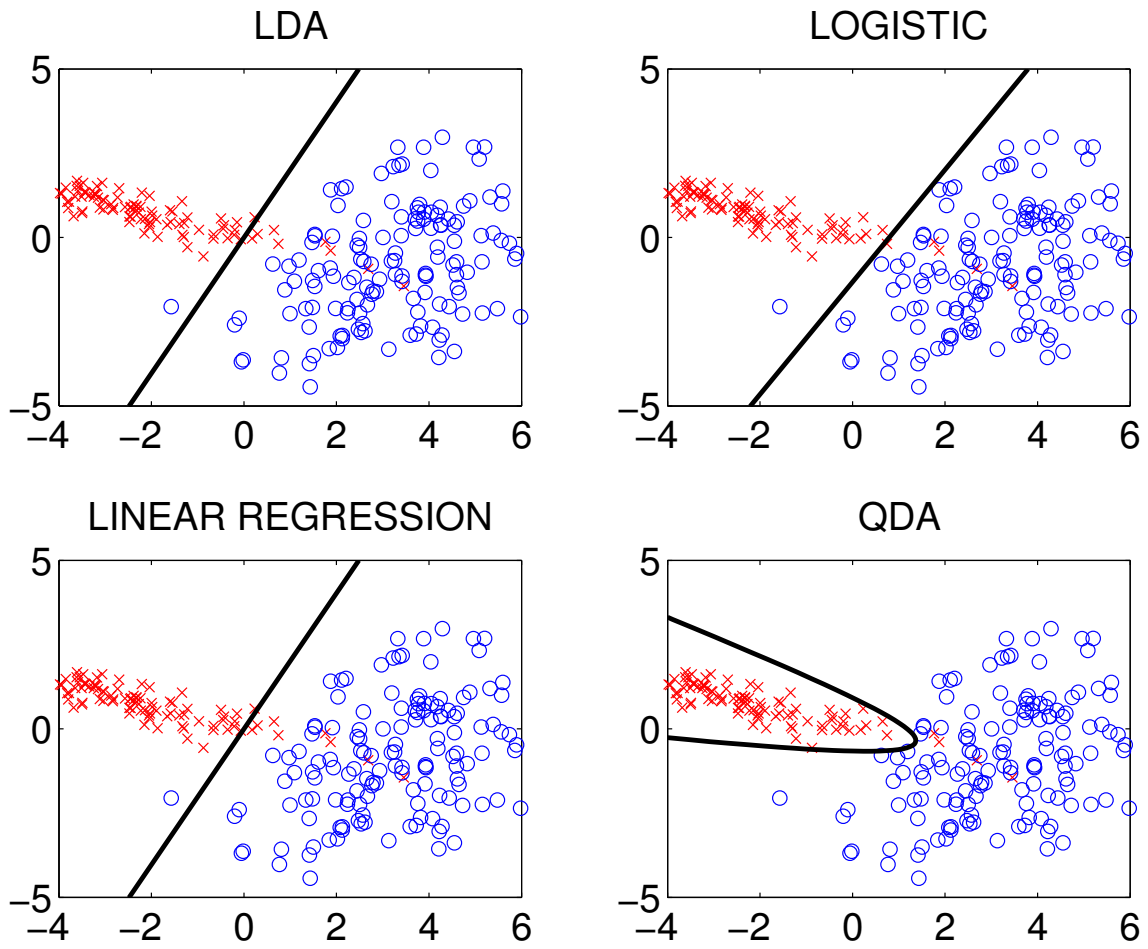
The points to note :

— All methods are behaving well, with the exception of logistic regression which overfits slightly. Indeed, it turns out that, by chance, the training data are separable, which explains why the training error is equal to 0 and why the magnitude of the learned parameters is so large.

— By contrast, the data are generated exactly according to the LDA model with each class corresponding to one of two Gaussians with the same covariance matrix. Since LDA is exactly the right model, it is not very surprising that this model leads to the smallest amount of error.

— The performance of QDA is essentially identical : if we were trying to learn in a higher dimensional space, QDA could suffer (compared to LDA which exploits the a priori knowledge that the covariance matrices are equal) given that QDA would have to learn more parameters, which could lead to some overfitting compared to LDA, but in dimension 2 this has no impact and the improvement of the training error observed does not translate into an improvement of the test error.

2.4 Model comparisons - Dataset B



Misclassification error in %

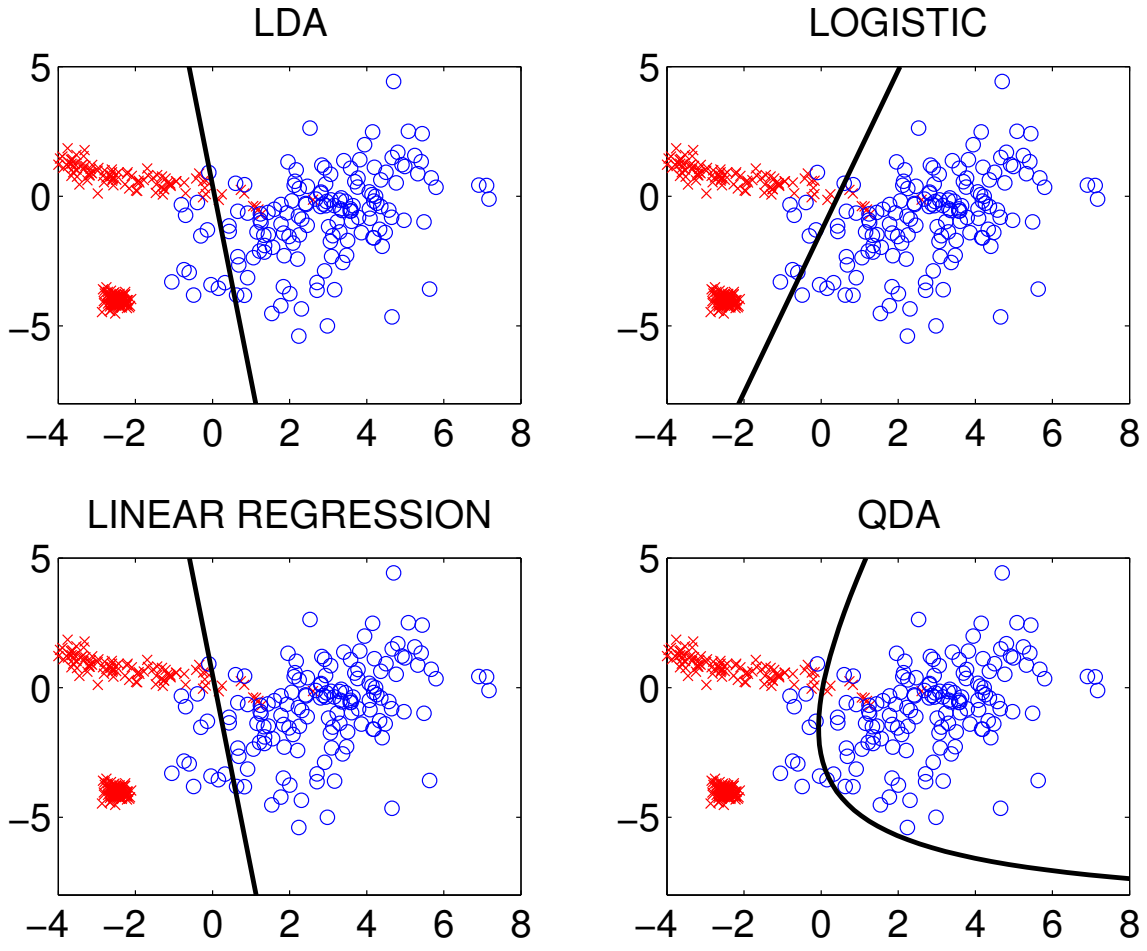
	Train	Test
LDA	3.0	4.2
Linear reg.	3.0	4.2
Logistic reg.	2.0	4.3
QDA	1.3	2.0

The main points are :

— The data are here generated by a model in the family corresponding to QDA : each class is a Gaussian cloud with its own covariance matrix ; moreover in this case the covariance matrices are very different. The best performance is clearly attained by QDA, which is natural given that it is the correct model, and that it is moreover the only model which is nonlinear with respect to the data.

— By comparison LDA and the regressions leading to linear classifiers have much lower precision since they are not able to separate the tip of the red class which is close to the blue class. For logistic regression or linear regression, this problem could be tackled by using a quadratic mapping of the data as in kernel methods (corresponding for linear regression to performing quadratic regression), which would be quite easy to do here since the input space is of low dimension.

2.5 Model comparisons - Dataset C



Misclassification error in %

	Train	Test
LDA	5.5	4.2
Linear reg.	5.5	4.2
Logistic reg.	4.0	2.3
QDA	5.3	3.8

The points to note :

- The data are generated here by a model which is quite different than LDA or QDA, in which one of the two classes contains two clusters. Hence the relatively high error rate of LDA and QDA. The presence of a small red cluster also biases linear regression, which is anyway not so well adapted to solving classification problems.
- By contrast, we see here the advantage of using a method like logistic regression that models directly $p(y|x)$ instead of modeling it indirectly via a model of $p(x)$ which could be wrong. Logistic regression yields a test error which is smaller than the train error ; this is unusual, but indicates that not overfitting is occurring.
- In almost all experiments (for datasets A, B and C), the test error is larger than the training error, which is the expected and the most common situation. In the cases in which the gap is large between these two quantities, it should be interpreted as a sign that (possibly mild) overfitting is occurring.

2.6 A remark on LDA and logistic regression

It can be noticed that the performance level of LDA and linear regression are relatively close to each other. It is perhaps even more striking that the corresponding linear separator seem to be aligned. This is in fact exactly true for binary classification : it is possible to prove, just using linear algebra, that the the vectors of parameters obtained by these two methods are actually colinear ! You can try to prove it...

3 Numerical values of the parameters

Data A

LDA

$$\mu_0 = \begin{pmatrix} 2.90 \\ -0.89 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.69 \\ 0.87 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 2.44 & -1.13 \\ -1.13 & 0.61 \end{pmatrix} \quad \beta = \begin{pmatrix} -6.62 \\ -9.35 \end{pmatrix} \quad \gamma = -0.14$$

Logistic regression

$$\beta = \begin{pmatrix} -648.82 \\ -1124.07 \end{pmatrix} \quad \gamma = -107.66$$

Note that in this case the data is perfectly separable so the minimum of the objective is only attained in the limit where $\|\beta\| \to \infty$. Therefore the value obtained depends on the optimization algorithm and the stopping criterion you chose. We considered that the returned solution was correct if it was perfectly separating the training data. The proper thing to do here would have been to add some regularization (but this was of course not expected since this is beyond the scope of this course).

Linear regression

$$\beta = \begin{pmatrix} -0.26 \\ -0.37 \end{pmatrix} \quad \gamma = -0.01$$

QDA

$$\mu_0 = \begin{pmatrix} 2.90 \\ -0.89 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.69 \\ 0.87 \end{pmatrix} \quad \Sigma_0 = \begin{pmatrix} 2.31 & -1.05 \\ -1.05 & 0.58 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2.70 & -1.30 \\ -1.30 & 0.69 \end{pmatrix}$$

$$Q = \begin{pmatrix} -1.52 & -3.03 \\ -3.03 & -5.72 \end{pmatrix} \quad \beta = \begin{pmatrix} -7.37 \\ -10.87 \end{pmatrix} \quad \gamma = -0.63$$

Data B

LDA

$$\mu_0 = \begin{pmatrix} 3.34 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 3.35 & -0.14 \\ -0.14 & 1.74 \end{pmatrix} \quad \beta = \begin{pmatrix} -1.92 \\ 0.95 \end{pmatrix} \quad \gamma = 0.00$$

Logistic regression

$$\beta = \begin{pmatrix} -1.71 \\ 1.02 \end{pmatrix} \quad \gamma = 1.35$$

Linear regression

$$\beta = \begin{pmatrix} -0.10 \\ 0.05 \end{pmatrix} \quad \gamma = 0.00$$

QDA

$$\mu_0 = \begin{pmatrix} 3.34 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix} \quad \Sigma_0 = \begin{pmatrix} 2.54 & 1.06 \\ 1.06 & 2.96 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 4.15 & -1.33 \\ -1.33 & 0.52 \end{pmatrix}$$

$$Q = \begin{pmatrix} -0.96 & -3.85 \\ -3.85 & -11.06 \end{pmatrix} \quad \beta = \begin{pmatrix} -2.28 \\ 1.46 \end{pmatrix} \quad \gamma = 3.37$$

Data C

LDA

$$\mu_0 = \begin{pmatrix} 2.79 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.94 \\ -0.96 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 2.88 & -0.63 \\ -0.63 & 5.20 \end{pmatrix} \quad \beta = \begin{pmatrix} -2.05 \\ -0.27 \end{pmatrix} \quad \gamma = 0.11$$

Logistic regression

$$\beta = \begin{pmatrix} -2.20 \\ 0.71 \end{pmatrix} \quad \gamma = 0.96$$

Linear regression

$$\beta = \begin{pmatrix} -0.13 \\ -0.02 \end{pmatrix} \quad \gamma = 0.01$$

QDA

$$\mu_0 = \begin{pmatrix} 2.79 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.94 \\ -0.96 \end{pmatrix} \quad \Sigma_0 = \begin{pmatrix} 2.90 & 1.25 \\ 1.25 & 2.92 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2.87 & -1.76 \\ -1.76 & 6.56 \end{pmatrix}$$

$$Q = \begin{pmatrix} 0.00 & -0.29 \\ -0.29 & 0.24 \end{pmatrix} \quad \beta = \begin{pmatrix} -2.67 \\ 0.35 \end{pmatrix} \quad \gamma = 0.11$$