第四次作业

Ext 4.1

Ex. 4.1 Show how to solve the generalized eigenvalue problem $\max a^T \mathbf{B} a$ subject to $a^T \mathbf{W} a = 1$ by transforming to a standard eigenvalue problem.

Consider using the idea of Lagrangian multipliers. And we define

$$\mathcal{L}(a; \lambda) = a^\top B a + \lambda(a^\top W a - 1)$$

And we taking the a derivative of this equation and set it to 0. Get

$$W^{-1} B a = \lambda a$$

so we can say that λ, a is a pair of eigenvalue and eigenvector.

Ext 4.2

Ex. 4.2 Suppose we have features $x \in \mathbb{R}^p$, a two-class response, with class sizes N_1, N_2, and the target coded as $-N/N_1, N/N_2$.

a

(a) Show that the LDA rule classifies to class 2 if

$$x^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(N_2/N_1),$$

and class 1 otherwise.

Following the equation (4.9) in textbook, we only need to notice that $\pi_1 = \frac{N_1}{N}$ and $\pi_2 = \frac{N_2}{N}$.

b

(b) Consider minimization of the least squares criterion

$$\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2. \tag{4.55}$$

Show that the solution $\hat{\beta}$ satisfies

$$\left[(N-2)\hat{\mathbf{\Sigma}} + N\hat{\mathbf{\Sigma}}_B\right]\beta = N(\hat{\mu}_2 - \hat{\mu}_1) \tag{4.56}$$

(after simplification), where $\hat{\mathbf{\Sigma}}_B = \frac{N_1 N_2}{N^2}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T$.

Follows the notation of textbook, and we additionally add a row 1 into X, so

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} N & \sum_{i=1}^N x_i^T \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i x_i^T \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}.$$

In order to get the minimum point, we need to make it satisfies:

$$\mathbf{X}^T\mathbf{X}\begin{pmatrix}\beta_0 \\ \beta\end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^N x_i^T \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i x_i^T \end{pmatrix}\begin{pmatrix}\beta_0 \\ \beta\end{pmatrix} = \mathbf{X}^T Y = \begin{pmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N y_i x_i \end{pmatrix}$$

And we solve this equation in to parts, we get:

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

and

$$\left[\sum_{i=1}^{N} x_i x_i^T - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i^T\right)\right]\beta = \sum_{i=1}^{N} y_i x_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i\right).$$

By using

$$\hat{\mu}_1 = \frac{1}{N_1}\sum_{i:g_i=1} x_i, \hat{\mu}_2 = \frac{1}{N_2}\sum_{i:g_i=2} x_i$$

$$\sum_{i=1}^{N} x_i = N_1\hat{\mu}_1 + N_2\hat{\mu}_2, \sum_{i=1}^{N} x_i^T = N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T$$

$$\sum_{i=1}^{N} y_i = N_1 t_1 + N_2 t_2, \sum_{i=1}^{N} y_i x_i = t_1 N_1\hat{\mu}_1 + t_2 N_2\hat{\mu}_2$$

and

$$\hat{\Sigma} = \frac{1}{N-2}\left(\sum_{i:g_i=1}(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^T + \sum_{i:g_i=2}(x_i-\hat{\mu}_2)(x_i-\hat{\mu}_2)^T\right)$$

The left side of this equation is equal to :

$$\begin{aligned}
&\sum_{i=1}^{N} x_i x_i^T - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i^T\right) \\
&= \sum_{i=1}^{N} x_i x_i^T - \frac{1}{N}\Big(N_1\hat{\mu}_1 + N_2\hat{\mu}_2\Big)\Big(N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T\Big) \\
&= (N-2)\hat{\Sigma} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T - \frac{1}{N}\Big(N_1\hat{\mu}_1 + N_2\hat{\mu}_2\Big)\Big(N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T\Big) \\
&= (N-2)\hat{\Sigma} + \frac{N_1 N_2}{N}(\hat{\mu}_2-\hat{\mu}_1)(\hat{\mu}_2-\hat{\mu}_1)^T \\
&= (N-2)\hat{\Sigma} + N\hat{\Sigma}_B,
\end{aligned}$$

and the right side is equal to :

$$
\begin{aligned}
&\sum_{i=1}^{N} y_i x_i - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i\right) \\
&= t_1 N_1 \hat{\mu}_1 + t_2 N_2 \hat{\mu}_2 - \frac{1}{N}\Big(N_1 t_1 + N_2 t_2\Big)\Big(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2\Big) \\
&= \frac{N_1 N_2}{N}(t_2 - t_1)(\hat{\mu}_2 - \hat{\mu}_1).
\end{aligned}
$$

And we can notice that $t_2 - t_1 = \frac{N^2}{N_1 N_2}$

So we finish the proof.

c

(c) Hence show that $\hat{\mathbf{\Sigma}}_B \beta$ is in the direction $(\hat{\mu}_2 - \hat{\mu}_1)$ and thus

$$\hat{\beta} \propto \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{4.57}$$

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

We can see that:

$$\hat{\Sigma}_B \beta = \frac{N_1 N_2}{N^2}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T \beta$$

and $\frac{N_1 N_2}{N^2}$ and $(\hat{\mu}_2 - \hat{\mu}_1)^T \beta$ is both constant. so :

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

d

(d) Show that this result holds for any (distinct) coding of the two classes.

We have use the coding of two classes after derive equation

So all that is needed is for $(t_2 - t_1)$ to be a constant to derive the same conclusion

e

(e) Find the solution $\hat{\beta}_0$ (up to the same scalar multiple as in (c), and hence the predicted value $\hat{f}(x) = \hat{\beta}_0 + x^T\hat{\beta}$. Consider the following rule: classify to class 2 if $\hat{f}(x) > 0$ and class 1 otherwise. Show this is not the same as the LDA rule unless the classes have equal numbers of observations.

Ext 4.3

Ex. 4.3 Suppose we transform the original predictors $\mathbf{X}$ to $\hat{\mathbf{Y}}$ via linear regression. In detail, let $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}$, where $\mathbf{Y}$ is the indicator response matrix. Similarly for any input $x \in \mathbb{R}^p$, we get a transformed vector $\hat{y} = \hat{\mathbf{B}}^T x \in \mathbb{R}^K$. Show that LDA using $\hat{\mathbf{Y}}$ is identical to LDA in the original space.

Assuming the encoding of $-N/N_1$ and N/N_2, by (b) we have

$$\begin{aligned}\hat{\beta}0 &= \frac{1}{N}\left(\sum i = 1^N y_i - \left(\sum_{i=1}^{N} x_i^T\right)\beta\right) \\ &= -\frac{1}{N}\left(\sum_{i=1}^{N} x_i^T\right)\hat{\beta} \\ &= -\frac{1}{N}\left(N_1\hat{\mu}_1^T + N_2\hat{\mu}_2^T\right)\hat{\beta}\end{aligned}$$

so that

$$\hat{f}(x) = \hat{\beta}_0 + x^T \hat{\beta} = \left[x^T - \frac{1}{N} \left(N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T \right) \right] \hat{\beta}$$

Since $\hat{\beta} \propto \hat{\Sigma}^{-1} \left(\hat{\mu}_2 - \hat{\mu}_1 \right)$, there exists $\lambda > 0$ (up to a scalar constant, i.e., we can flip the classification sign if $\lambda < 0$) such that $\hat{\beta} = \lambda \hat{\Sigma}^{-1} \left(\hat{\mu}_2 - \hat{\mu}_1 \right)$. Therefore, $\hat{f}(x) > 0$ is equivalent to

$$\left[x^T - \frac{1}{N} \left(N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T \right) \right] \hat{\Sigma}^{-1} \left(\hat{\mu}_2 - \hat{\mu}_1 \right) > 0,$$

which is equivalent to LDA rule (1) when $N_1 = N_2$. When $N_1 \neq N_2, \log \left(N_2 / N_1 \right) \neq 0$ in (1) so they are not equivalent.