Probabilistic Graphical Models : Homework 1

Matthis Maillard

October 2018

1 Learning in discrete graphical model

By maximizing $\log(p_{\pi,\theta}(x,y))$ we find estimators for π and θ:

$$\hat{\pi}_m = \frac{\sum_{i=1}^{M} z_i^m}{n} \text{ where } z_i^m = 1 \text{ if } z_i = m, 0 \text{ otherwise}$$

$$\hat{\theta}_{m,k} = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\sum_{i=1}^{n}\sum_{k=1}^{K} x_i^k z_i^m} \text{where } x_i^k = 1 \text{ if } x_i = k, 0 \text{ otherwise}$$

2 Linear Classification

2.1 LDA

We maximize the log-likelihood of $p_{\pi,\mu_1,\mu_2,\sigma}(x,y)$ to find the estimators of the parameters. We obtain :

$$\hat{\pi} = \frac{1}{N}\sum_{n=1}^{N} y_n$$

$$\hat{\mu_1} = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} y_n}$$

$$\hat{\mu_2} = \frac{\sum_{n=1}^{N} (1-y_n) x_n}{\sum_{n=1}^{N} (1-y_n)}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}[y_n(x_n-\mu_1)(x_n-\mu_1)^T + (1-y_n)(x_n-\mu_2)(x_n-\mu_2)^T]$$

2.2 QDA

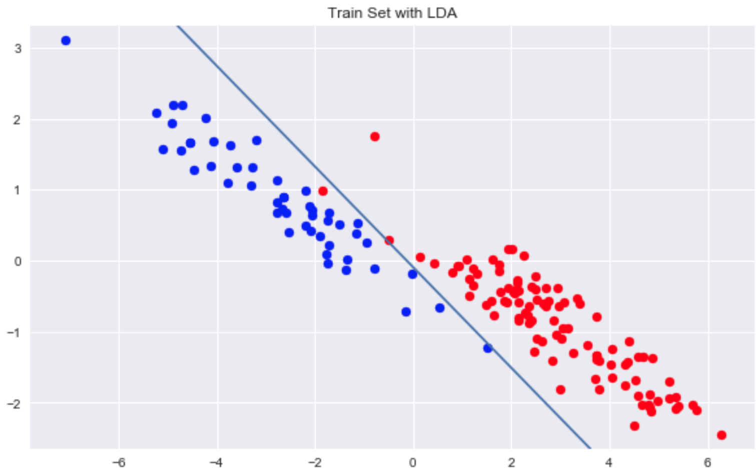
With the assumption that the covariance matrix of the classes are different, we get :

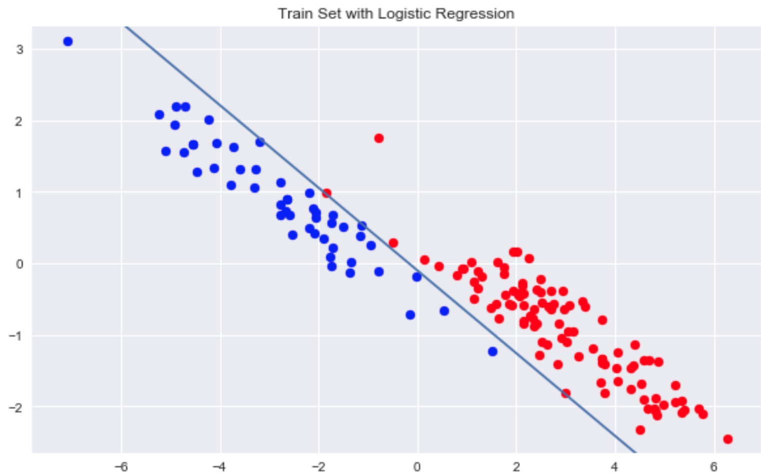
$$\hat{\Sigma}_1 = \frac{1}{N}\sum_{n=1}^{N} y_n(x_n-\mu_1)(x_n-\mu_1)^T$$

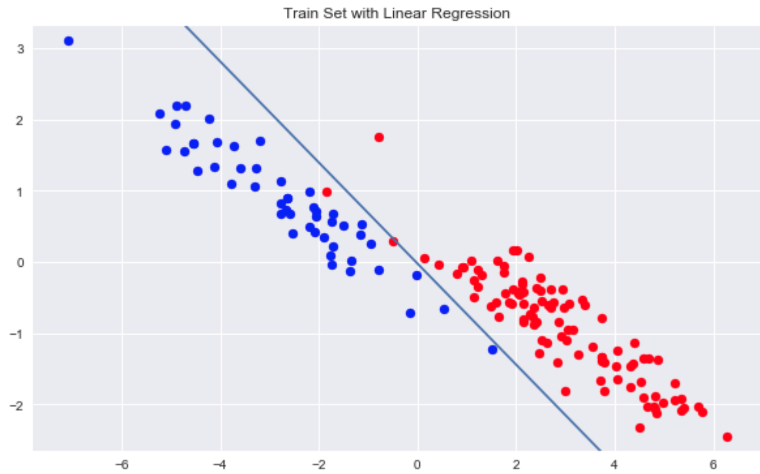
$$\hat{\Sigma}_2 = \frac{1}{N}\sum_{n=1}^{N} (1-y_n)(x_n-\mu_2)(x_n-\mu_2)^T]$$

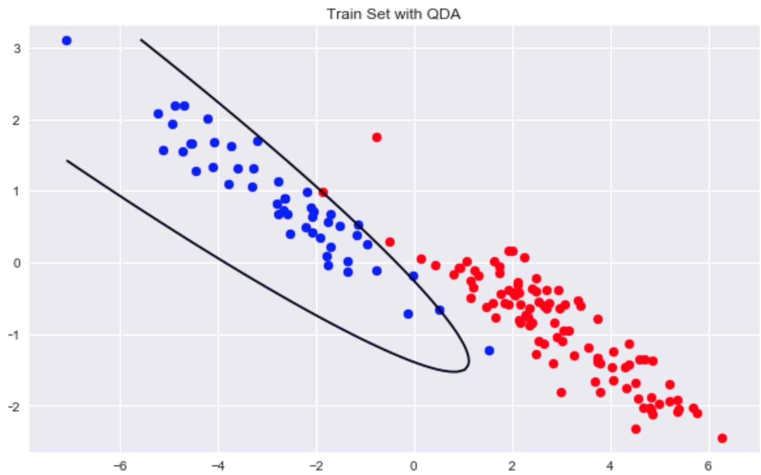
All proofs are at the end of the document.

Dataset A :





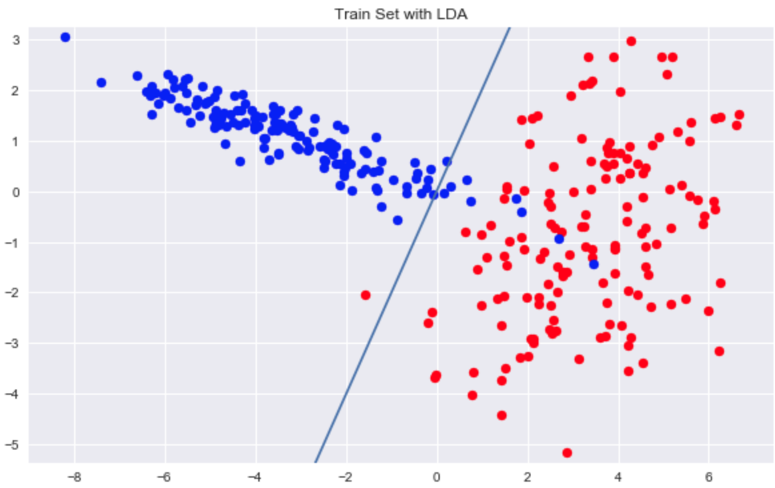


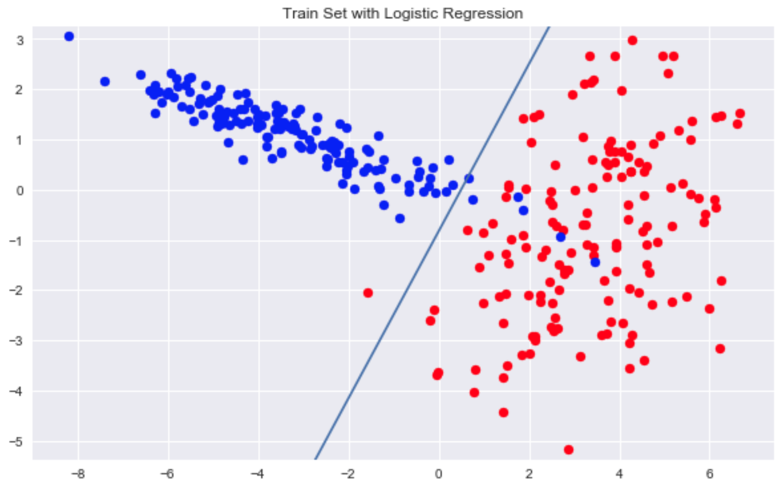


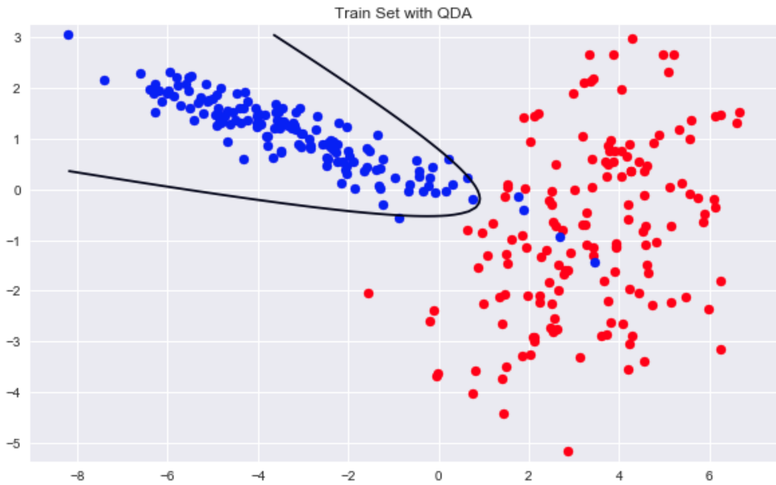
Classifier	Train errors	Test errors
LDA	0.013333	0.020000
Logistic_Regression	0.000000	0.034667
Linear_regression	0.013333	0.020667
QDA	0.020000	0.030667

With this dataset, the logistic regression has perfectly separated the training data but its prediction score on the test set is the lowest. This is because the logistic regression model aims to separate the data with a hyperplane but does not consider the noise of data, hence, it is overfitting it. The assumption that the classes have similar covariance matrix seems correct since the LDA model has approximately the same error in the training set and the test set. Plus, the model has the best prediction score. As both classes have the same distribution and are well separated, the linear regression is a good solution and does almost as well as LDA. The QDA classifier has the worst training error and is just a bit better than Logistic Regression on the test set. Indeed, the data seems to be separable by a hyperplane, so doing it with a quadratic form is not appropriate. There is almost no difference between linear regression and LDA, if we look at the separating hyperplanes they also look the same.

Dataset B :



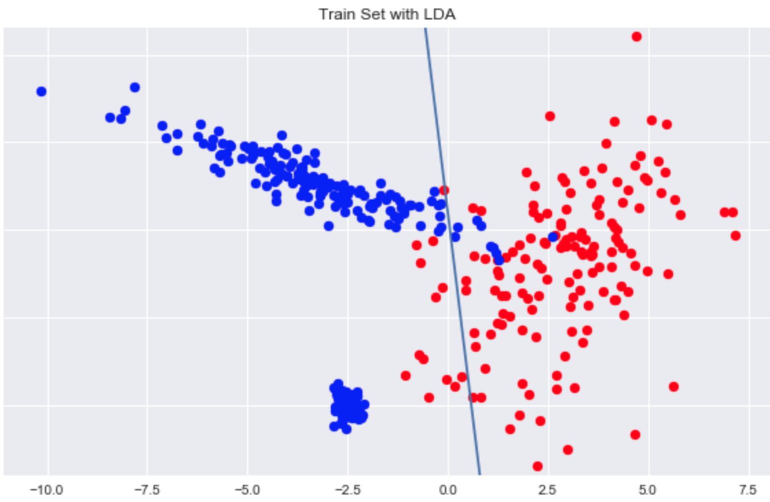


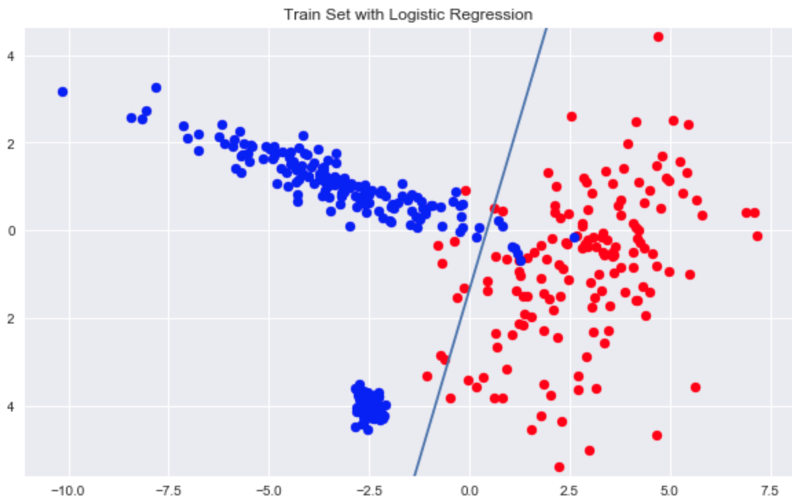


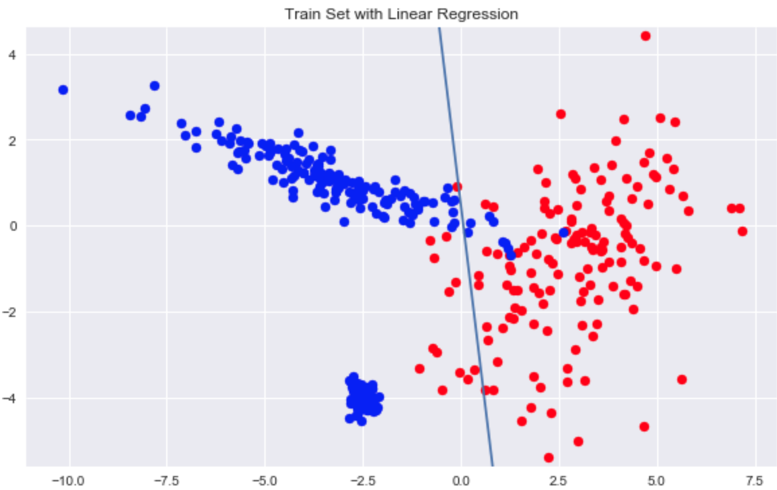
Classifier	Train errors	Test errors
LDA	0.030000	0.0415
Logistic_Regression	0.020000	0.0430
Linear_regression	0.030000	0.0415
QDA	0.016667	0.0215

With the dataset B, logistic regression is still doing the best training score and the worst testing score. LDA, logistic regression and linear regression have about the same testing score. This is because the classes are not separable by a hyperplane. Some points of the blue class are mixed with the red class and it does not look like noise. For the LDA model, the assumption that classes have the same covariance matrix is clearly false, the red class is more spread than the blue one. As the covariance matrix are different and that it has a quadratic form, the QDA model is better than the others. The model can correctly wrap data. Again, linear regression and LDA give the same results, here the hyperplanes exactly look the same.

Dataset C :







Classifier	Train errors	Test errors
LDA	0.055	0.042333
Logistic_Regression	0.040	0.022667
Linear_regression	0.055	0.042333
QDA	0.050	0.041000

In this case, logistic regression beats all the other classifier in both the training set and the test set. Though the dataset look like dataset B, the fact that the blue class is separated in two groups makes it difficult for QDA to fit the data as in B. Hence, the conic wraps the red dataset but the result is not as good as in B and, here, the logistic regression does better. LDA and Linear regression have the same errors in both datasets and have the same separating lines.

Matthis

Maillard

Probabilistic Graphical Models

Homework 1

I Learning in discrete graphical models

Let $(x_i, z_i)_{i=1,n}$ be an i.i.d sample of observations.

We have that:

$$p_{\theta,\pi}(x_1 = x_1, \dots, x_n = x_n, z_1 = z_1, \dots, z_n = z_n) = \prod_{i=1}^{n} p_{\theta,\pi}(x_i, z_i) = \prod_{i=1}^{n} p_{\pi}(z_i)\, p_{\theta}(x_i | z_i)$$

To find the maximum likelihood estimator, we compute the log likelihood.

$$\log p_{\theta,\pi}(x, z) = \sum_{i=1}^{n} \log(p_{\pi}(z_i)) + \log(p_{\theta}(x_i | z_i))$$

$$= \sum_{i=1}^{n} \log\left(\prod_{m=1}^{M} \pi_m^{z_i^m}\right) + \log\left(\prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{m,k}^{x_i^k z_i^m}\right), \text{ with } z_i^m = \begin{cases} 1 & \text{if } z_i = m \\ 0 & \text{otherwise} \end{cases}$$

$$= \sum_{i=1}^{n} \left[\sum_{m=1}^{M} z_i^m \log \pi_m + \sum_{m=1}^{M} \sum_{k=1}^{K} x_i^k z_i^m \log \theta_{m,k}\right] \quad \text{and} \quad x_i^k = \begin{cases} 1 & \text{if } x_i = k \\ 0 & \text{otherwise} \end{cases}$$

As we are dealing with probabilities, we have the constraints that:

$$\sum_{m=1}^{M} \pi_m = 1 \quad \text{and} \quad \sum_{k=1}^{K} \theta_{m,k} = 1 \quad \text{for all } m \in \{1, M\}$$

To find $\underset{\theta,\pi}{\operatorname{argmax}} \log p_{\theta,\pi}(x, z)$, we compute the Lagrangian.

$$L(\pi, \theta, \lambda, \nu) = \sum_{i=1}^{n} \left[\sum_{m=1}^{M} z_i^m \log \pi_m + \sum_{m=1}^{M} \sum_{k=1}^{K} x_i^k z_i^m \log \theta_{m,k}\right] + \lambda\left(\sum_{m=1}^{M} \pi_m - 1\right) + \sum_{m=1}^{M} \nu_m \left(\sum_{k=1}^{K} \theta_{m,k} - 1\right)$$

We set the gradients to zero

$$\left(\nabla_{\pi} L(\pi, \theta, \lambda, \nu)\right)_m = \frac{\sum_{i=1}^{n} z_i^m}{\pi_m} - \lambda, \quad \frac{\sum_{i=1}^{n} z_i^m}{\pi_m} - \lambda = 0 \Rightarrow \pi_m = \frac{\sum_{i=1}^{n} z_i^m}{\lambda}$$

$$\left(\nabla_{\theta} L(\pi, \theta, \lambda, \nu)\right)_{m,k} = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\theta_{m,k}} - \nu \Rightarrow \theta_{m,k} = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\nu}$$

Then, $\sum_{m=1}^{M} \pi_m = 1 \iff \sum_{m=1}^{M} \frac{\sum_{i=1}^{n} z_i^m}{\lambda} = 1 \iff \frac{\sum_{i=1}^{n} 1}{\lambda} = 1 \iff \lambda = n$

plus, $\sum_{k=1}^{K} \theta_{m,k} = 1 \iff \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\nu} = 1 \iff \nu = \sum_{i=1}^{n} \sum_{k=1}^{K} x_i^k z_i^m$

So, the MLE gives us $\pi_m^* = \frac{1}{n} \sum_{i=1}^{n} z_i^m$ and $\theta_{m,k}^* = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\sum_{i=1}^{n} \sum_{k=1}^{K} x_i^k z_i^m}$

2. Linear classification

1. a LDA

The set of parameters is $\Theta = [0,1] \times \mathbb{R}^2 \times \mathbb{R}^2 \times S_2(\mathbb{R})^+$ with $\theta = (\pi, \mu_1, \mu_2, \Sigma)$

The joint probability of x and y is:

$$p_\theta(x,y) = p_\theta(y)\, p_\theta(x|y) \propto \left(\frac{\pi}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}\right)^y \left(\frac{1-\pi}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}\right)^{1-y}$$

Thus, $\log p_\theta(x,y) = C + \sum_{n=1}^{N} y_n \left(\log \pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_1)^T \Sigma^{-1}(x_n-\mu_1)\right) + (1-y_n)\left(\log(1-\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu_2)^T \Sigma^{-1}(x_n-\mu_2)\right)$

$C \in \mathbb{R}$,

To find its maximum, we set the gradients to zero.

$\nabla_\pi \log p_\theta(x,y) = \sum_{n=1}^{N} \frac{y_n}{\pi} - \frac{1-y_n}{1-\pi}$, $\nabla_\pi \log p_\theta(x,y) = 0 \iff \pi^* = \frac{1}{N}\sum_{n=1}^{N} y_n$

To find the gradient of $f: \mu \to (x-\mu)^T \Sigma^{-1}(x-\mu)$, we compute for $h \in \mathbb{R}^2$

$$\begin{aligned} f(\mu+h) - f(\mu) &= (x-\mu-h)^T \Sigma^{-1}(x-\mu-h) - (x-\mu)^T \Sigma^{-1}(x-\mu) \\ &= -2h^T \Sigma^{-1}(x-\mu) + h^T \Sigma^{-1} h \\ &= \langle h | -2\Sigma^{-1}(x-\mu)\rangle + o(\|h\|) \end{aligned}$$

So $\nabla_\mu f(\mu) = -2\Sigma^{-1}(x-\mu)$

Hence, $\nabla_{\mu_1} \log p_\theta(x,y) = \sum_{n=1}^{N} y_n(-2\Sigma^{-1}(x_n-\mu_1))$, $\nabla_{\mu_1} \log p_\theta(x,y) = 0 \implies \mu_1^* = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} y_n}$

With the same method, we find that $\mu_2^* = \frac{\sum_{n=1}^{N} (1-y_n) x_n}{\sum_{n=1}^{N} (1-y_n)}$

To find the $\nabla_{\Sigma^{-1}} \log p_\theta(x,y)$, we compute the gradient of $\log(\det(A))$ with $A \in S_n(\mathbb{R})$

Let $H \in S_n(\mathbb{R})$, $\log(\det(A+H)) = \log(\det(A^{1/2}A^{1/2} + H))$

$$= \log(\det(A^{\frac{1}{2}}(I + A^{-\frac{1}{2}}HA^{-\frac{1}{2}})A^{\frac{1}{2}}))$$

$$= \log(\det(A(I + A^{-\frac{1}{2}}HA^{-\frac{1}{2}})))$$

$A^{-\frac{1}{2}}HA^{-\frac{1}{2}}$ is symmetric, thus by applying the spectral theorem, we rewrite $A^{-\frac{1}{2}}HA^{-1/2}$ as $P^T D P$ with D a diagonal matrix and P an orthogonal matrix.

$$\log\det((A(I + A^{-\frac{1}{2}}HA^{-\frac{1}{2}}))) = \log\det A + \log\det(P^T(I + D)P)$$

$$= \log\det A + \sum_{i=1}^{n} \log(1+\lambda_i) \text{ with } (\lambda_i)_i \text{ the eigenvalues of } A^{-\frac{1}{2}}HA^{-\frac{1}{2}}$$

when $\|H\| \to 0$: $\sum_{i=1}^{n} \log(1+\lambda_i) = \sum_{i=1}^{n} \lambda_i + o(\|H\|)$

$$= \mathrm{Tr}(A^{-\frac{1}{2}}HA^{-\frac{1}{2}})$$

$$= \mathrm{Tr}(A^{-1}H)$$

Hence, the gradient of $\log\det(A)$ is A^{-1}.

Plus, the gradient of $\Sigma^{-1} \to (x-\mu)^T\Sigma^{-1}(x-\mu)$ with respect to Σ^{-1} is $(x-\mu)(x-\mu)^T$

To compute the gradient of $\log p_\theta(x,y)$ w.r.t Σ^{-1}, we rewrite $\log\det\Sigma$ as $-\log\det(\Sigma^{-1})$.

thus, $\nabla_{\Sigma^{-1}} \log p_\theta(x,y) = \sum_{n=1}^{N}\left[y_n\left(\frac{1}{2}\Sigma - \frac{1}{2}(x_n-\mu_1)(x_n-\mu_1)^T\right) + (1-y_n)\left(\frac{1}{2}\Sigma - \frac{1}{2}(x_n-\mu_2)(x_n-\mu_2)^T\right)\right]$

By setting it to zero, we obtain $\Sigma^* = \frac{1}{N}\sum_{n=1}^{N}\left[y_n(x_n-\mu_1)(x_n-\mu_1)^T + (1-y_n)(x_n-\mu_2)(x_n-\mu_2)^T\right]$

$$p(y=1|x) = \frac{p(x|y=1)\,p(y=1)}{p(x|y=1)\,p(y=1) + p(x|y=0)\,p(y=0)} = \frac{1}{1 + \frac{p(x|y=0)\,p(y=0)}{p(x|y=1)\,p(y=1)}}$$

$$= \frac{1}{1 + \frac{1-\pi}{\pi} e^{-(w^Tx+b)}}$$

where $w \in \mathbb{R}^2$, $b \in \mathbb{R}^2$ depend on Σ, μ_1, μ_2.

For the form of logistic regression, $p(y=1|x) = \dfrac{1}{1+e^{-(w'^T x + b')}}$

where $w' \in \mathbb{R}^2$ and $b' \in \mathbb{R}^2$

The LDA and the logistic regression are similar, the difference is that in the logistic regression the classes are assumed to have the same probability.

5.a QDA $\quad \Theta = [0,1] \times \mathbb{R}^2 \times \mathbb{R}^2 \times S_2(\mathbb{R}) \times S_2(\mathbb{R}) \quad \theta = (\pi, \mu_1, \mu_2, \Sigma_1, \Sigma_2)$

By resoning as in 1.a for LDA, we find that:

$$\log p_\theta(x,y) \propto \mp C + \sum_{n=1}^{N} y_n \left(\log \pi - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x_n-\mu_1)^T \Sigma_1^{-1}(x_n-\mu_1)\right) + (1-y_n)\left(\log(1-\pi) - \frac{1}{2}\log|\Sigma_2^{-1}| - \frac{1}{2}(x_n-\mu_2)^T \Sigma_2^{-1}(x_n-\mu_2)\right)$$

and $\quad \pi^* = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N} y_n$

$$\mu_1^* = \frac{\sum_{n=1}^{N} y_n x_n}{\sum_{n=1}^{N} y_n}, \quad \mu_2^* = \frac{\sum_{n=1}^{N}(1-y_n)x_n}{\sum_{n=1}^{N}(1-y_n)}$$

By setting the gradient w.r.t Σ_1^{-1} to zero, we get:

$$\frac{1}{2}\sum_{n=1}^{N} y_n \left(\Sigma_1 - (x_n-\mu_1)(x_n-\mu_1)^T\right) = 0 \iff \Sigma_1^* = \frac{1}{N}\sum_{n=1}^{N} y_n (x_n-\mu_1)(x_n-\mu_1)^T$$

Also, for Σ_2: $\quad \Sigma_2^* = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(1-y_n)(x_n-\mu_2)(x_n-\mu_2)^T$