باسمه تعالی

گزارش کار تمرین سوم درس یادگیری ماشین – جناب آقای دکتر باباعلی

ایمان کیانیان

۱. مقدمه

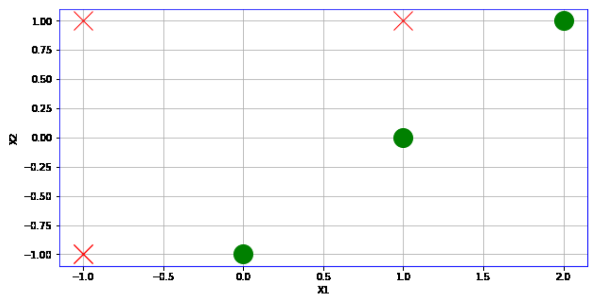
در این تمرین به بررسی LDA میپردازیم. در سوال اول باید با داده هایی که به ما داده شده است ، یک classifier با استفاده از خط LDA بدست آوریم. در قسمت دوم میخواهیم تاثیر feature extraction بر روی دقت کلاس بندی را بررسی کنیم. در ادامه توضیحاتی در خصوص این تمرین و نتایج بدست آمده ارائه میشود.

۲. سوال اول

در این سوال داده های زیر به ما داده شده است . دو کلاس C0 و C1 داریم. داده های مربوط به هر کلاس به صورت زیر است:

$$C_0 = \left\{ \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix} \right\}$$
$$C_1 = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\}$$

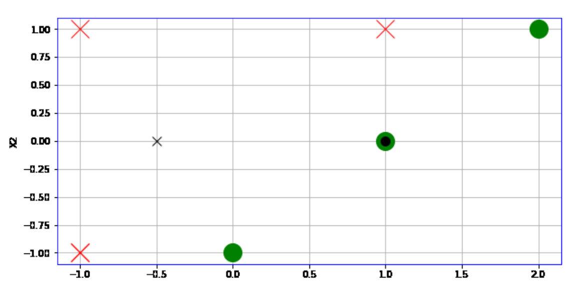
کلاس اول شامل ۳ داده و کلاس دوم شامل ۴ داده است. ضربدر های قرمز داده های کلاس C1 و دایره های سبز داده های کلاس C0 هستند.



میانگین داده های هر کلاس را به سادگی میتوانیم محاسبه کنیم.

$$\mu_1 = \frac{1}{3} \times \begin{pmatrix} 0+1+2 \\ -1+0+1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$
$$\mu_2 = \frac{1}{4} \times \begin{pmatrix} 1-1-1-1 \\ 1+1-1-1 \end{pmatrix} = \begin{pmatrix} -0.5 \\ 0 \end{pmatrix}$$

حال با استفاده از matplotlib داده ها و میانگین را رسم میکنیم. میانگین با نقطه های مشکی نشان داده شده است:



در مرحله بعدی به سادگی میتوانیم S_1 و S_2 را محاسبه کنیم. به شکل زیر این ماتریس هارا محاسبه میکنیم:

C_1

1) $\begin{pmatrix} 0 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ $\quad t_0 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \begin{pmatrix} -1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

2) $\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ $\quad t_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$

3) $\begin{pmatrix} 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\quad t_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$

$$S_1 = t_0 + t_1 + t_2 = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$$

S_2

1) $\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} -0{,}5 \\ 0 \end{pmatrix} = \begin{pmatrix} 1{,}5 \\ 1 \end{pmatrix} \Rightarrow t_0 = \begin{pmatrix} 1{,}5 \\ 1 \end{pmatrix} \begin{pmatrix} 1{,}5 & 1 \end{pmatrix} = \begin{pmatrix} 2{,}25 & 1{,}5 \\ 1{,}5 & 1 \end{pmatrix}$

2) $\begin{pmatrix} -1 \\ 1 \end{pmatrix} - \begin{pmatrix} -0{,}5 \\ 0 \end{pmatrix} = \begin{pmatrix} -0{,}5 \\ 1 \end{pmatrix} \Rightarrow t_1 = \begin{pmatrix} -0{,}5 \\ 1 \end{pmatrix} \begin{pmatrix} -0{,}5 & 1 \end{pmatrix} = \begin{pmatrix} 0{,}25 & -0{,}5 \\ -0{,}5 & 1 \end{pmatrix}$

3) $\begin{pmatrix} -1 \\ -1 \end{pmatrix} - \begin{pmatrix} -0{,}5 \\ 0 \end{pmatrix} = \begin{pmatrix} -0{,}5 \\ -1 \end{pmatrix} \Rightarrow t_2 = \begin{pmatrix} -0{,}5 \\ -1 \end{pmatrix} \begin{pmatrix} -0{,}5 & -1 \end{pmatrix} = \begin{pmatrix} 0{,}25 & 0{,}5 \\ 0{,}5 & 1 \end{pmatrix}$

4) - - - ——— $\longrightarrow t_3 = \begin{pmatrix} 0{,}25 & 0{,}5 \\ 0{,}5 & 1 \end{pmatrix}$

$$\Longrightarrow S_2 = t_0 + t_1 + t_2 + t_3 = \begin{pmatrix} 3 & 2 \\ 2 & 2 \end{pmatrix}$$

سپس S_w را با استفاده از مجموع S_1 و S_2 محاسبه میکنیم.

$$S_w = \begin{bmatrix} 5 & 4 \\ 4 & 6 \end{bmatrix}$$

و سپس محاسبه ی S_b :

$$S_b = \begin{bmatrix} 2/25 & 0 \\ 0 & 0 \end{bmatrix}$$

به راحتی میتوان W را حساب کرد که برابر مقدار زیر است :

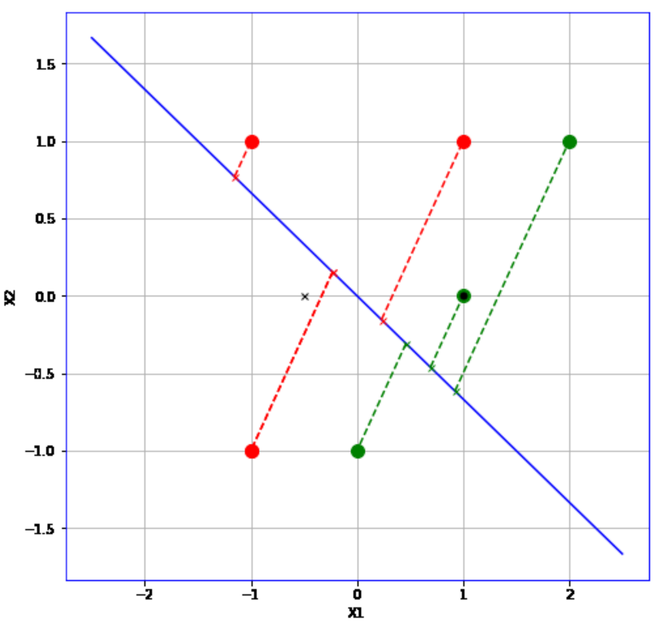
$$W = \begin{bmatrix} 0.64285714 \\ -0.42857143 \end{bmatrix}$$

حال با استفاده از رابطه زیر میتوانیم خط مربوط به LDA را رسم کنیم:

$$y = mx + b \xRightarrow{b=0} y = mx$$

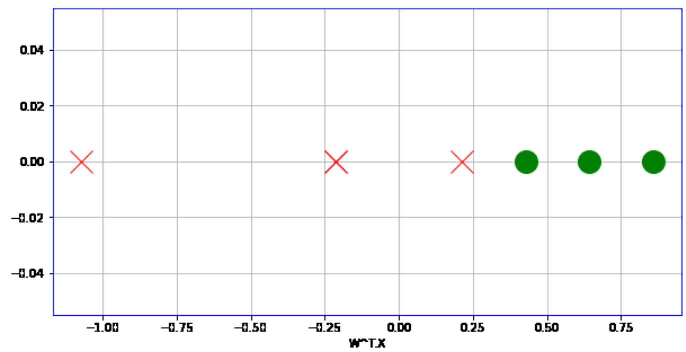
$$m = \frac{W[1]}{W[0]} \Longrightarrow y = \frac{-0.42857143}{0.64285714} x$$

همینطور به سادگی میتوانیم داده ها را به روی خط LDA تصویر کنیم:



در شکل مشخص شده است که داده های دو کلاس به طور خیلی خوبی در راستای LDA از هم جدا شده اند.

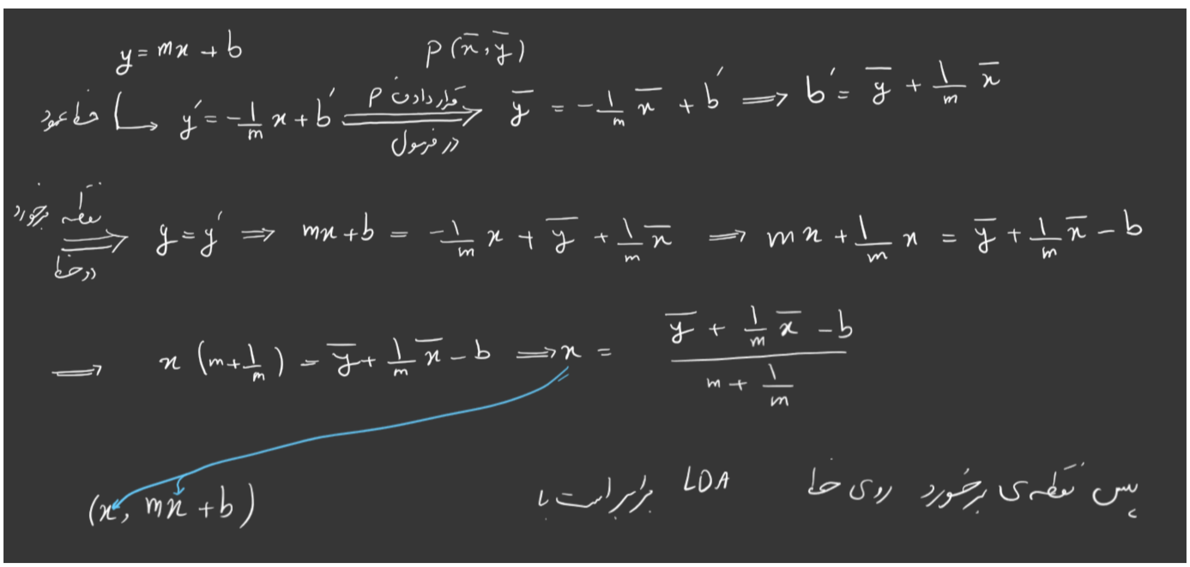
به طور واضح تر میتوان دید:



به سادگی با تعریف یک threshold مثل ۰,۲۵ میتوانیم کلاس بندی را انجام دهیم:

$$y = \begin{cases} 0 & if\ W^T X > 0.25 \\ 1 & if\ W^T X \le 0.25 \end{cases}$$

برای پیدا کردن نقطه ای از روی خط مربوط به LDA که تصویر نقاط است از فرمول خط عمود استفاده کردیم.



۳. سوال دوم

در این سوال میخواهیم تاثیر عمل feature extraction با استفاده از LDA را روی دقت دسته بندی چه تاثیری می گذارد. انتظار ما این است چون یک فیچر را از بین میبریم پس قطعا یک سری اطلاعات را از دست میدهیم . بنابراین دقت کمی کاهش می یابد. در ادامه به بررسی این مسئله میپردازیم.

ابتدا دقت را روی دسته بندی پرسپترون بدون اعمال LDA برای داده های آموزشی و تست گرفتیم که به شکل زیر است:

N = 4000	Predicted : 0	Predicted : 1	
Actual : 0	TN = 1833	FP = 174	**2007**
Actual : 1	FN = 149	TP = 1844	**1993**
	1982	***2018***	

N = 1000	Predicted : 0	Predicted : 1	
Actual : 0	TN = 468	FP = 35	**503**
Actual : 1	FN = 40	TP = 457	**497**
	508	***492***	

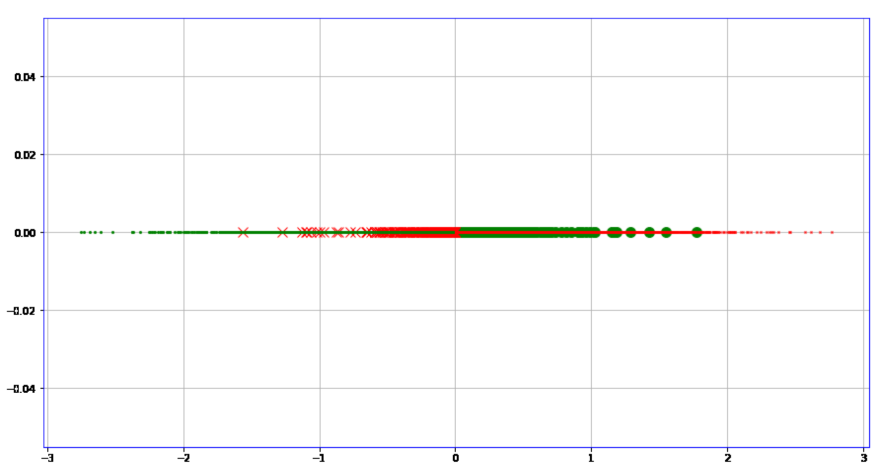
بنابراین دقت برای داده های اموزشی برابر ۰/۹۱۹۲۵ و برای داده های تست برابر ۰/۹۲۵ است (همانطور که در گزارش قبل هم ارائه شد).

حال با استفاده از فرمول داده شده برای استاندارد سازی داده ها ، داده های خومان را استاندارد سازی کردیم:

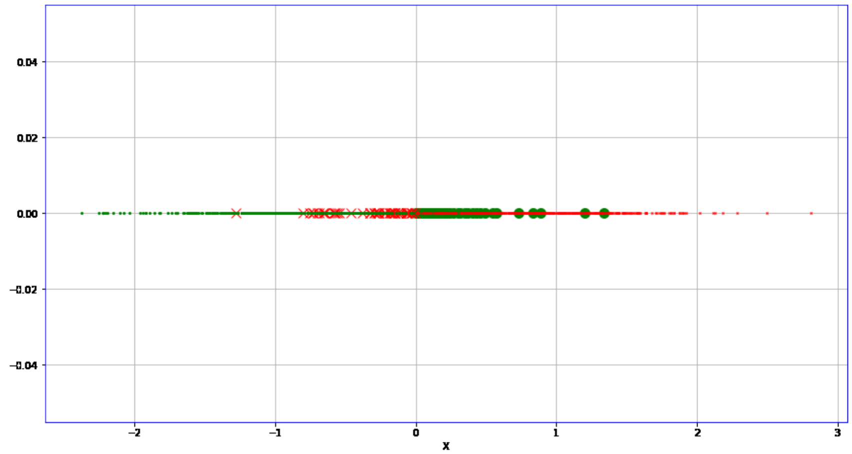
$$X_{standard} = \frac{X - \mu(X)}{\sqrt{Var(X)}}$$

سپس همانند قسمت اول ، LDA را اعمال میکنیم با این تفاوت که در پایان مقادیر ویژه را بدست می آوریم. بردار ویژه متناظر با مقدار ویژه بزرگتر را به عنوان W انتخاب میکنیم. حال تمام داده ها را در آن ضرب میکنیم و داده های جدید که شامل یک feature هستند را تشکیل میدهیم. سپس با استفاده از این تک feature با استفاده از پرسپترون دسته بندی را انجام میدهیم.

برای داده های آموزشی دقت پس از دسته بندی برابر ۰,۸۹۸۲۵ است و نمایش داده ها بصورت زیر است:



داده های misclassify شده به وضوع مشخص است. دایره های سبز بزرگ داده های misclassify شده است که در کلاس اول است ولی مدل ما داده ها را کلاس دو شناسایی میکند چون از ۰ بزرگتر است.
همچنین داده های قرمز (ضربدر های قرمز) بزرگ ، که در سمت چپ عکس میبینید misclassify شده اند چون داده ها کلاس واقعی دوم هستند ولی کلاس اول پیش بینی شده اند که غلط است.
تعداد داده های misclassify شده برای داده های آموزشی ۴۰۷ عدد از ۴۰۰۰ عدد است.
دقت در داده های تست بعد از اعمال LDA برابر ۰,۹۰۵ و تعداد داده های misclassify شده ۹۵ از ۱۰۰۰ تا داده است و به شکل زیر قابل نمایش است:

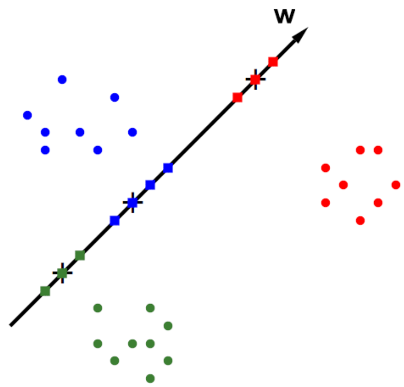


همانند شکل قبلی ضربدر های قرمز بزرگ (سمت چپ) داده های misclassify شده از کلاس دوم و سبز های بزرگ در سمت راست داده های misclassify شده از کلاس اول هستند.
آنهایی که کوچک هستند درست classify شده اند.

	Perceptron without LDA and Standardization	**Perceptron after applying LDA and Standardization**
Accuracy on Training Data	0.91925	0.89825
Accuracy on Test Data	0.925	0.905

که واضح است دقت بعد از اعمال LDA کمی کمتر شده است. این نتیجه مورد انتظار ما بود . اگر داده های ما نویز کمی داشته باشند دقت بعد از کم کردن feature ها کاسته خواهد شد که در این مورد چنین اتفاقی رخ میدهد و نتیجه کاملا واضح است.

۴. سوال سوم



بدست آوردن میانه های هر کلاس کار سختی نیست. کافی است یک ماتریس M به تعداد کلاس ها ستون و به تعداد فیچر ها (ورودی) سطر داشته باشیم. هر ستون از این ماتریس نمایش میانه است. مثلا ستون i ام μ_i میباشد.

$$M = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_c]$$

که c تعداد کلاس های ما هستند. هر μ_i یک بردار n تایی است. n تعداد فیچر های داده های ماست.

برای بدست آوردن S_i بصورت زیر عمل میکنیم:

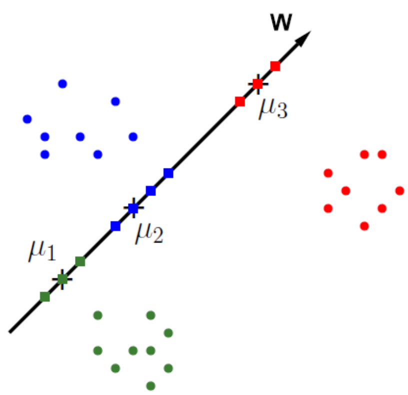
$$S_i = \sum_{x \in C_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T \quad \text{for} \quad i = 1 \; to \; c$$

حال هر کدام از S_i ها را داریم. به راحتی میتوانیم S_W را محاسبه کنیم:

$$S_W = \sum_{i=1}^{c} S_i$$

میخواهیم فاصله یا اختلاف بین میانه هارا بیشینه کنیم :

$$\sum_{i=1}^{c} (\mu_i - \bar{\mu})^2 = \frac{1}{c} \sum_{i \le \ell} (\mu_i - \mu_\ell)^2 , \quad \text{where } \bar{\mu} = \frac{1}{c} \sum_{i=1}^{c} \mu_i$$



برای محاسبه S_b داریم:

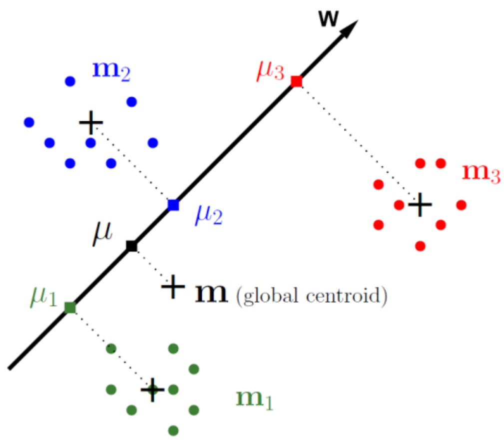
$$S_b = \sum_{i=1}^{c} n_i (\mu_i - \mu)^2 , \qquad \text{where} \quad \mu = \frac{1}{n} \sum_{i=1}^{c} n_i \mu_i$$

چون میانگین وزن دار (μ) ، تصویر مرکز گلوبال (m) داده آموزشی بر روی w است پس:

$$w^T m = w^T \left(\frac{1}{n} \sum_{i=1}^{n} x_i \quad \right) = w^T \left(\frac{1}{n} \sum_{j=1}^{c} n_j m_j \right) = \frac{1}{n} \sum_{j=1}^{c} n_j \mu_j$$

در مقابل، میانگین ساده چنین تفسیر هندسی ای ندارد:

$$\bar{\mu} = \frac{1}{c} \sum_{j=1}^{c} \mu_j = \frac{1}{c} \sum_{j=1}^{c} w^T m_j = w^T \left(\frac{1}{c} \sum_{j=1}^{c} m_j \right)$$



میتوانیم پراکندگی بین کلاس ها را (در فضای w) به صورت زیر ساده کنیم:

$$\sum_{j=1}^{c} n_j (\mu_j - \mu)^2 = \sum n_j \left(w^T (m_j - m\ ) \right)^2$$

$$= \sum n_j w^T (m_j - m)(m_j - m)^T w$$

$$= w^T \left(\sum n_j (m_j - m)(m_j - m)^T \right) w$$

$$= w^T S_b w.$$

تابع هدف ما به صورت زیر بود:

$$J(w) = \frac{\sum n_j (\mu_j - \mu)^2}{\sum s_j^2}$$

حال ما به مقدار پایین میرسیم :

$$J(w) = \frac{w^T S_b w}{w^T S_\omega w}$$

هدف ما ماکزیمم کردن مقدار بالا است یعنی باید فاصله بیرون کلاسی بیشترین و فاصله درون کلاسی کمترین باشد .

$$\max_{w:\|w\|=1} J(w) = \max_{w:\|w\|=1} \frac{w^T S_b w}{w^T s_w w}$$

محاسبه ی مشتق با حالت دو کلاسه تفاوتی ندارد :

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\frac{\partial \mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\partial \mathbf{w}} \times \mathbf{w}^T \mathbf{S}_W \mathbf{w} - \frac{\partial \mathbf{w}^T \mathbf{S}_W \mathbf{w}}{\partial \mathbf{w}} \times \mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\left(\mathbf{w}^T \mathbf{S}_W \mathbf{w}\right)^2} = \frac{(2\mathbf{S}_B \mathbf{w})\mathbf{w}^T \mathbf{S}_W \mathbf{w} - (2\mathbf{S}_W \mathbf{w})\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\left(\mathbf{w}^T \mathbf{S}_W \mathbf{w}\right)^2}$$

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$$

همانند حالت دو کلاسه داریم :

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w} \xrightarrow{\text{If } \boldsymbol{S}_W \text{ is full-rank}} \boldsymbol{S}_W^{-1} \boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{w}$$

حال چون در حالت دو کلاسه w با w حال حاضر فرق دارد پس باید :

$$\lambda w = S_\omega^{-1} S_b w = s_\omega^{-1} \sum_{j=1}^{c} n_j (m_j - m)(m_j - m)^T w$$

بطوریکه $(m_j - m)^T w$ یک عدد اسکالر است.