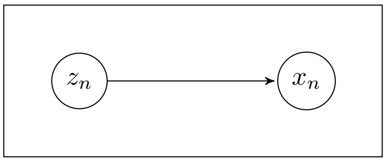
Master MVA
Probabilistic Graphical Models
DM 1 - 31/10/2016
Youssef Achari Berrada

1 Learning in discrete graphical models

In this example, we have a simple model with two nodes, represented by a plate:



z and x are discrete variables taking respectively M and K different values with $p(z_n = m) = \pi_m$ and $p(x_n = k|z_n = m) = \theta_{km}$. First, we observe that $\sum_{m=1}^{M} \pi_m = 1$ and $\sum_{k=1}^{K} \theta_{mk} = 1$ for each m.Then we have:

$$\begin{aligned} p_{\theta,\pi}(z_n, x_n) &= p_\pi(z_n).p_\theta(x_n|z_n) \\ &= \prod_{m=1}^{M} \left(\pi_m^{z_n^m} \prod_{k=1}^{K} \left(\theta_{mk}^{x_n^k} \right) \right) \end{aligned}$$

So the probability of the observation x:

$$\begin{aligned} p_{\theta,\pi}(z, x) &= \prod_{n=1}^{N} \prod_{m=1}^{M} \left(\pi_m^{z_n^m} \prod_{k=1}^{K} \left(\theta_{mk}^{x_n^k} \right) \right) \\ &= \prod_{m=1}^{M} \left[\prod_{n=1}^{N} (\pi_m^{z_n^m}) \prod_{k=1}^{K} \left(\prod_{n=1}^{N} (\theta_{mk}^{x_n^k}) \right) \right] \\ &= \prod_{m=1}^{M} \left[\pi_m^{\sum_{n=1}^{N} z_n^m} \prod_{k=1}^{K} \left(\theta_{mk}^{\sum_{n=1}^{N} x_n^k} \right) \right] \end{aligned}$$

To calculate the maximum likelihood estimates, we take the logarithm of the last equation to obtain the log likelihood:

$$\begin{aligned} l_{\pi,\theta}(z, x) &= \sum_{m=1}^{M} \left[\left(\sum_{n=1}^{N} z_n^m \right) \log(\pi_m) + \sum_{k=1}^{K} \left(\sum_{n=1}^{N} x_n^k \right) \log(\theta_{mk}) \right] \\ &= \sum_{m=1}^{M} \left(\sum_{n=1}^{N} z_n^m \right) \log(\pi_m) + \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\sum_{n=1}^{N} x_n^k \right) \log(\theta_{mk}) \end{aligned}$$

The maximization over (π, θ) of the log likelyhood is a constrained optimization problem for which we use the Lagrange multipliers with $M + 1$ constraints:

$$\tilde{l}_{\pi,\theta}(z, x) = \sum_{m=1}^{M} \left(\sum_{n=1}^{N} z_n^m \right) \log(\pi_m) + \sum_{m=1}^{M} \sum_{k=1}^{K} \left(\sum_{n=1}^{N} x_n^k \right) \log(\theta_{mk}) + \lambda(1 - \sum_{m=1}^{M} \pi_m) + \sum_{m=1}^{M} \gamma_m(1 - \sum_{k=1}^{K} \theta_{mk})$$

The derivative over π_m:

$$\frac{\partial \tilde{l}_{\pi,\theta}(z,x)}{\partial \pi_m} = \frac{\sum_{n=1}^{N} z_n^m}{\pi_m} - \lambda$$

The derivative over θ_{mk}:

$$\frac{\partial \tilde{l}_{\pi,\theta}(z,x)}{\partial \theta_{mk}} = \frac{\sum_{n=1}^{N} x_n^k}{\theta_{mk}} - \gamma_m$$

By setting these partial derivatives equal to zero, we obtain:

$$\sum_{n=1}^{N} z_n^m = \lambda \pi_m \qquad \underbrace{\Longrightarrow}_{\text{After summing over m}} \qquad \lambda = N$$

$$\sum_{n=1}^{N} x_n^k = \gamma_m \theta_{mk} \qquad \underbrace{\Longrightarrow}_{\text{After summing over k}} \qquad \gamma_m = N$$

Finally, by substitution, we obtain:

$$\hat{\pi}_{m,ML} = \frac{1}{N} \sum_{n=1}^{N} z_n^m$$

$$\hat{\theta}_{mk.ML} = \frac{1}{N} \sum_{n=1}^{N} x_n^k$$

Noting that $\sum_{n=1}^{N} z_n^m$ is the count of the number of times that the mth value is observed in z and also $\sum_{n=1}^{N} x_n^k$ is the count of the number of times that the kth value is observed in x. We see that the maximum likelihood estimate is a sample proportion.

2 Linear classification

2.1 Generative model (LDA):

(a) First, we know that $y \sim \textit{Bernoulli}(\pi)$ and $x|y = i \sim \textit{Normal}(\mu_i, \Sigma)$. So we can express:

$$p(y=1) = \pi \quad , \quad p(y=0) = 1 - \pi \quad , \quad p_{\mu_i,\Sigma}(X = x|y = i) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x - \mu_i)\right)$$

For $i \in \{0, 1\}$, and if $X_1, ..., X_N \sim \textit{Normal}(\mu_i, \Sigma)$,iid., we have seen in the lecture 2 that:

$$\hat{\mu}_i = \frac{1}{N} \sum_{k=1}^{N} x_k$$

$$\hat{\Sigma}_i = \frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{\mu}_i)^T (x_k - \hat{\mu}_i)$$

In the experiment, the fisher assumption is considered by taken $\hat{\Sigma} = \dfrac{\hat{\Sigma}_1 + \hat{\Sigma}_0}{2}$.

(b)

$$\begin{aligned} p(y=1|X=x) &= \frac{p(X=x|y=1)p(y=1)}{p(X=x|y=1)p(y=1)+p(X=x|y=0)p(y=0)} \\ &= \frac{\pi\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\pi\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)+(1-\pi)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)} \\ &= \frac{1}{1+\frac{1-\pi}{\pi}\exp\left(-\frac{1}{2}[(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]\right)} \end{aligned}$$

We can proove that:

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

Then we conclude that:

$$p(y=1|X=x)=\frac{1}{1+\frac{1-\pi}{\pi}\exp\left(-(x-\frac{\mu_0+\mu_1}{2})^T\Sigma^{-1}(\mu_1-\mu_0)\right)}$$

(c) So we conclude that:

$$p(y=1|X=x)>0.5 \Leftrightarrow (x-\frac{\mu_0+\mu_1}{2})^T\Sigma^{-1}(\mu_1-\mu_0)>\log(\frac{1-\pi}{\pi})$$

which define a halfspace that depend on the value of π.

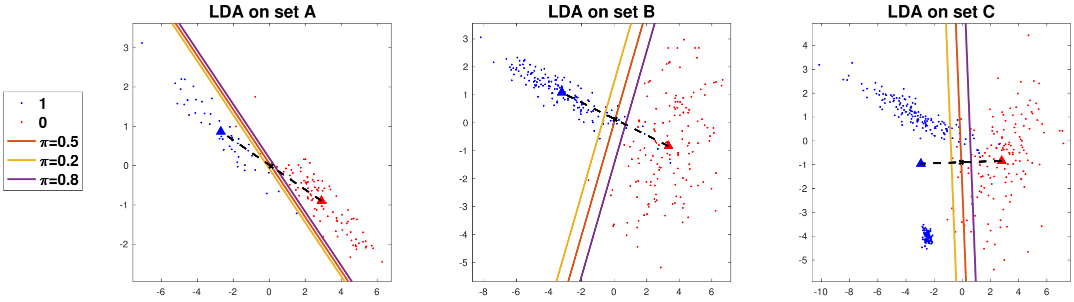


Figure 1: LDA on the train data with different value of π.

Table 1: Learnt parameters for LDA

	A	B	C
μ_0	$\begin{pmatrix} 2.8997 \\ -0.8939 \end{pmatrix}$	$\begin{pmatrix} 3.3407 \\ -0.8355 \end{pmatrix}$	$\begin{pmatrix} 2.7930 \\ -0.8384 \end{pmatrix}$
μ_1	$\begin{pmatrix} -2.6923 \\ 0.8660 \end{pmatrix}$	$\begin{pmatrix} -3.2167 \\ 1.0831 \end{pmatrix}$	$\begin{pmatrix} -2.9423 \\ -0.9578 \end{pmatrix}$
Σ	$\begin{pmatrix} 2.5468 & -1.1927 \\ -1.1927 & 0.6427 \end{pmatrix}$	$\begin{pmatrix} 3.3687 & -0.1361 \\ -0.1361 & 1.7497 \end{pmatrix}$	$\begin{pmatrix} 2.8996 & -0.2574 \\ -0.2574 & 4.7676 \end{pmatrix}$

2.2 Logistic Regression:

(a) $f(x) = w^T x + b$ with the parameters $w \in \mathbb{R}^2$ and $b \in \mathbb{R}$. In order to include the constant term, one can do the following transformation on f :

$$f(x) = \begin{pmatrix} b \\ w \end{pmatrix}^T \begin{pmatrix} 1 \\ x \end{pmatrix}$$

and then consider $\tilde{w} = \begin{pmatrix} b \\ w \end{pmatrix} \in \mathbb{R}^3$.

After learning the parameters w and b using the *Iterative Reweighted Linear Square* method.

Table 2: Learnt parameters by logistic regression.

Logistic Regression	Train A	Train B	Train C
w_1	-989.0780	-6.4128	-7.1788
w_2	-637.7548	1.6720	1.6198
b	-215.6995	1.3706	2.0556

(b)

$$p(y = 1|x) = 0.5 \Leftrightarrow \begin{pmatrix} b \\ w \end{pmatrix}^T \begin{pmatrix} 1 \\ x \end{pmatrix} = 0$$
$$\Leftrightarrow w^T x + b = 0$$

which define a line in $\mathbb{R}^2$.

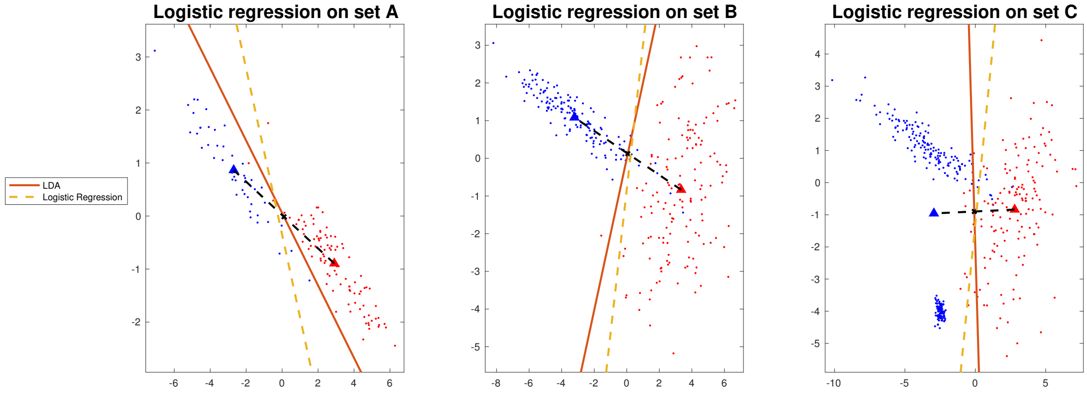


Figure 2: Logistic regression on the training data versus LDA.

2.3 Linear Regression:

Table 3: Learnt parameters by linear regression.

Linear Regression	Train A	Train B	Train C
w_1	-0.8118	-0.3920	-0.4161
w_2	-0.4266	0.0846	-0.0388
b	0.3333	0.5000	0.6250

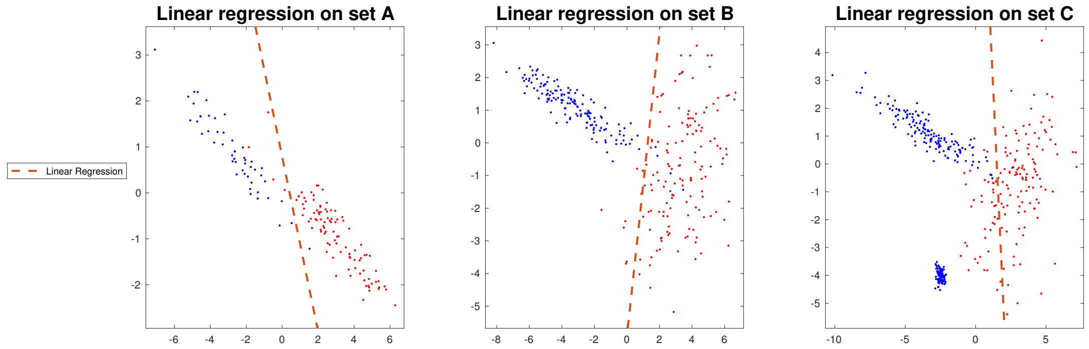


Figure 3: Linear regression on the training data.

2.4 Testing:

Table 4: Misclassification Error

		A	B	C
LDA	Train	**0.0133**	0.0300	0.0500
	Test	0.0240	**0.0415**	0.0387
Logisctic Regression	Train	0.0333	**0.0267**	**0.0400**
	Test	**0.0180**	0.0440	**0.0257**
Linear Regression	Test	0.0400	0.0400	0.0975
	Train	0.0360	0.0670	0.0780

As we see in the table above, the logistic model is the best one, in particular on the set C where the model performs far better than the other two models.
The LDA perform well on the set A because indeed, the fisher assumption is verified on that particular set.
The LDA and Logistic model yield to very similar results on the set B. The Linear model performs bad on the models where there is outliers like on the set C, because it does not take into account the distribution of the data.

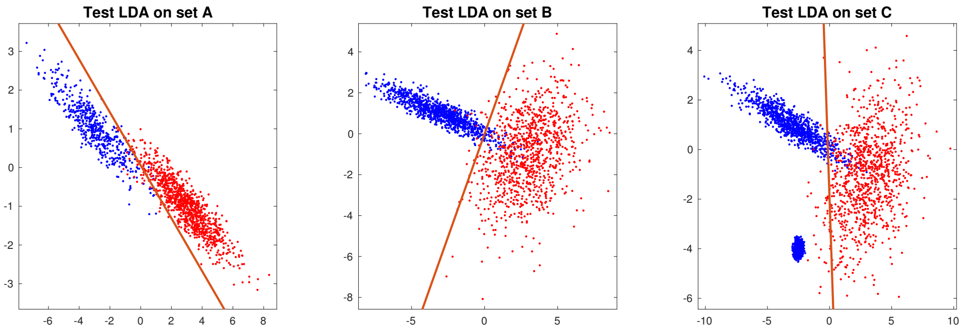


Figure 4: Testing LDA.

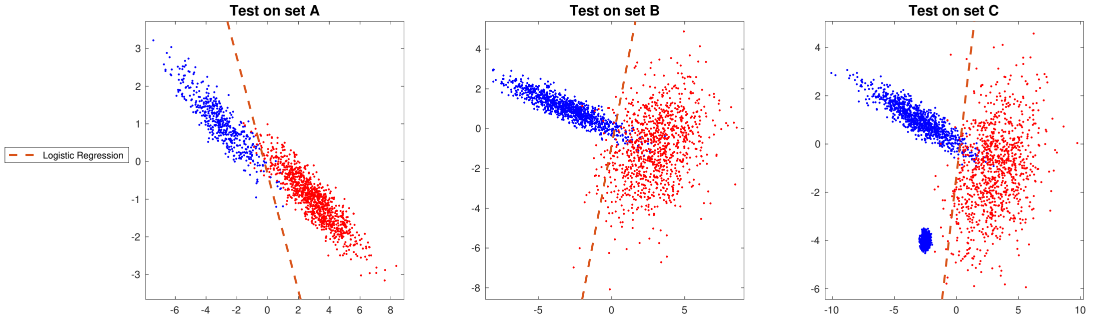


Figure 5: Testing logistic regression.

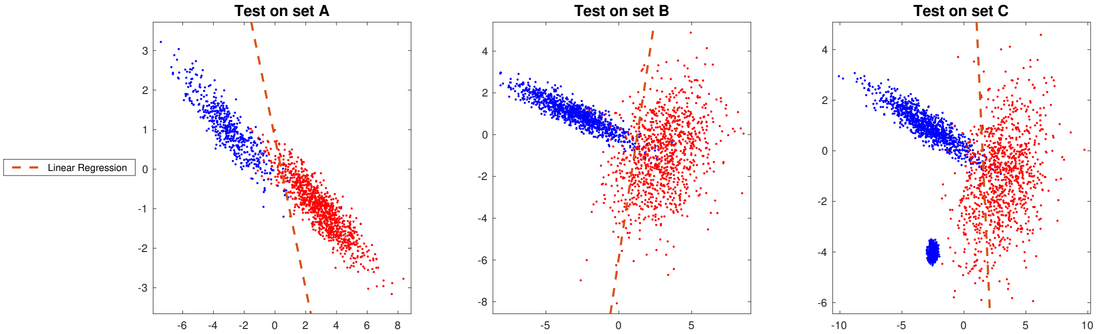


Figure 6: Testing linear regression.

2.5 QDA model:

(a) The parameters μ_0 and μ_1 are the same as in the LDA model. See table 1.

Table 5: Learnt parameters for QDA

	A	B	C
Σ_0	$\begin{pmatrix} 2.5468 & -1.1927 \\ -1.1927 & 0.6427 \end{pmatrix}$	$\begin{pmatrix} 3.3687 & -0.1361 \\ -0.1361 & 1.7497 \end{pmatrix}$	$\begin{pmatrix} 2.8996 & -0.2574 \\ -0.2574 & 4.7676 \end{pmatrix}$
Σ_1	$\begin{pmatrix} 2.5468 & -1.1927 \\ -1.1927 & 0.6427 \end{pmatrix}$	$\begin{pmatrix} 3.3687 & -0.1361 \\ -0.1361 & 1.7497 \end{pmatrix}$	$\begin{pmatrix} 2.8996 & -0.2574 \\ -0.2574 & 4.7676 \end{pmatrix}$

(b)

$$\begin{aligned} p(y=1|X=x) &= \frac{p(X=x|y=1)p(y=1)}{p(X=x|y=1)p(y=1)+p(X=x|y=0)p(y=0)} \\ &= \frac{\pi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)}{\pi \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)+(1-\pi)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)} \\ &= \frac{1}{1+\frac{1-\pi}{\pi}\exp\left(-\frac{1}{2}[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)]\right)} \end{aligned}$$

$$p(y=1|X=x)=0.5 \Leftrightarrow (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)=2\log(\frac{1-\pi}{\pi})$$

which define a conic.

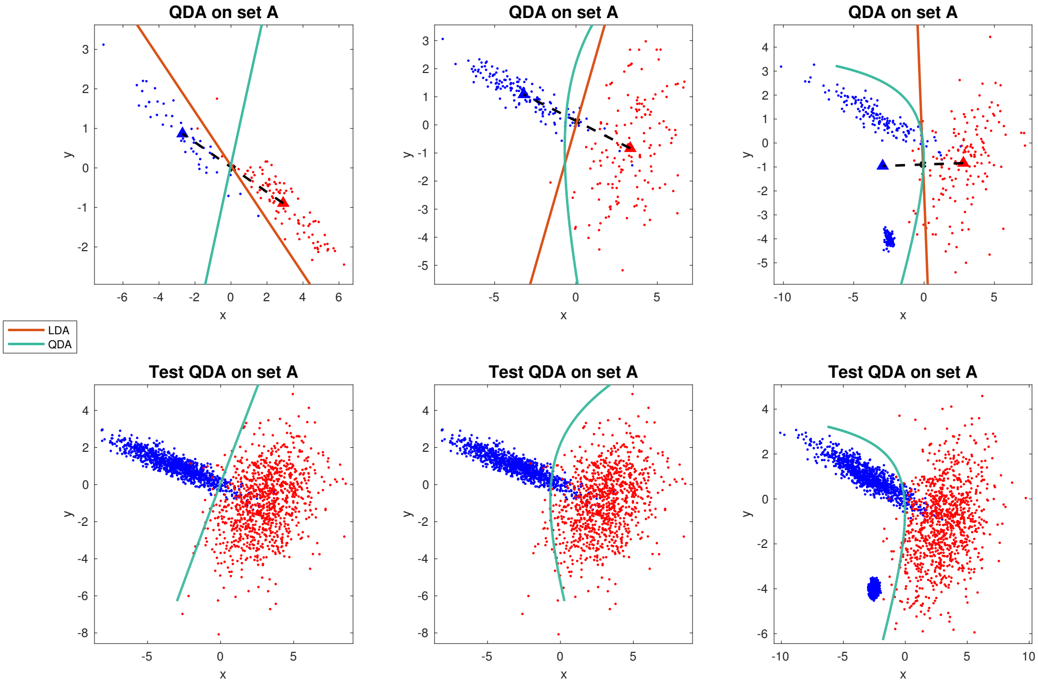


Figure 7: QDA vs LDA.

(c) Here is the misclassification error on the QDA training and test sets.

Table 6: QDA Misclassification Error

		A	B	C
QDA	Train	0.0133	0.0233	0.0525
	Test	0.0213	0.0235	0.0370

(d) If we compare the misclassification errors between the QDA and LDA models, they are very similar on set A, because the QDA and the LDA are equivalent on set A as $\Sigma_1 \sim \Sigma_2$.
The QDA model is better than the LDA model for the set B, because the QDA model now take into account the non similarity between Σ_1 and Σ_2. We can notice a high improvement on the test set B by the QDA model indeed.
We could notice that the QDA classification has less error that the LDA classification on the training set C and in opposite the QDA classification performs lower than the LDA one on the test set C. This is basically what happens when we have overfitting. And it is normal because the QDA model involve quadratic terms on x and y but the LDA involve only linear terms on x and y. But in this experiment, the QDA model is better that the LDA one on both the training and the test set C.