M2 MVA, ENS CACHAN

Probabilistic Graphical Models - Homework Assignment 1

Chia-Man Hung

October 27, 2016

1 LEARNING IN DISCRETE GRAPHICAL MODELS

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.

Consider N observations $X_1, X_2, ..., X_N$ of x and N observations $Z_1, Z_2, ..., Z_N$ of z. We denote $x_i (i = 1, 2, ..., N)$ (resp. z_i) the K-dimensional (resp. M-dimensional) vectors of 0s and 1s representing X_i (resp. Z_i) for which the event $X_i = k$ (resp. $Z_i = m$) corresponds to the event $\{x_{ik} = 1 and x_{il} = 0, \forall l \neq k\}$ (resp. $\{z_{im} = 1 and z_{il} = 0, \forall l \neq m\}$).

First, we compute the likelihood.

$$\begin{aligned}
\mathscr{L}(\pi, \theta) &= p(x_1, ..., x_N, z_1, ..., z_N; \pi, \theta) \\
&= \prod_{i=1}^{N} p(x_i, z_i; \pi, \theta) \\
&= \prod_{i=1}^{N} p(x_i|z_i; \pi, \theta) p(z_i; \pi, \theta) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{m=1}^{M} \theta_{mk}^{x_{ik} z_{im}} \pi_m^{z_{im}}
\end{aligned} \tag{1.1}$$

It is more convenient to work with the log-likelihood.

$$\begin{aligned} l(\pi,\theta) &= log\mathscr{L}(\pi,\theta) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} x_{ik}z_{im}log\theta_{mk} + K\sum_{i=1}^{N}\sum_{m=1}^{M} z_{im}log\pi_m \\ &= \sum_{k=1}^{K}\sum_{m=1}^{M}(\sum_{i=1}^{N} x_{ik}z_{im})log\theta_{mk} + K\sum_{m=1}^{M}(\sum_{i=1}^{N} z_{im})log\pi_m \end{aligned} \tag{1.2}$$

We need to minimize

$$f(\pi,\theta) = -\sum_{k=1}^{K}\sum_{m=1}^{M}(\sum_{i=1}^{N} x_{ik}z_{im})log\theta_{mk} - K\sum_{m=1}^{M}(\sum_{i=1}^{N} z_{im})log\pi_m \tag{1.3}$$

subject to the constraints $\sum_{m=1}^{M}\pi_m = 1, \sum_{k=1}^{K}\theta_{mk} = 1, \forall m \in \{1,...,M\}$.

The Lagrangian of this problem is

$$L(\pi,\theta,\lambda,\mu) = -\sum_{k=1}^{K}\sum_{m=1}^{M}(\sum_{i=1}^{N} x_{ik}z_{im})log\theta_{mk} - K\sum_{m=1}^{M}(\sum_{i=1}^{N} z_{im})log\pi_m + \lambda(\sum_{m=1}^{M}\pi_m - 1) + \sum_{m-1}^{M}\mu_m(\sum_{k=1}^{K}\theta_{mk} - 1) \tag{1.4}$$

Clearly, as $n_k \geq 0, k = 1,...,K$ and $n'_m \geq 0, m = 1,...,M$, f is convex and this problem is a convex optimization problem. Moreover, it is trivial that there exist π,θ verifying the constraints, so by Slater's constraint qualification, the problem has strong duality property. Therefore, we have

$$\underset{\pi,\theta}{min} f(\pi,\theta) = \underset{\lambda,\mu}{max}\,\underset{\pi,\theta}{min} L(\pi,\theta,\lambda,\mu) \tag{1.5}$$

As $L(\pi,\theta,\lambda,\mu)$ is convex with respect to π,θ, to find $\underset{\pi,\theta}{min} L(\pi,\theta,\lambda,\mu)$, it suffices to take derivatives with respect to π_m,θ_{mk}. This yields

$$\frac{\partial L}{\partial \pi_m} = -K\frac{\sum_{i=1}^{N} z_{im}}{\pi_m} + \lambda = 0, m = 1,...,M.$$

or

$$\pi_m = \frac{K}{\lambda}\sum_{i=1}^{N} z_{im}, m = 1,...,M. \tag{1.6}$$

Substituting this into the constraint $\sum_{m=1}^{M}\pi_m = 1$ we obtain $\sum_{m=1}^{M}\sum_{i=1}^{N} z_{im} = \frac{\lambda}{K}$, yielding $\lambda = KN$.
From this and the previous equation, we get finally

$$\widehat{\pi}_m = \frac{\sum_{i=1}^{N} z_{im}}{N}, m = 1,...,M. \tag{1.7}$$

$$\frac{\partial L}{\partial \theta_{mk}} = -\frac{\sum_{i=1}^{N} x_{ik} z_{im}}{\theta_{mk}} + \mu_m = 0, m = 1, ..., M, k = 1, ..., K.$$

or

$$\theta_{mk} = \frac{\sum_{i=1}^{N} x_{ik} z_{im}}{\mu_m}, m = 1, ..., M, k = 1, ..., K. \tag{1.8}$$

Similarly, $1 = \sum_{k=1}^{K} \theta_{mk} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{N} x_{ik} z_{im}}{\mu_m} = \frac{\sum_{i=1}^{N} z_{im}}{\mu_m}, m = 1, ..., M$, yielding $\mu_m = \sum_{i=1}^{N} z_{im}$

$$\widehat{\theta}_{mk} = \frac{\sum_{i=1}^{N} x_{ik} z_{im}}{\sum_{i=1}^{N} z_{im}}, m = 1, ..., M, k = 1, ..., K. \tag{1.9}$$

2 LINEAR CLASSIFICATION

1. Generative model (LDA)
y ~Bernoulli(π), $x|y = i$ Normal(μ_i, Σ)
(a)

$$p(y) = \pi^y (1-\pi)^{1-y} \tag{2.1}$$

$$p(x|y=0) = \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)) \tag{2.2}$$

$$p(x|y=1) = \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)) \tag{2.3}$$

$$\begin{aligned} l(\pi, \mu_0, \mu_1, \Sigma) &= log \prod_{i=1}^{N} p(x_i, y_i; \pi, \mu_0, \mu_1, \Sigma) \\ &= log \prod_{i=1}^{N} p(x_i|y_i; \mu_0, \mu_1, \Sigma) p(y_i; \pi) \\ &= \sum_{i=1}^{N} (log \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{1}{2}(x_i - \mu_{y_i})^T \Sigma (x_i - \mu_{y_i})) \\ &\quad + y_i log\pi + (1-y_i) log(1-\pi)) \\ &= -Nlog2\pi - \frac{N}{2} log(det\Sigma) - \frac{1}{2} \sum_{i=1}^{N} (x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}) \\ &\quad + \sum_{i=1}^{N} (y_i log\pi + (1-y_i) log(1-\pi)) \end{aligned} \tag{2.4}$$

$$\frac{\partial l}{\partial \mu_0} = \sum_{i=1}^{N} \mathbb{1}\{y_i = 0\} \Sigma^{-1} (\mu_0 - x_i) = 0 \Rightarrow \widehat{\mu}_0 = \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = 0\} x_i}{\sum_{i=1}^{N} \mathbb{1}\{y_i = 0\}} \tag{2.5}$$

Similarly,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N} \mathbb{1}\{y_i = 1\} x_i}{\sum_{i=1}^{N} \mathbb{1}\{y_i = 1\}} \tag{2.6}$$

$$\frac{\partial l}{\partial \pi} = \sum_{i=1}^{N} \frac{y_i}{\pi} - \frac{1-y_i}{1-\pi} = 0 \Rightarrow \hat{\pi} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{2.7}$$

To compute the gradient of l with respect to Σ, we follow the lecture notes 1.4.7. By replacing the scalar $(x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i})$ with $Trace((x_i - \mu_{y_i})^T \Sigma^{-1} (x_i - \mu_{y_i}))$ and by computing $\bigtriangledown log(detA) = A^{-1}$, we obtain

$$\bigtriangledown_{\Sigma^{-1}}(l) = -\frac{N}{2}\Sigma + \frac{N}{2}\tilde{\Sigma} \tag{2.8}$$

where $\tilde{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T$.
Finally, we obtain

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T \tag{2.9}$$

(b)

$$\begin{aligned}
p(y=1|x) &= \frac{p(y=1,x)}{p(x)} \\
&= \frac{p(x,y=1)}{p(x,y=0)+p(x,y=1)} \\
&= \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \\
&= \frac{\mathcal{N}(x;\mu_1,\Sigma)\pi}{\mathcal{N}(x;\mu_0,\Sigma)(1-\pi)+\mathcal{N}(x;\mu_1,\Sigma)\pi} \\
&= \frac{\pi exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}{(1-\pi)exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))+\pi exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))} \\
&= \frac{1}{1+\frac{1-\pi}{\pi}exp((-\frac{1}{2})(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))} \\
&= \frac{1}{1+exp(-(\alpha+\beta x))}
\end{aligned} \tag{2.10}$$

where $\alpha = log\frac{\pi}{1-\pi} + \frac{1}{2}(\mu_0+\mu_1)^T\Sigma(\mu_0-\mu_1), \beta = (\mu_1-\mu_0)^T\Sigma^{-1}$.

This is exactly the form the the logistic regression use to model $p(y=1|x)$. Note that the offset α can be absorbed into x by considering $\begin{pmatrix} x \\ 1 \end{pmatrix}$ instead of x.

(c)

$$
\begin{aligned}
p(y=1|x) = 0.5 &\Leftrightarrow exp(-(\alpha+\beta x)) = 1 \\
&\Leftrightarrow \alpha + \beta x = 0 \\
&\Leftrightarrow x_1 = \frac{-\alpha - \beta_0 x_0}{\beta_1}
\end{aligned} \tag{2.11}
$$

Figure 2.1: LDA + Training data. Left: A, Middle: B, Right: C

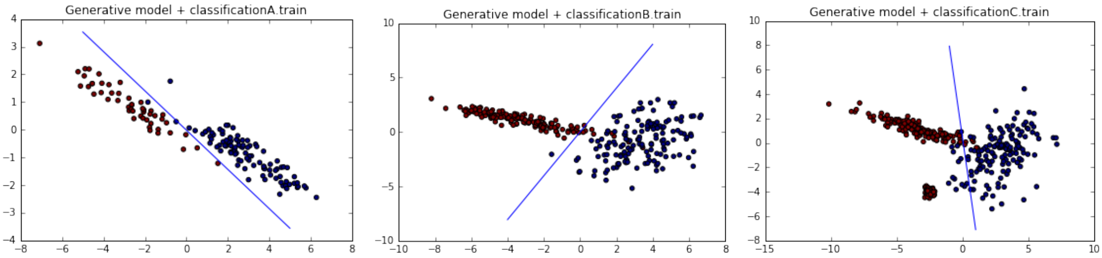
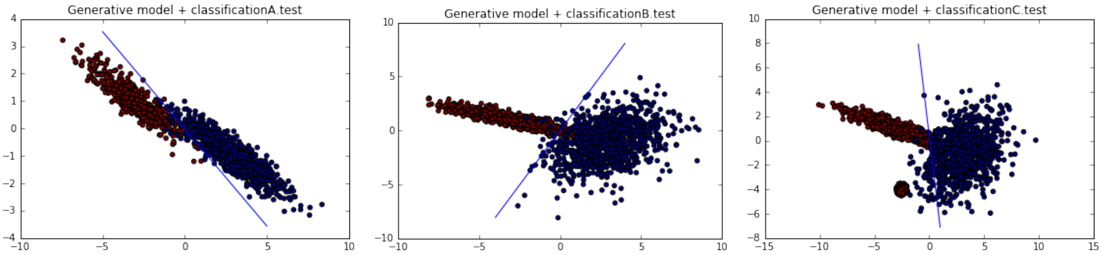


Figure 2.2: LDA + Test data. Left: A, Middle: B, Right: C



2. Logistic regression

(a)

$$w_{logreg,A} = \begin{pmatrix} -2337 \\ -3908 \end{pmatrix}, b_{logreg,A} = -511$$

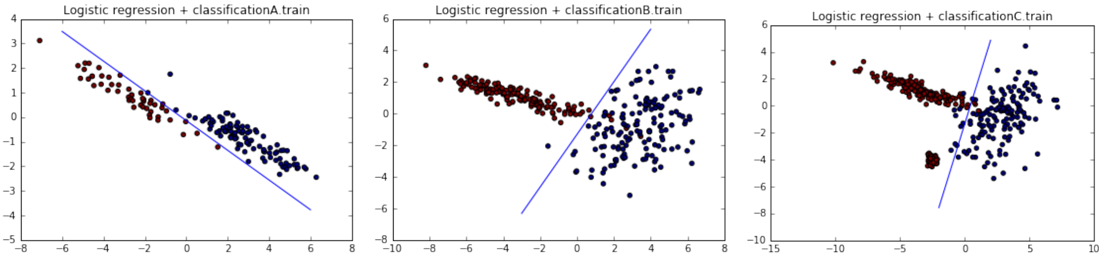
$$w_{logreg,B} = \begin{pmatrix} -1.71 \\ 1.02 \end{pmatrix}, b_{logreg,B} = 1.35$$

$$w_{logreg,C} = \begin{pmatrix} -2.20 \\ 0.709 \end{pmatrix}, b_{logreg,C} = 0.959$$

(b)

$$
\begin{aligned}
p(y=1|x) = 0.5 &\Leftrightarrow \sigma(w^T x + b) = 0.5 \\
&\Leftrightarrow exp(-(w^T x + b)) = 1 \\
&\Leftrightarrow w^T x + b = 0 \\
&\Leftrightarrow x_1 = \frac{-b - w_0 x_0}{w_1}
\end{aligned} \tag{2.12}
$$

Figure 2.3: Logistic regression + Training data. Left: A, Middle: B, Right: C

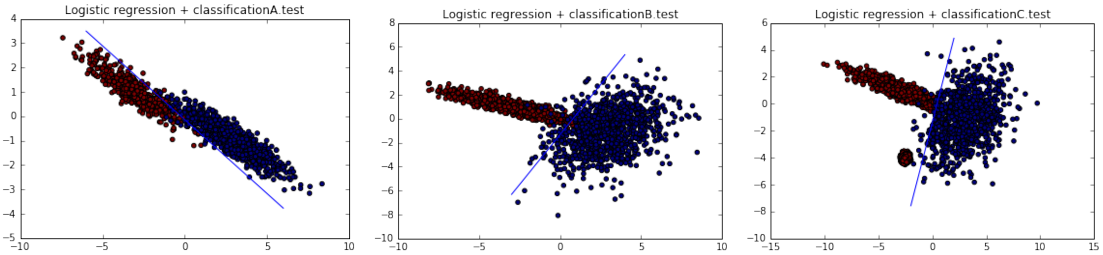


Logistic regression + classificationA.train

Logistic regression + classificationB.train

Logistic regression + classificationC.train

Figure 2.4: Logistic regression + Test data. Left: A, Middle: B, Right: C



Logistic regression + classificationA.test

Logistic regression + classificationB.test

Logistic regression + classificationC.test

3. Linear regression

(a)

$$w_{linreg,A} = \begin{pmatrix} -0.264 \\ -0.373 \end{pmatrix}, b_{linreg,A} = 0.492$$

$$w_{linreg,B} = \begin{pmatrix} -0.104 \\ 0.0518 \end{pmatrix}, b_{linreg,B} = 0.500$$

$$w_{linreg,C} = \begin{pmatrix} -0.128 \\ -0.0170 \end{pmatrix}, b_{linreg,C} = 0.508$$

(b)

$$\begin{aligned} p(y=1|x) = 0.5 &\Leftrightarrow \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(1-w^T x - b)^2}{2\sigma^2}) = \frac{1}{2} \\ &\Leftrightarrow (1-w^T x - b)^2 = 2\sigma^2 log\sqrt{\frac{2}{\pi\sigma^2}} \\ &\Leftrightarrow x_1 = \frac{1 - b - \sqrt{2\sigma^2 log\sqrt{\frac{2}{\pi\sigma^2}}} - w_0 x_0}{w_1} \end{aligned} \tag{2.13}$$

4.

(a)

Generative model:

Figure 2.5: Linear regression + Training data. Left: A, Middle: B, Right: C

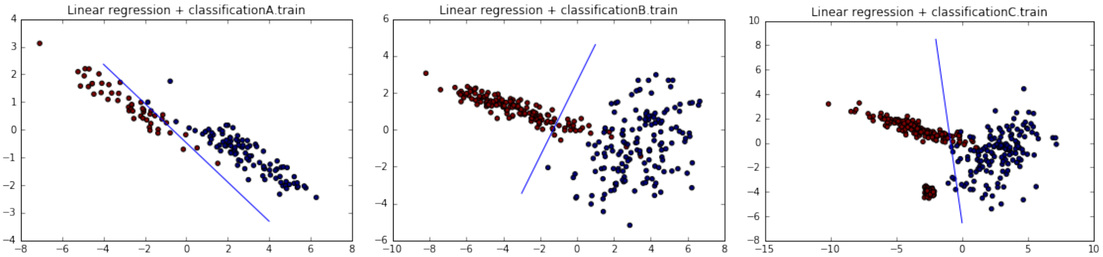
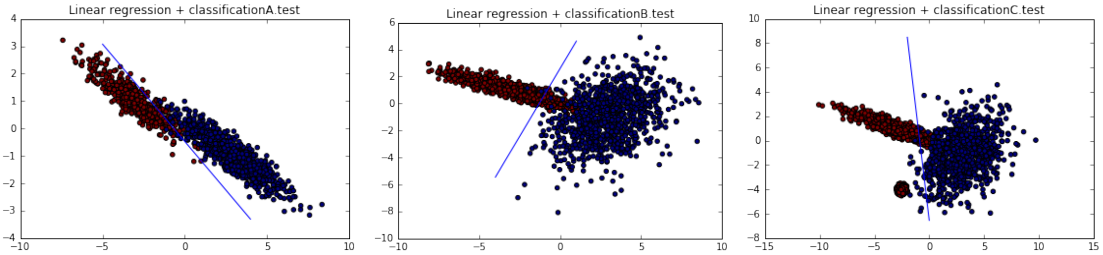


Figure 2.6: Linear regression + Test data. Left: A, Middle: B, Right: C



$$
\begin{aligned}
p(y=1|x) > 0.5 &\Leftrightarrow exp(-(\alpha+\beta x)) < 1 \\
&\Leftrightarrow \alpha+\beta x > 0
\end{aligned}
\tag{2.14}
$$

Logistic regression:

$$
\begin{aligned}
p(y=1|x) > 0.5 &\Leftrightarrow \sigma(w^T x + b) > 0.5 \\
&\Leftrightarrow exp(-(w^T x + b)) < 1 \\
&\Leftrightarrow exp(w^T x + b) > 1 \\
&\Leftrightarrow w^T x + b > 0
\end{aligned}
\tag{2.15}
$$

Linear regression:

$$
p(y=1|x) > 0.5 \Leftrightarrow \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(1-w^T x - b)^2}{2\sigma^2}) > \frac{1}{2}
\tag{2.16}
$$

(b)
It makes sense that the misclassification error be smaller on the training data than on the test data, since the model is trained based on the training data. However, this is not always the case. In Case C, we observe that the error on the training data is greater than that on the test data. This is probably due to a small group of points on the left-bottom side.

If $p(x|y)$ is a multivariate Gaussian with shared Σ, then $p(y|x)$ necessarily follows a logistic function. However, if $p(x|y)$ is a logistic function, $p(y|x)$ does not necessarily follow a multivariate Gaussian distribution. This means that the LDA model makes stronger assumptions

Table 2.1: Misclassification error among different models

data \model	LDA	LogReg	LinReg	N
A.train	0.0133	0.0	0.04	150
A.test	0.02	0.0333	0.0473	1500
B.train	0.03	0.02	0.08	300
B.test	0.0415	0.043	0.0965	2000
C.train	0.055	0.04	0.0725	400
C.test	0.0423	0.0227	0.061	3000

than the logistic regression model. When these assumptions are correct, the LDA model performs better than logistic regression. This is the case of A (the two groups of points roughly share the same Σ), but not of B or C.

In contrast, logistic regression makes weaker assumptions and is more robust to deviations from modeling assumptions. Also, when the data set is large, logistic regression performs better. This can be observed by comparing B and C.

In general, linear regression performs worse than the other two models.

Remark: In Case A, the training data is separable. When we use the Newton method to approximate the minimum of gradient l, the Hessian becomes non invertible at some point and we have to stop before that happens.

5. QDA model
We finally relax the assumption that the covariance matrices for the two classes are the same.

$$p(y) = \pi^y(1-\pi)^{1-y} \tag{2.17}$$

$$p(x|y=0) = \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)) \tag{2.18}$$

$$p(x|y=1) = \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)) \tag{2.19}$$

$$
\begin{aligned}
l(\pi,\mu_0,\mu_1,\Sigma_0,\Sigma_1) &= log\prod_{i=1}^{N} p(x_i,y_i;\pi,\mu_0,\mu_1,\Sigma) \\
&= log\prod_{i=1}^{N} p(x_i|y_i;\mu_0,\mu_1,\Sigma_0,\Sigma_1)p(y_i;\pi) \\
&= \sum_{i=1}^{N}(log\frac{1}{2\pi\sqrt{det\Sigma_{y_i}}}exp(-\frac{1}{2}(x_i-\mu_{y_i})^T\Sigma_{y_i}(x_i-\mu_{y_i})) \\
&\quad + y_i log\pi + (1-y_i)log(1-\pi)) \\
&= -Nlog2\pi - \frac{1}{2}\sum_{i=1}^{N} log(det\Sigma_{y_i}) - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu_{y_i})^T\Sigma_{y_i}^{-1}(x_i-\mu_{y_i}) \\
&\quad + \sum_{i=1}^{N}(y_i log\pi + (1-y_i)log(1-\pi))
\end{aligned}
\tag{2.20}
$$

$$
\frac{\partial l}{\partial \mu_0} = \sum_{i=1}^{N}\mathbb{1}\{y_i=0\}\Sigma_0^{-1}(\mu_0-x_i) = 0 \Rightarrow \widehat{\mu}_0 = \frac{\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}x_i}{\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}} \tag{2.21}
$$

Similarly,

$$
\widehat{\mu}_1 = \frac{\sum_{i=1}^{N}\mathbb{1}\{y_i=1\}x_i}{\sum_{i=1}^{N}\mathbb{1}\{y_i=1\}} \tag{2.22}
$$

$$
\frac{\partial l}{\partial \pi} = \sum_{i=1}^{N}\frac{y_i}{\pi} - \frac{1-y_i}{1-\pi} = 0 \Rightarrow \widehat{\pi} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{2.23}
$$

To compute the gradient of l with respect to Σ, we follow the lecture notes 1.4.7. By replacing the scalar $(x_i-\mu_{y_i})^T\Sigma^{-1}(x_i-\mu_{y_i})$ with $Trace((x_i-\mu_{y_i})^T\Sigma^{-1}(x_i-\mu_{y_i}))$ and by computing $\bigtriangledown log(detA) = A^{-1}$, we obtain

$$
\bigtriangledown_{\Sigma_0^{-1}}(l) = -\frac{1}{2}\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}\Sigma_0 + \frac{1}{2}\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}(x_i-\mu_0)(x_i-\mu_0)^T = 0 \tag{2.24}
$$

We obtain

$$
\widehat{\Sigma_0} = \frac{\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}(x_i-\mu_0)(x_i-\mu_0)^T}{\sum_{i=1}^{N}\mathbb{1}\{y_i=0\}} \tag{2.25}
$$

Similarly,

$$
\widehat{\Sigma_1} = \frac{\sum_{i=1}^{N}\mathbb{1}\{y_i=1\}(x_i-\mu_1)(x_i-\mu_1)^T}{\sum_{i=1}^{N}\mathbb{1}\{y_i=1\}} \tag{2.26}
$$

(a)

$\widehat{\mu_{0,A}} = \begin{pmatrix} 2.90 \\ -0.894 \end{pmatrix}, \widehat{\mu_{1,A}} = \begin{pmatrix} -2.69 \\ 0.866 \end{pmatrix}, \widehat{\pi_A} = 0.333, \widehat{\Sigma_{0,A}} = \begin{pmatrix} 2.31 & -1.05 \\ -1.05 & 0.576 \end{pmatrix}, \widehat{\Sigma_{1,A}} = \begin{pmatrix} 2.70 & -1.30 \\ -1.30 & 0.690 \end{pmatrix}$

$$\widehat{\mu_{0,B}} = \begin{pmatrix} 3.34 \\ -0.835 \end{pmatrix}, \widehat{\mu_{1,B}} = \begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix}, \widehat{\pi_B} = 0.5, \widehat{\Sigma_{0,B}} = \begin{pmatrix} 2.54 & 1.06 \\ 1.06 & 2.96 \end{pmatrix}, \widehat{\Sigma_{1,B}} = \begin{pmatrix} 4.15 & -1.33 \\ -1.33 & 0.516 \end{pmatrix}$$

$$\widehat{\mu_{0,C}} = \begin{pmatrix} 2.79 \\ -0.838 \end{pmatrix}, \widehat{\mu_{1,C}} = \begin{pmatrix} -2.94 \\ -0.958 \end{pmatrix}, \widehat{\pi_C} = 0.625, \widehat{\Sigma_{0,C}} = \begin{pmatrix} 2.90 & 1.25 \\ 1.25 & 2.92 \end{pmatrix}, \widehat{\Sigma_{1,C}} = \begin{pmatrix} 2.87 & -1.76 \\ -1.76 & 6.56 \end{pmatrix}$$

(b)

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \\ &= \frac{\mathcal{N}(x;\mu_1,\Sigma_1)\pi}{\mathcal{N}(x;\mu_0,\Sigma_0)(1-\pi)+\mathcal{N}(x;\mu_1,\Sigma_1)\pi} \\ &= \frac{\pi exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{(1-\pi)exp(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0))+\pi exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))} \\ &= \frac{1}{1+\frac{1-\pi}{\pi}exp((-\frac{1}{2})(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))} \end{aligned} \tag{2.27}$$

$$\begin{aligned} p(y=1|x)=0.5 \Leftrightarrow \frac{1-\pi}{\pi}exp((-\frac{1}{2})(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)) = 1 \\ \Leftrightarrow (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) = 2log\frac{\pi}{1-\pi} \end{aligned} \tag{2.28}$$

Figure 2.7: QDA + Training data. Left: A, Middle: B, Right: C

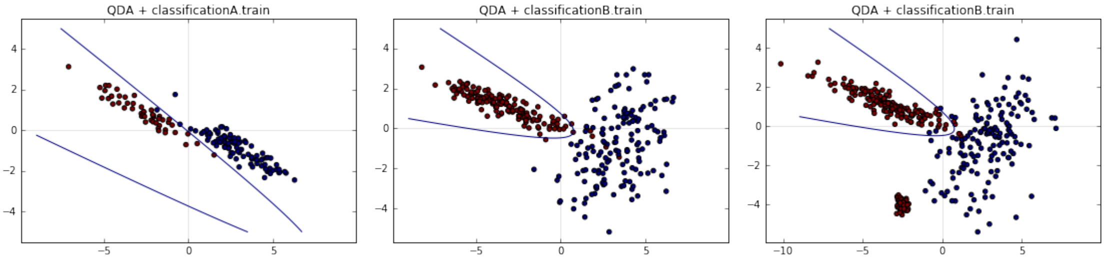
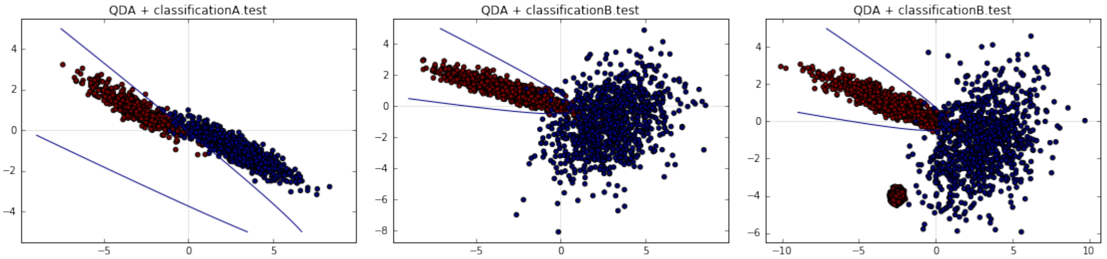


Figure 2.8: QDA + Test data. Left: A, Middle: B, Right: C



(c)

$$
\begin{aligned}
p(y=1|x) > 0.5 &\Leftrightarrow \frac{1-\pi}{\pi} exp((-\frac{1}{2})(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)+\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)) < 1 \\
&\Leftrightarrow (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)-(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) < 2log\frac{\pi}{1-\pi}
\end{aligned}
\tag{2.29}
$$

Table 2.2: Misclassification error among different models

data \model	LDA	LogReg	LinReg	QDA	N
A.train	0.0133	0.0	0.04	0.00667	150
A.test	0.02	0.0333	0.0473	0.02	1500
B.train	0.03	0.02	0.08	0.0233	300
B.test	0.0415	0.043	0.0965	0.04	2000
C.train	0.055	0.04	0.0725	0.0525	400
C.test	0.0423	0.0227	0.061	0.0575	3000

(d)
For the QDA model, the misclassification error is smaller on the training data than on the test data as previously. It performs slightly better than the LDA model as expected, since it makes weaker assumptions. It is comparable to the logistic regression in Case A and B but not in Case C, since the red points clearly do not follow a Gaussian distribution.