HW4 Solution, ECE7252, February 25, 2008

1. The discriminant function for two classes, $k = 1, 2$, based on Eq. (4.10) is simply

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k. \tag{1}$$

(a) The LDA rule of deciding a new sample x to be classified as class 2 is that $\delta_2(x) > \delta_1(x)$, or

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

If we estimate the unknown parameters, μ_1, μ_2, Σ, π_1 and π_2 according to the equations below Eq. (4.10) with $\hat{\pi}_k = N_k/N$, we have

$$x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2} \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1] - \log(\frac{N_2}{N}) + \log(\frac{N_1}{N}). \tag{3}$$

(b) Minimizing the square error $\sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i \right)^2$

$$\sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i \right)^2 = \sum_{i=1}^{N_1} \left(y_i - \beta_0 - \beta^T x_i \right)^2 + \sum_{i=1}^{N_2} \left(y_j - \beta_0 - \beta^T x_j \right)^2$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta^T x_i \right)^2 \propto \sum_{i=1}^{N_1} \left(y_i - \beta_0 - \beta^T x_i \right) + \sum_{i=1}^{N_2} \left(y_j - \beta_0 - \beta^T x_j \right) = 0$$

$$= \frac{-N}{N_1} N_1 - N_1 \beta_0 - N_1 \beta^T \hat{\mu}_1 + \frac{N}{N_2} N_2 - N_2 \beta_0 - N_2 \beta^T \hat{\mu}_2$$

$$\boxed{\beta_0 = -\frac{N_1 \beta^T \hat{\mu}_1 + N_2 \beta^T \hat{\mu}_2}{N_1 + N_2} = -\frac{N_1 \beta^T \hat{\mu}_1 + N_2 \beta^T \hat{\mu}_2}{N} = -\frac{N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T}{N} \beta}$$

$$\sum_{i=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2=\sum_{i=1}^{N_1}\left(y_i-\beta_0-\beta^T x_i\right)^2+\sum_{i=1}^{N_2}\left(y_j-\beta_0-\beta^T x_j\right)^2$$

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{N}\left(y_i-\beta_0-\beta^T x_i\right)^2\propto\sum_{i=1}^{N_1}\left(y_i-\beta_0-\beta^T x_i\right)x_i+\sum_{i=1}^{N_2}\left(y_j-\beta_0-\beta^T x_j\right)x_j=0$$

$$=\sum_{i=1}^{N_1}y_i x_i-\sum_{i=1}^{N_1}\beta_0 x_i-\sum_{i=1}^{N_1}\beta x_i x_i^T+\sum_{i=1}^{N_2}y_j x_j-\sum_{i=1}^{N_2}\beta_0 x_j-\sum_{i=1}^{N_2}\beta x_j x_j^T$$

$$=\frac{-N}{N_1}N_1\hat{\mu}_1-\beta_0 N_1\hat{\mu}_1-\left[\left(N_1-1\right)\hat{\mathbf{\Sigma}}+N_1\hat{\mu}_1\hat{\mu}_1^T\right]\beta+$$

$$\frac{N}{N_2}N_2\hat{\mu}_2-\beta_0 N_2\hat{\mu}_2-\left[\left(N_2-1\right)\hat{\mathbf{\Sigma}}+N_2\hat{\mu}_2\hat{\mu}_2^T\right]\beta$$

$$=N\left(\hat{\mu}_2-\hat{\mu}_1\right)-\left(N-2\right)\hat{\mathbf{\Sigma}}\beta-\frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_B\beta$$

$$\therefore\boxed{\left[\left(N-2\right)\hat{\mathbf{\Sigma}}-\frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_B\right]\beta=N\left(\hat{\mu}_2-\hat{\mu}_1\right)}$$

Where we use

$$-\beta_0 N_1\hat{\mu}_1-N_1\hat{\mu}_1\hat{\mu}_1^T\beta-\beta_0 N_2\hat{\mu}_2-N_2\hat{\mu}_2\hat{\mu}_2^T\beta$$

$$=\frac{N_1\beta^T\hat{\mu}_1+N_2\beta^T\hat{\mu}_2}{N}\left(N_1\hat{\mu}_1+N_2\hat{\mu}_2\right)-N_1\hat{\mu}_1\hat{\mu}_1^T\beta-N_2\hat{\mu}_2\hat{\mu}_2^T\beta$$

$$=\frac{N_1^2\hat{\mu}_1\hat{\mu}_1^T+N_1N_2\hat{\mu}_1\hat{\mu}_2^T+N_1N_2\hat{\mu}_1\hat{\mu}_2^T+N_2^2\hat{\mu}_2\hat{\mu}_2^T-NN_1\hat{\mu}_1\hat{\mu}_1^T\beta-NN_2\hat{\mu}_2\hat{\mu}_2^T\beta}{N}$$

$$=-\frac{N_1N_2}{N}\left(\hat{\mu}_1\hat{\mu}_1^T-2\hat{\mu}_1\hat{\mu}_2^T+\hat{\mu}_2\hat{\mu}_2^T\right)=-\frac{N_1N_2}{N}\left(\hat{\mu}_2-\hat{\mu}_1\right)\left(\hat{\mu}_2-\hat{\mu}_1\right)^T$$

$$\triangleq-\frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_B$$

(c) From (b),

$$\hat{\mathbf{\Sigma}}_B = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T$$

$$\hat{\mathbf{\Sigma}}_B \beta = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T \beta = (\hat{\mu}_2 - \hat{\mu}_1)\left[(\hat{\mu}_2 - \hat{\mu}_1)^T \beta\right]$$

$$(\hat{\mu}_2 - \hat{\mu}_1) \times scalar \,//\, (\hat{\mu}_2 - \hat{\mu}_1)$$

Thus,

$$\left[(N-2)\hat{\mathbf{\Sigma}} + \frac{N_1 N_2}{N}\hat{\mathbf{\Sigma}}_B\right]\beta = N(\mu_2 - \mu_1)$$

$$(N-2)\hat{\mathbf{\Sigma}}\beta = (N + scalar)(\mu_2 - \mu_1)$$

$$\beta = \frac{(N + const)}{N-2}\hat{\mathbf{\Sigma}}^{-1}(\mu_2 - \mu_1) \propto \hat{\mathbf{\Sigma}}^{-1}(\mu_2 - \mu_1)$$

(d) Now, assume we have arbitrary label A and B for class 1 and 2, respectively:

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta^T x_i\right)^2 = \sum_{i=1}^{N_1}\left(y_i - \beta_0 - \beta^T x_i\right)^2 + \sum_{i=1}^{N_2}\left(y_j - \beta_0 - \beta^T x_j\right)^2$$

$$\frac{\partial}{\partial \beta}\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta^T x_i\right)^2 \propto \sum_{i=1}^{N_1}\left(y_i - \beta_0 - \beta^T x_i\right) + \sum_{i=1}^{N_2}\left(y_j - \beta_0 - \beta^T x_j\right) = 0$$

$$= AN_1 - N_1\beta_0 - N_1\beta^T\hat{\mu}_1 + BN_2 - N_2\beta_0 - N_2\beta^T\hat{\mu}_2$$

$$\beta_0 = -\frac{N_1\beta^T\hat{\mu}_1 + N_2\beta^T\hat{\mu}_2 - AN_1 - BN_2}{N_1 + N_2}$$

$$\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta^T x_i\right)^2 = \sum_{i=1}^{N_1}\left(y_i - \beta_0 - \beta^T x_i\right)^2 + \sum_{i=1}^{N_2}\left(y_j - \beta_0 - \beta^T x_j\right)^2$$

$$\frac{\partial}{\partial \beta}\sum_{i=1}^{N}\left(y_i - \beta_0 - \beta^T x_i\right)^2 \propto \sum_{i=1}^{N_1}\left(y_i - \beta_0 - \beta^T x_i\right)x_i + \sum_{i=1}^{N_2}\left(y_j - \beta_0 - \beta^T x_j\right)x_j = 0$$

$$= \sum_{i=1}^{N_1} y_i x_i - \sum_{i=1}^{N_1}\beta_0 x_i - \sum_{i=1}^{N_1}\beta x_i x_i^T + \sum_{i=1}^{N_2} y_j x_j - \sum_{i=1}^{N_2}\beta_0 x_j - \sum_{i=1}^{N_2}\beta x_j x_j^T$$

$$= AN_1\hat{\mu}_1 - \beta_0 N_1\hat{\mu}_1 - \left[(N_1 - 1)\hat{\mathbf{\Sigma}} + N_1\hat{\mu}_1\hat{\mu}_1^T\right]\beta +$$

$$BN_2\hat{\mu}_2 - \beta_0 N_2\hat{\mu}_2 - \left[(N_2 - 1)\hat{\mathbf{\Sigma}} + N_2\hat{\mu}_2\hat{\mu}_2^T\right]\beta$$

$$= -(N-2)\hat{\Sigma}\beta - \frac{N_1 N_2}{N}\hat{\Sigma}_B\beta + AN_1\hat{\mu}_1 + BN_2\hat{\mu}_2$$
$$- \frac{AN_1^2\hat{\mu}_1 + BN_1N_2\hat{\mu}_1 + AN_1N_2\hat{\mu}_2 + BN_2^2\hat{\mu}_2}{N}$$
$$= -(N-2)\hat{\Sigma}\beta - \frac{N_1 N_2}{N}\hat{\Sigma}_B\beta + \frac{AN_1N_2\hat{\mu}_1 - BN_1N_2\hat{\mu}_1 - AN_1N_2\hat{\mu}_2 + BN_1N_2\hat{\mu}_2}{N}$$
$$= -(N-2)\hat{\Sigma}\beta - \frac{N_1 N_2}{N}\hat{\Sigma}_B\beta + \frac{BN_1N_2 - AN_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1) = 0$$
$$\therefore \boxed{\left[(N-2)\hat{\Sigma} + \frac{N_1N_2}{N}\hat{\Sigma}_B\right]\beta = \frac{BN_1N_2 - AN_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)}$$

which has exactly the same form as those obtained in (b), except now we have $\frac{BN_1N_2 - AN_1N_2}{N}$ rather than N, Thus the properties in (c) still hold

Where we use

$$-\beta_0 N_1\hat{\mu}_1 - N_1\hat{\mu}_1\hat{\mu}_1^T\beta - \beta_0 N_2\hat{\mu}_2 - N_2\hat{\mu}_2\hat{\mu}_2^T\beta$$
$$= -\frac{N_1N_2}{N}\hat{\Sigma}_B - \frac{AN_1^2\hat{\mu}_1 + BN_1N_2\hat{\mu}_1 + AN_1N_2\hat{\mu}_2 + BN_2^2\hat{\mu}_2}{N}$$

(e) Clearly, the result of LDA has the following intercept:

$$-\frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) + \log\left(\frac{N_2}{N_1}\right)$$

from (b), we know $\beta_0 = -\frac{N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T}{N}\beta$ and from (c), we know

$\beta \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ thus

$$\beta_0 \propto -\frac{N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T}{N}\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

Obviously, only when $N_1 = N_2 = N/2$, the first term will become $-\frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T$, and this will also let logarithm term in LDA go to 0.

2. HTF Exercise 4.6, part (a)

Clearly, separability means one class will always have positive value when inner product with β, and the other class will always generate negative value. The

normalization will make all data point to have norm equal to 1. It is also clear that we can choose class label arbitrarily so that the class with positive value when take inner product with some β be labeled as +1, and the other can be labeled as -1. Since multiply a β by a minus sign will still generate the same hyper plane $\beta^T x^*=0$. so we have this freedom to make choice. Without losing of generality, we assume class one should give negative value and class two should give positive value when taking inner product. By this convention,

$$y_1\beta^T z_1 = -\beta^T z_1 = -\|\beta\|\|z_1\|\cos(\angle\beta, z_1) = -\|\beta\|\cos(\angle\beta, z_1) > 0,$$

$$\Rightarrow \|\beta\| \geq -\|\beta\|\cos(\angle\beta, z_1) > 0$$

$$y_2\beta^T z_2 = \beta^T z_2 = \|\beta\|\|z_1\|\cos(\angle\beta, z_2) = \|\beta\|\cos(\angle\beta, z_2) > 0$$

$$\Rightarrow \|\beta\| \geq \|\beta\|\cos(\angle\beta, z_2) > 0$$

where we use

$$\beta^T z_1 < 0 \Rightarrow -1 \leq \cos(\angle\beta, z_1) < 0$$

$$\beta^T z_2 > 0 \Rightarrow 0 < \cos(\angle\beta, z_1) \leq 1$$

Clearly, assume original β has norm 1, if we scale β by

$$\frac{1}{\min\left\{\left|\cos(\angle\beta, z_1)\right|, \left|\cos(\angle\beta, z_2)\right|\right\}}$$ for all data from 2 classes,

we will have

$$y_1\beta_{opt}^T z_1 = -\left\|\beta_{opt}\right\|\cos(\angle\beta, z_1) = -\frac{\cos(\angle\beta, z_1)}{\min\left\{\left|\cos(\angle\beta, z_1)\right|, \left|\cos(\angle\beta, z_2)\right|\right\}} \geq 1$$

$$y_1\beta_{opt}^T z_1 = \left\|\beta_{opt}\right\|\cos(\angle\beta, z_2) = -\frac{\cos(\angle\beta, z_2)}{\min\left\{\left|\cos(\angle\beta, z_1)\right|, \left|\cos(\angle\beta, z_2)\right|\right\}} \geq 1$$

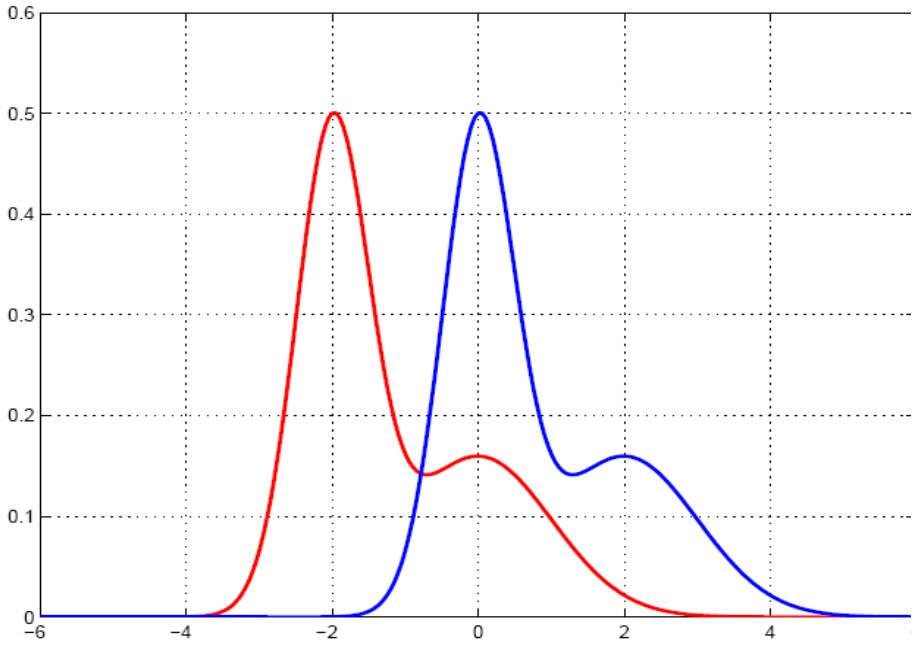
This is fine because scaling the normal vector will not affect the decision hyper plane.

3.

- Each conditional density is a mixture of two normals:

– Class 1 (red): $0.6N(-2, \frac{1}{4}) + 0.4N(0, 1)$.
– Class 2 (blue): $0.6N(0, \frac{1}{4}) + 0.4N(2, 1)$.

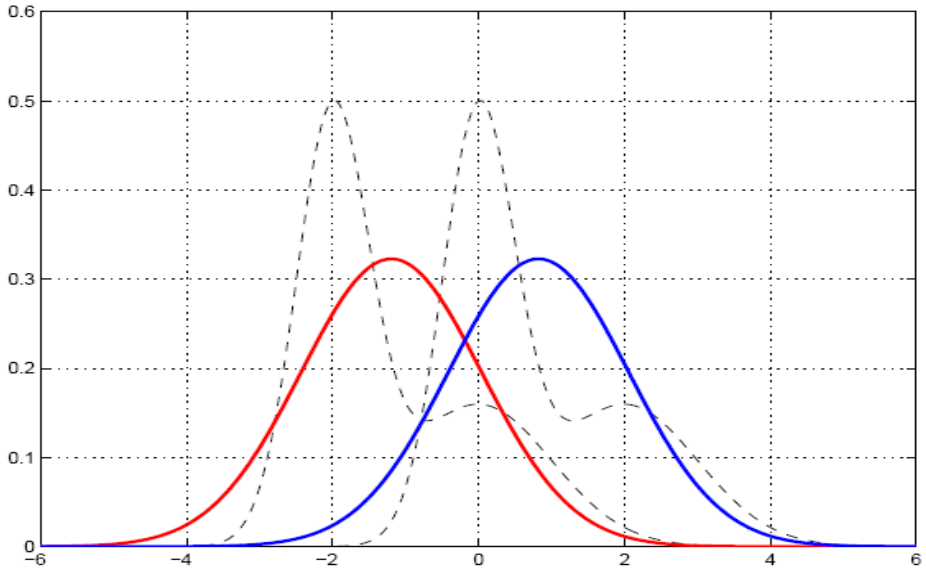
- The class-conditional densities are shown below



LDA Result

- Training data set: 2000 samples for each class.
- Test data set: 1000 samples for each class.
- The estimation by LDA: $\hat{\mu}_1 = -1.1948$, $\hat{\mu}_2 = 0.8224$, $\hat{\sigma}^2 = 1.5268$. Boundary value between the two classes is $(\hat{\mu}_1 + \hat{\mu}_2)/2 = -0.1862$.
- The classification error rate on the test data is 0.2315.

- Based on the true distribution, the Bayes (optimal) boundary value between the two classes is -0.7750 and the error rate is 0.1765.



Logistic Regression Result

- Linear logistic regression obtains

$$\beta = (-0.3288, -1.3275)^T .$$

The boundary value satisfies $-0.3288 - 1.3275X = 0$, hence equals -0.2477.

- The error rate on the test data set is 0.2205.
- The estimated posterior probability is:

$$Pr(G = 1 \mid X = x) = \frac{e^{-0.3288-1.3275x}}{1 + e^{-0.3288-1.3275x}} .$$