HW6 Solution, ECE7252 Spring 2008 (March 24)

1. HTF Exercise 12.1

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \left[1 - y_i f(x_i)\right]_+ + \lambda\|\beta\| = \min_{\beta_0,\beta} \sum_{i=1}^{N} \left\{1 - y_i\left[h(x_i)^T \beta + \beta_0\right]\right\}_+ + \lambda\|\beta\| \quad (12.25)$$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N} \xi_i \quad \text{subject to } \xi_i \geq 0,\ y_i\left(x_i^T\beta + \beta_0\right) \geq 1 - \xi_i \ \forall i \quad (12.8)$$

Clearly, all constraints and objectives listed above are convex.
Applying the Lagrange function on (12.8), since the constraints in (12.8) satisfy Slater condition ($-\xi_i < 0,\ 1 - \xi_i - y_i\left(x_i^T\beta + \beta_0\right) < 0,\ \forall i$ is solvable) and is below bounded, the Strong duality of convex programming duality theorem gives the optimal solution of the dual problem is equal to that of the prime problem. Consider the dual problem of (12.8)
That is, $\sup_\lambda\{\inf_\beta[L(12.8)]\} = \inf_\beta\{\sup_\lambda [L(12.8)]\}$, with all non-strict constraints, inf = min, sup = max.

$$\max_{\lambda_1,\lambda_2} \frac{1}{2}\|\beta^*\|^2 + \gamma\sum_{i=1}^{N} \xi_i \ - \sum_{i=1}^{N} \lambda_{1i}\xi_i + \sum_{i=1}^{N} \lambda_{2i}\left[1 - \xi_i - y_i\left(x_i^T\beta^* + \beta_0^*\right)\right]$$

Where β^* and β_0^* are optimal solution of the supermom of the Lagrange function. We can see that for this convex programming to be optimal, examine the extreme points where all constraints with β are active (i.e. equality holds),
$1 - \xi_i - y_i\left(x_i^T\beta + \beta_0\right) = 0 \Rightarrow 1 - y_i\left(x_i^T\beta + \beta_0\right) = \xi_i \geq 0$
That is, we should only pick the λ_{2i} terms which correspond to non-negative $1 - y_i\left(x_i^T\beta + \beta_0\right)$, and leave others 0. Since finding β that will result in minimum of norm β times some constant λ is exactly the same as finding β that will results in smallest half squared norm of it if λ is non-negative. Thus we can conclude (12.25) and (12.8) will generate same results:

Choose $\lambda = \dfrac{1}{2\gamma}$, and pick $\lambda_{2i} = 1$ for all non - negative $\left\{1 - y_i\left[x_i^T\beta + \beta_0\right]\right\}$.

Since the optimal solution occur when constraints are active, $1 - y_i\left[x_i^T\beta + \beta_0\right] = \xi_i$ when $\left\{1 - y_i\left[x_i^T\beta + \beta_0\right]\right\} \geq 0$

$$\Rightarrow \sum_{i=1}^{N} \left\{1 - y_i\left[x_i^T\beta + \beta_0\right]\right\}_+ = \sum_{i=1}^{N} \xi_i \Leftrightarrow \arg\min_{\beta,\beta_0} \sum_{i=1}^{N} \left\{1 - y_i\left[x_i^T\beta + \beta_0\right]\right\}_+ + \lambda\|\beta\| = \arg\min_{\beta,\beta_0} \sum_{i=1}^{N} \xi_i + \lambda\|\beta\|$$

$$= \arg\min_{\beta,\beta_0} \sum_{i=1}^{N} \xi_i + \lambda\|\beta\|^2 = \arg\min_{\beta,\beta_0} \gamma\sum_{i=1}^{N} \xi_i + \frac{1}{2}\|\beta\|^2$$

2. HTF Exercise 12.9

We can simply plug in the transformation into decision rule of LDA in (4.29), since for Gaussian distribution, the linear transform by a matrix A, its mean will go from μ to A μ, and variance will go from Σ to $A\Sigma A^T$. Thus for the new variable x*, the equation for decision rule can be expressed in x as follows:

$$\log\left(\frac{\pi_k}{\pi_K}\right)-\frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)\hat{\mathbf{B}}\left[\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\Sigma}\hat{\mathbf{B}}\right]^{-1}\hat{\mathbf{B}}^{\mathrm{T}}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)+\mathbf{x}^T\hat{\mathbf{B}}\left[\hat{\mathbf{B}}^{\mathrm{T}}\boldsymbol{\Sigma}\hat{\mathbf{B}}\right]^{-1}\hat{\mathbf{B}}^T(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)$$

Assume there is no redundant class, that is, column rank of the indicator matrix $\hat{\mathbf{Y}}$ = K, then the matrix $\hat{\mathbf{B}}$ will not be rank deficient. (i.e. $\exists\hat{\mathbf{B}}^{\dagger}\rightarrow\hat{\mathbf{B}}\hat{\mathbf{B}}^{\dagger}=\mathbf{I}_p$ and $\left[\hat{\mathbf{B}}^T\right]^{\dagger}\rightarrow\left[\hat{\mathbf{B}}^T\right]^{\dagger}\hat{\mathbf{B}}^T=\mathbf{I}_K$), thus we can write:

$$\log\left(\frac{\pi_k}{\pi_K}\right)-\frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)\hat{\mathbf{B}}\hat{\mathbf{B}}^{\dagger}\boldsymbol{\Sigma}^{-1}\left[\hat{\mathbf{B}}^T\right]^{\dagger}\hat{\mathbf{B}}^{\mathrm{T}}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)+\mathbf{x}^T\hat{\mathbf{B}}\hat{\mathbf{B}}^{\dagger}\boldsymbol{\Sigma}^{-1}\left[\hat{\mathbf{B}}^T\right]^{\dagger}\hat{\mathbf{B}}^T(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)$$

$$=\log\left(\frac{\pi_k}{\pi_K}\right)-\frac{1}{2}(\boldsymbol{\mu}_k+\boldsymbol{\mu}_K)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_K)$$

which is exactly the same as original LDA.