Assignment Two

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

October 8, 2017

1 Fisher LDA

Given the class variable, the data are assumed to be Gaussians with different means for different classes but with the same covariance matrix.

1.1 Derive the form of the maximum likelihood estimator for this model

Given $Y \sim Bernoulli(\pi), X \mid Y = j \sim \mathcal{N}(\mu_j, \Sigma)$. We first obtain the log likelihood $l(\theta \mid D)$, where D are the data points $\{x^{(i)}, y^{(i)}\}_{i=1}^N$ and $\theta = (\pi, \mu_0, \mu_1, \Sigma)$, $\pi \in [0, 1]$, $\mu_0, \mu_1 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. We suppose WLOG that $y^{(1)} = \cdots = y^{(N_0)} = 0$ for $N_0 < N$ and $y^{(N_0+1)} = \cdots = y^{(N)} = 1$.

$$
\begin{aligned}
l(\theta \mid D) &= \ln P(D \mid \theta) \\
&= \sum_{i=1}^{N} \ln P(x^{(i)}, y^{(i)} \mid \theta) \\
&= \sum_{i=1}^{N} \ln P(x^{(i)} \mid y^{(i)} \mid \theta) + \ln P(y^{(i)} \mid \theta) \\
&\propto -\frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{N_0} (x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0) - \frac{1}{2} \sum_{i=N_0+1}^{N} (x^{(i)} - \mu_1)^T \Sigma^{-1} (x^{(i)} - \mu_1) \\
&+ (N - N_0) \ln \pi + N_0 \ln(1 - \pi)
\end{aligned}
$$

1.1.1 MLE of π

To get the MLE of π, we find the stationary points of $l(\theta \mid D)$:

$$\begin{aligned}\frac{\partial l(\theta \mid D)}{\partial \pi} &= \frac{\partial}{\partial \pi}(N - N_0)\ln \pi + N_0 \ln(1-\pi) \\ &= \frac{N-N_0}{\pi} - \frac{N_0}{1-\pi}\end{aligned}$$

Setting this to 0 yields:

$$\frac{N-N_0}{\pi} = \frac{N_0}{1-\pi} \Rightarrow N\pi = N - N_0 \Rightarrow \pi = \frac{N-N_0}{N}$$

We can confirm that this is a minimum since

$$\frac{\partial^2 l(\theta \mid D)}{\partial \pi^2} = -\frac{N-N_0}{\pi^2} - \frac{N_0}{(1-\pi)^2} < 0$$

1.1.2 MLE of μ_0, μ_1

We look for stationary points for candidate MLE solutions of μ_0.

$$\frac{\partial l(\theta \mid D)}{\partial \mu_0} = \sum_{j=1}^{N_0} -\frac{1}{2}\frac{\partial}{\partial \mu_0}(x^{(j)} - \mu_0)^T \Sigma^{-1}(x^{(j)} - \mu_0) \tag{1.1}$$

To solve this, we use the chain rule. Let $f : \mathbb{R}^d \to \mathbb{R}^d$, $f(x) = x - \mu_0$ and let $g : \mathbb{R}^d \to \mathbb{R}$, $g(x) = x^T \Sigma^{-1} x$ where Σ^{-1} is symmetric and positive semi definite. From lecture results we have that $d_f(x) = I$ and $d_g(x) = 2x^T\Sigma^{-1}$. We can then compute

$$d_{g \circ f}(x) = d_g(f(x)) \cdot d_f(x) = 2(x - \mu_0)^T \Sigma^{-1} \cdot I \tag{1.2}$$

Upon substituting (1.2) into (1.1) and equating to zero we have that:

$$\frac{\partial l(\theta \mid D)}{\partial \mu_0} = \sum_{j=1}^{N_0} -((x^{(j)} - \mu_0)^T \Sigma^{-1})^T = \Sigma^{-1}(x^{(j)} - \mu_0) = 0 \tag{1.3}$$

We left multiply each side by $(\Sigma^{-1})^{-1}$ (which exists since Σ^{-1} symmetric and positive definite), and get:

$$\sum_{j=1}^{N_0} -(x^{(j)} - \mu_0) = 0 \Rightarrow \mu_0 = \frac{1}{N_0}\sum_{j=1}^{N_0} x^{(j)} = \bar{x}_0 \tag{1.4}$$

An identical computation yields $\mu_1 = \frac{1}{N-N_0}\sum_{j=N_0+1}^{N} x^{(j)} = \bar{x}_1$. Hence the MLE estimates for each class mean is the sample mean of each class.

1.1.3 MLE of Σ

We compute the MLE for Σ^{-1} instead of for Σ using the invariance of the MLE. We substitue the MLE estimate for μ_0, μ_1.

$$\frac{\partial l(\theta \mid D)}{\partial \Sigma^{-1}} = \frac{\partial}{\partial \Sigma^{-1}} \left(\frac{-N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^{N_0} (x^{(i)} - \bar{x}_0)^T \Sigma^{-1} (x^{(i)} - \bar{x}_0) - \frac{1}{2} \sum_{i=N_0+1}^{N} (x^{(i)} - \bar{x}_1)^T \Sigma^{-1} (x^{(i)} - \bar{x}_1) \right)$$

Differentiating the first term yields

$$\begin{aligned} \frac{\partial}{\partial \Sigma^{-1}} \ln |\Sigma| &= \frac{\partial}{\partial \Sigma^{-1}} - \ln |\Sigma^{-1}| && (1.5) \\ &= -\Sigma && (1.6) \end{aligned}$$

Where the derivative evaluated in (1.6) comes from a result proved in class.
Differentiating the second term yields:

$$\begin{aligned} \frac{\partial}{\partial \Sigma^{-1}} \sum_{i=1}^{N_0} (x^{(i)} - \bar{x}_0)^T \Sigma^{-1} (x^{(i)} - \bar{x}_0) &= \frac{\partial}{\partial \Sigma^{-1}} \sum_{i=1}^{N_0} tr\left((x^{(i)} - \bar{x}_0)^T \Sigma^{-1} (x^{(i)} - \bar{x}_0) \right) && (1.7) \\ &= \frac{\partial}{\partial \Sigma^{-1}} \sum_{i=1}^{N_0} tr\left((x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T \Sigma^{-1} \right) && (1.8) \\ &= \frac{\partial}{\partial \Sigma^{-1}} tr\left(\sum_{i=1}^{N_0} (x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T \Sigma^{-1} \right) && (1.9) \end{aligned}$$

Where (1.7) is using that a scalar is the Trace of a 1D matrix, (1.8) uses that the Trace is invariant under cyclic permutations, and (1.9) uses that the Trace is a linear operator. Finally, if we let $S_0 = \frac{1}{N_0} \sum_{i=1}^{N_0} (x^{(i)} - \mu_0)(x^{(i)} - \mu_0)^T$ we can rewrite (1.9) as

$$\begin{aligned} \frac{\partial}{\partial \Sigma^{-1}} \sum_{i=1}^{N_0} (x^{(i)} - \bar{x}_0)^T \Sigma^{-1} (x^{(i)} - \bar{x}_0) &= \frac{\partial}{\partial \Sigma^{-1}} N_0 tr\left(S_0 \Sigma^{-1} \right) && (1.10) \\ &= N_0 S_0 && (1.11) \end{aligned}$$

An equivalent computation yields:

$$\begin{aligned} & \frac{\partial}{\partial \Sigma^{-1}} \sum_{i=N_0+1}^{N} (x^{(i)} - \bar{x}_1)^T \Sigma^{-1} (x^{(i)} - \bar{x}_1) && (1.12) \\ &= \frac{\partial}{\partial \Sigma^{-1}} (N - N_0) tr\left(S_1 \Sigma^{-1} \right) && (1.13) \\ &= (N - N_0) S_1 && (1.14) \end{aligned}$$

Where $S_1 = \frac{1}{N-N_0}\sum_{i=N_0+1}^{N}(x^{(i)}-\mu_1)(x^{(i)}-\mu_1)^T$. Susbtituting the results from (1.6), (1.10) and (1.11) into the derivative with respect to Σ^{-1} we have that:

$$\frac{\partial l(\theta \mid D)}{\partial \Sigma^{-1}} = \frac{N}{2}\Sigma - \frac{1}{2}N_0 S_0 - \frac{1}{2}(N-N_0)S_1 \tag{1.15}$$

And so $\frac{\partial l(\theta|D)}{\partial \Sigma^{-1}} = 0 \Rightarrow N\Sigma = N_0 S_0 + (N-N_0)S_1 \Rightarrow \Sigma = \frac{N_0}{N}S_0 + \frac{N-N_0}{N}S_1$.
We leave the proof that the stationary points are maxima to the reader.

1.2 Derive $p(y=1|x)$

$$\begin{aligned}
p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} && (1.16)\\
&= \frac{\pi e^{\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}}{(1-\pi)e^{\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right\}}+\pi e^{\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}} && (1.17)\\
&= \frac{\pi e^{\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}}{(1-\pi)e^{\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right\}}+\pi e^{\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}}} && (1.18)\\
&= \frac{e^{\left\{\ln\pi-\mu_1^T\Sigma^{-1}x-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1)\right\}}}{e^{\left\{\ln(1-\pi)-\mu_0^T\Sigma^{-1}x-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0)\right\}}+e^{\left\{\ln\pi-\mu_1^T\Sigma^{-1}x-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1)\right\}}} && (1.19)
\end{aligned}$$

Where (1.19) comes from cancelling the quadratic in x from the numerator and denominator.
If we let

$$\beta_0 = \begin{bmatrix} -\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + ln(1-\pi) \\ -\Sigma^{-1}\mu_0 \end{bmatrix} \tag{1.20}$$

and

$$\beta_1 = \begin{bmatrix} -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\pi) \\ -\Sigma^{-1}\mu_1 \end{bmatrix} \tag{1.21}$$

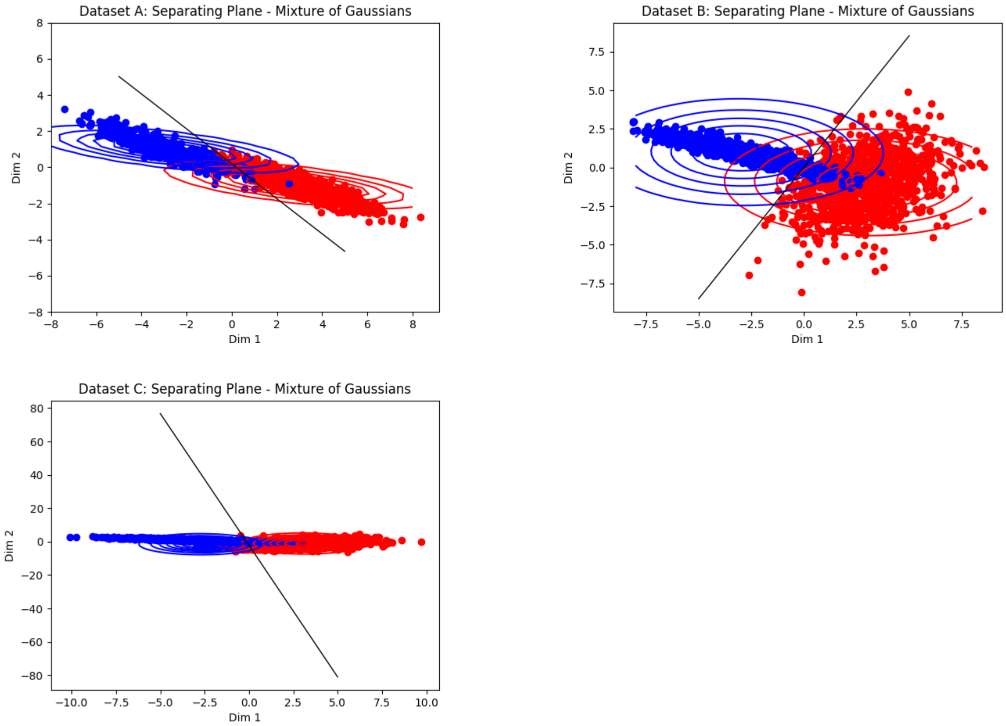
Then if we augment x to have a first component equal to one, we can rewrite (1.19) as

$$\frac{e^{\beta_1^T x}}{e^{\beta_0^T x}+e^{\beta_1^T x}} \tag{1.22}$$

Which we recognize has the same form as a Logistic Regression. Upon dividing by $e^{\beta_1^T x}$, we get that $P(y=1 \mid x) = \sigma\left(\beta_0^T x - \beta_1^T x\right)$. We notice that the decision boundary is linear in x. to find it, we note that $\sigma(z) = 0.5$ iff $z = 0$, and so we solve for $\beta_0^T x - \beta_1^T x = 0$. For $(x^1, x^2) \in \mathbb{R}^2$, this would be:

$$x^2 = -\frac{\beta_0^0 - \beta_1^0 + (\beta_0^1 - \beta_1^1)x^1}{\beta_0^1 - \beta_1^1} \tag{1.23}$$

Where $\beta_i^T := (\beta_i^0, \beta_i^1)$. We evaluating the above using the MLE estimates for 3 datasets to get figure 1.1.


Dataset A: Separating Plane - Mixture of Gaussians

Dataset B: Separating Plane - Mixture of Gaussians

Dataset C: Separating Plane - Mixture of Gaussians

Figure 1.1: Three different Datasets fit with MLE estimates from Mixture of Gaussian models. The seperating plane is in black, and contours from each gaussian are in red and blue.

Table 2.1: Logistic Regression Parameter Values

Dataset	b	w_1	w_2
A	-0.604	-3.802	-3.935
B	0.744	-1.622	0.933
C	1.342	-2.330	0.711

Table 3.1: Linear Regression Parameter Values

Dataset	b	w_1	w_2
A	0.464	-0.206	-0.211
B	0.499	-0.105	0.062
C	-0.130	-0.016	0.540

2 Logistic Regression

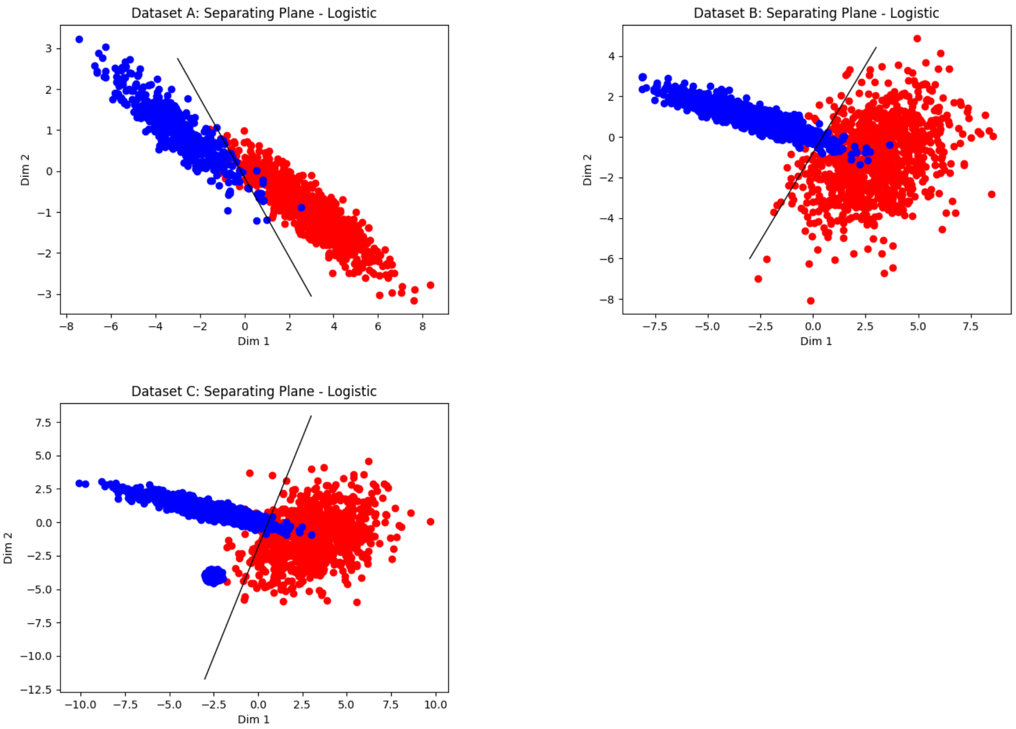
We now implement logistic regression to learn an affine function $f(x) = w^T x + b$ using the IRLS algorithm applied on each of the above datasets. The learning stopped when the norm of the difference of the parameters fell below 0.001. The estimated parameters are displayed in table 2.1. The corresponding decision boundary is presented in figure 2.1.

3 Linear Regression

We now implement linear regression by solving the normal equations. The estimated parameters are displayed in table 3.1. The corresponding decision boundary is presented in figure 3.1.

4 Comparing the Approaches

We compute the misclassification error on the training and test data. The results are presented in table 4.1.
(b) Compare the performances of the different methods on the three datasets. Is the misclas- siffication error larger, smaller, or similar on the training and test data? Why? Which methods yield very similar/dissimilar results ? Which methods yield the best results on the different datasets ? Provide an interpretation.



Dataset A: Separating Plane - Logistic

Dataset B: Separating Plane - Logistic

Dataset C: Separating Plane - Logistic

Figure 2.1: Three different Datasets fit with IRLS estimates from Logistic Regression model. The seperating plane is in black.

Table 4.1: Test error for the models applied to the datasets, with ground truth values in brackets

Dataset	MoG	Logistic	Linear
A	0.02 (0.026)	0.027 (0.017)	0.027 (0.016)
B	0.03 (0.0415)	0.023 (0.038)	0.03 (0.0415)
C	0.0387 (0.0387)	0.026 (0.026)	0.043 (0.043)

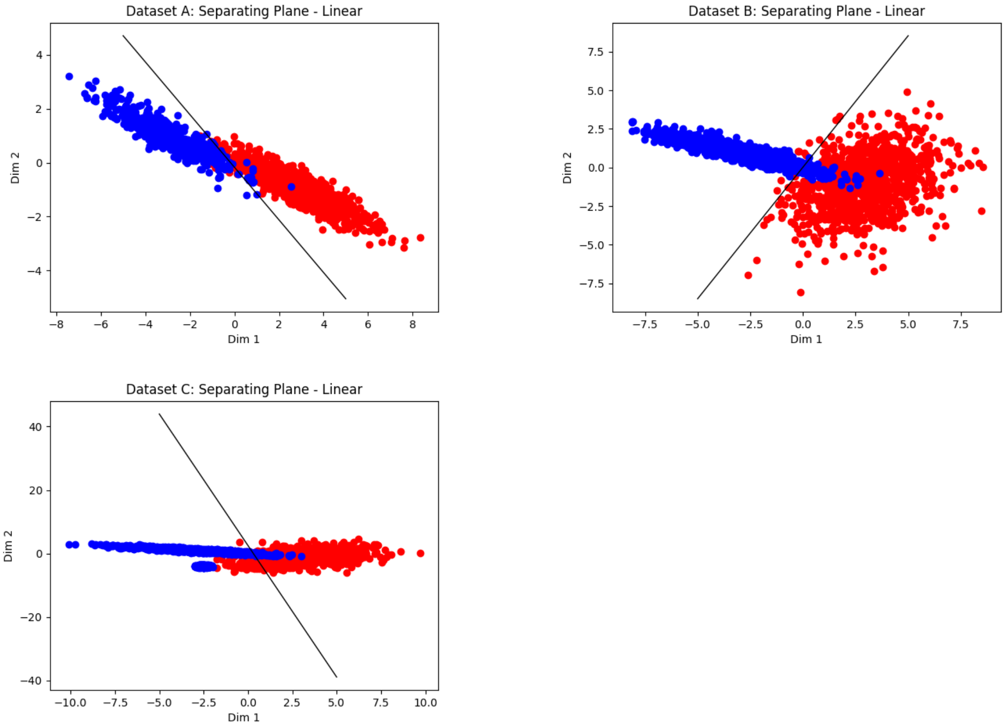


Figure 3.1: Three different Datasets fit with the Linear Regression model. The seperating plane is in black.