Ryan Farr – rlf238
October 21st, 2018
CS 5785 – Applied Machine Learning

Homework 2 – Written Portion

1. Here we have a prime example of where to apply the Lagrange multiplier with the constraint $a^T W a = 1$ entered as $\lambda(a^T W a - 1)$:
$$l(\lambda) = a^T B a - \lambda(a^T W a - 1)$$

   The minimum can be found when the derivative of this function is 0

$$\frac{\partial l}{\partial a} = Ba + B^T a - \lambda(Wa - W^T a) = (B + B^T)a - \lambda(W + W^T)a = 0$$

$$\lambda a = (W^T + W)^{-1}(B^T + B)a$$
   Which is the standard eigenvalue problem.

2. With features $x \in \mathbb{R}^p$, a two-class response and class sizes N_1, N_2 coded as $-\frac{N}{N_1}, \frac{N}{N_2}$.
   a. We know $\delta_1(x) = x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1 \Sigma^{-1} \mu_1 + \log \pi_1$ and $\delta_2(x) = x^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_2 \Sigma^{-1} \mu_2 + \log \pi_2$ and that we will classify as class 2 when $\delta_1 < \delta_2$

$$x^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_2 \Sigma^{-1} \mu_2 + \log \pi_2 > x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1 \Sigma^{-1} \mu_1 + \log \pi_1$$

$$x^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_2 \Sigma^{-1} \mu_2 + \log \frac{N_2}{N} > x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1 \Sigma^{-1} \mu_1 + \log \frac{N_1}{N}$$

$$x^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_2 \Sigma^{-1} \mu_2 > x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1 \Sigma^{-1} \mu_1 + \log \frac{N_1}{N} - \log \frac{N_2}{N}$$

$$x^T \Sigma^{-1} (\mu_2 - \mu_1) > \frac{1}{2}\mu_2 \Sigma^{-1} \mu_2 + x^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_1 \Sigma^{-1} \mu_1 + \log \frac{N_2}{N_1}$$

$$x^T \Sigma^{-1} (\mu_2 - \mu_1) > \frac{1}{2}(\mu_2 + \mu_1) \Sigma^{-1} (\mu_2 - \mu_1) + \log \frac{N_2}{N_1}$$
   b. Because this is a two-class problem we can create $U = U_1 + U_2$ where $U1 \in \mathbb{R}^N$ is a vector of 1s where the sample is of class 1 and 0 otherwise and vice versa for U_2.

$$\sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2$$

Can be rewritten as

$$(Y-\beta_0 U - X\beta)^T(Y-\beta_0 U - X\beta)$$

Where $Y = -\frac{N}{N_1}U_1 + \frac{N}{N_2}U_2$. To minimize, we take the derivative and set it to 0.

$$\frac{\partial RSS}{\partial \beta_0} = 2U^T U\beta_0 - 2U^T(Y - X\beta) = 2N\beta_0 - 2U^T(Y-X\beta) = 0$$

$$2N\beta_0 = 2U^T(Y-X\beta)$$

$$B_0 = \frac{1}{N}U^T(Y-X\beta)$$

And

$$\frac{\partial RSS}{\partial \beta} = 2X^T X\beta - 2X^T Y + 2\beta_0 X^T U = 0$$

$$2X^T X\beta - 2X^T Y + 2\left(\frac{1}{N}U^T(Y-X\beta)\right)X^T U = 0$$

$$X^T X\beta - X^T Y + \frac{1}{N}U^T Y X^T U - \frac{1}{N}U^T X\beta X^T U = 0$$

$$X^T X\beta - \frac{1}{N}X^T U U^T X\beta = X^T Y - \frac{1}{N}X^T U U^T Y$$

$$(X^T X - \frac{1}{N}X^T U U^T X)\beta = X^T Y - \frac{1}{N}\, X^T U U^T Y$$

Note that $X^T U_1 = N_1\mu_1$ and the same for class 2.

$$X^T X - \frac{1}{N}(X^T U U^T X)\beta = X^T X - \frac{1}{N}(N_1^2\mu_1\mu_1^T + N_2^2\mu_2\mu_2^T + N_1N_2\mu_1\mu_2^T + N_1N_2\mu_2\mu_1^T)$$

We know that the covariance matrix is

$$(N-2)\Sigma = \sum(x_i-\mu_1)(x_i-\mu_i)^T + \sum(x_i-\mu_2)(x_i-\mu_2)^T = X^T X - N_1\mu_1\mu_1^T - N_1\mu_2\mu_2^T$$

From the problem we'll define $\Sigma_\beta\beta = \frac{N_1N_2}{N^2}(\mu_2-\mu_1)(\mu_2-\mu_1)^T$

$$(N-2)\Sigma + \frac{N_1N_2}{N}\Sigma_\beta$$
$$= X^T X + \left(\frac{N_1N_2}{N} - N_1\right)\mu_1\mu_1^T + \left(\frac{N_1N_2}{N} - N_2\right)\mu_2\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T - \frac{N_1N_2}{N}\mu_1\mu_2^T$$
$$= X^T X - \frac{N_1^2}{N}\mu_1\mu_1^T - \frac{N_2^2}{N}\mu_2\mu_2^T - \frac{N_1N_2}{N}\mu_2\mu_1^T - \frac{N_1N_2}{N}\mu_1\mu_2^T$$

$$= \frac{1}{N}(N_1^2\mu_1\mu_1^T + N_2^2\mu_2\mu_2^T + N_1N_2\mu_1\mu_2^T + N_1N_2\mu_2\mu_1^T)$$

Note that $X^TY = X^T\left(-\frac{N}{N_1}U_1 + \frac{N}{N_2}U_2\right) = \frac{-\frac{N}{N_1}N_1^2+\frac{N}{N_2}N_1N_2}{N}\mu_1 + \frac{\frac{N}{N_2}N_2^2-\frac{N}{N_1}N_1N_2}{n}\mu_2 = \frac{N_1N_2}{N}\left(-\frac{N}{N_1}-\frac{N}{N_2}\right)(\mu_1-\mu_2)$ and $-\frac{N}{N_1}-\frac{N}{N_2} = \frac{N(N_1+N_2)}{N_1N_2} = -\frac{N^2}{N_1N_2}$ thus $\frac{N_1N_2}{N}\left(-\frac{N}{N_1}-\frac{N}{N_2}\right)(\mu_1-\mu_2) = N(\mu_2-\mu_1)$. Therefore:

$$\left((N-2)\,\Sigma + \frac{N_1N_2}{N}\Sigma_\beta\right)\beta = N(\mu_2-\mu_1)$$

c. We know that

$$\Sigma_\beta\beta \propto (\mu_2-\mu_1)(\mu_2-\mu_1)^T\beta = k(\mu_2-\mu_1)$$

Where $k = (\mu_2-\mu_1)^T\beta$. And thus, $\Sigma_\beta\beta$ is in the direction $(\mu_2-\mu_1)$. We can plug this into our equation in part b

$$(N-2)\Sigma\beta + k(\mu_2-\mu_1) = N(\mu_2-\mu_1)$$

$$\beta \propto \Sigma^{-1}(\mu_2-\mu_1)$$

And thus, the least-squares regression coefficient is identical to the LDA coefficient up to a scalar multiple.

d. If we say $a_1 = -\frac{N}{N_1}$ and $a_2 = \frac{N}{N_2}$ then we have

$$\left((N-2)\,\Sigma + \frac{N_1N_2}{N}\Sigma_\beta\right)\beta = \frac{N_1N_2}{N}(a_1-a_2)(\mu_1-\mu_2)$$

Which holds for any encodings for a_1 and a_2.

e. Note that $U^TY = -\frac{N}{N_1}N_1 + \frac{N}{N_2}N_2 = 0$. Therefore:

$$\beta_0 = \frac{1}{N}U^T(Y-X\beta) = \frac{1}{N}X\beta = -\frac{N_1}{N}\mu_1^T\beta - \frac{N_2}{N}\mu_2^T\beta = -\left(\frac{N_1}{N}\mu_T - \frac{N_2}{N}\mu_2^T\right)\beta$$

With the decision rule $f(x) = \beta_0 + x^T\beta$ we get:

$$f(x) = \beta_0 + x^T\beta = -\left(\frac{N_1}{N}\mu_1^T + \frac{N_2}{N}\mu_2^T\right)\beta + x^T\beta$$

If we're classifying to class 2 when $f(x) > 0$ then

$$x^T\beta > \left(\frac{N_1}{N}\mu_1^T + \frac{N_2}{N}\mu_2^T\right)\beta$$

And from part c we know $\beta \propto \Sigma^{-1}(\mu_2-\mu_1)$

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \left(\frac{N_1}{N}\mu_1^T + \frac{N_2}{N}\mu_2^T\right)\Sigma^{-1}(\mu_2-\mu_1)$$

And if $N_1 = N_2$ then

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{N_1}{N}(\mu_1^T+\mu_2^T)\Sigma^{-1}(\mu_2-\mu_1) - 0$$
$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{N_1}{N}(\mu_1^T+\mu_2^T)\Sigma^{-1}(\mu_2-\mu_1) - \log(1)$$

$$x^T\Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}(\mu_1^T+\mu_2^T)\Sigma^{-1}(\mu_2-\mu_1) - \log\left(\frac{N_1}{N_2}\right)$$

Which is the same as the LDA classification rule for class 2.

3. Given M

```
[ 1  0  3]
[ 3  7  2]
[ 2 -2  8]
[ 0 -1  1]
[ 5  8  7]
```

a. M^TM is given by

```
[ 39  57  60]
[ 57 118  53]
[ 60  53 127]
```

And MM^T is given by

```
[ 10   9  26   3  26]
[  9  62   8  -5  85]
[ 26   8  72  10  50]
[  3  -5  10   2  -1]
[ 26  85  50  -1 138]
```

b. The eigenvalues for M^TM are
214.670489196, 9.32587340685e-15, 69.3295108039

Note that the middle element is essentially 0.

The eigenvalues for MM^T are
214.670489196, -8.881784197e-16, 69.3295108039, -3.34838280783e-15, 7.47833226761e-16

Note that all but 214.67048916 and 69.3295108 are essentially 0.

c. The eigenvectors for M^TM are
214.670489196: [0.42615127 0.61500884 0.66344497]
69.3295108039: [-0.01460404 -0.72859799 0.68478587]

The eigenvectors for MM^T are
214.670489196: [-0.16492942 -0.47164732 -0.33647055 -0.00330585 -0.79820031]
69.3295108039: [0.24497323 -0.45330644 0.82943965 0.16974659 -0.13310656]

d. We know that Σ is the diagonal matrix of the square root of $M^T M$'s eigenvalues

```
[ 14.65163776   0.        ]
[  0.           8.32643446]
```

And V is composed of the eigenvectors of $M^T M$, thus V^T is

```
[ 0.42615127  0.90453403 -0.01460404]
[ 0.66344497 -0.30151134  0.68478587]
```

And we know that $M = U\Sigma V^T$ which means $U = \mathrm{MV}\Sigma^{-1}$ which is

```
[ 0.16492942  0.24497323]
[ 0.47164732 -0.45330644]
[ 0.33647055  0.82943965]
[ 0.00330585  0.16974659]
[ 0.79820031 -0.13310656]
```

e. We can find the one-dimensional approximation to M by taking $U[:,:1]\Sigma[:1,:1]V^T[:1,:]$ which gives:

```
[ 1.02978864  1.48616035  1.60320558]
[ 2.94487812  4.24996055  4.58467382]
[ 2.10085952  3.031898    3.27068057]
[ 0.02064112  0.02978864  0.0321347 ]
[ 4.9838143   7.19249261  7.75895028]
```

NOTE: All matrix arithmetic was done using python. See *main.ipynb* "Written 3.A" through "Written 3.E" for calculations.