Análise do Discriminante Linear de Fisher, Fisher's LDA

Murilo Henrique Gomes - nºUSP: 10289015
Giuliano Pantarotto Semente - n° USP: 10288949
Murilo Krebsky - nº USP: 11297847
SME0243 - Teoria Espectral de Matrizes
Universidade de São Paulo

13 de Julho, 2021

Roteiro da Apresentação.

Introdução:

Entre os objetivos do curso, está o de aprender o método de análise de componentes principais (PCA), e desenvolvê-lo em aplicações familiares ao contexto da matemática aplicada.

Tão importante quando aprender um método é também aprender suas limitações. No presente caso, analisaremos os resultados obtidos pelo PCA em um problema de classificação no qual as direções de maior variabilidade não são suficientes para distinguir a classe a qual um dado elemento da amostra pertence. Tais resultados serão então comparados com aqueles obtidos por intermédio da análise do discriminante Linear de Fisher.

Problema Inicial:

Para ilustrarmos a situação problema, iremos assumir uma amostra de n pontos no $\mathbb{R}^2$ particionada em duas classes distintas $\mathcal{C}_1$ e $\mathcal{C}_2$. Isto é, teremos a amostra $\mathbf{X} = \{x_1, \ldots, x_n\}$ cujos elementos $x_i \in \mathbb{R}^2$ e $\forall x_i \in \mathbf{X} \mid x_i \in \mathcal{C}_1$ ou $x_i \in \mathcal{C}_2$, com $\mathcal{C}_1 \bigcap \mathcal{C}_2 = \varnothing$.

Sejam as classes $\mathcal{C}_1 = \{(1,2),(2,3),(3,3),(4,5),(5,5)\}$ e $\mathcal{C}_2 = \{(1,0),(2,1),(3,1),(3,2),(5,3),(6,5)\}$ a dispersão dos dados está ilustrada a seguir

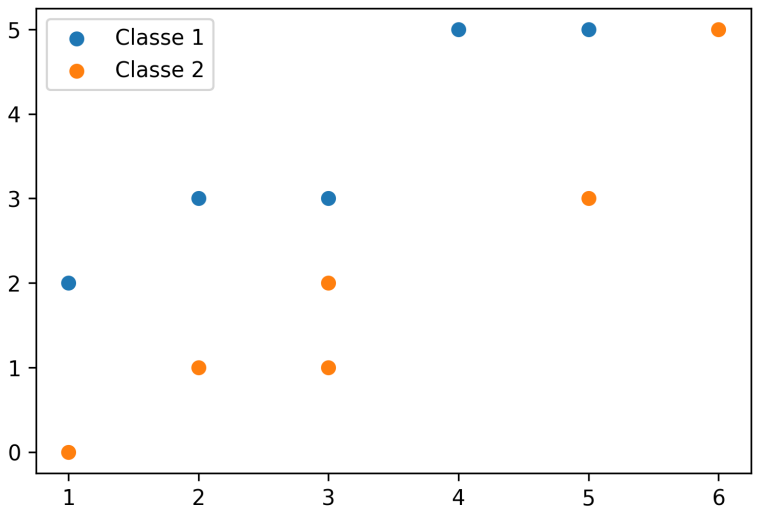

Figure 1: Representação das classes no plano cartesiano

Projetando os dados na direção de maior variabilidade fornecida pela análise de componentes principais sob os dados de $\mathbf{X}$, teremos a seguinte representação:

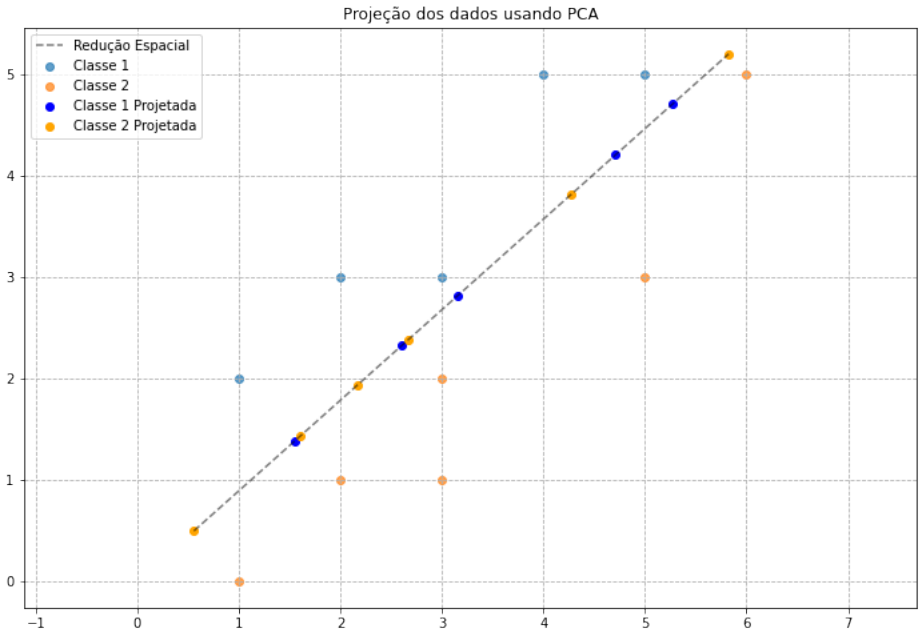

Figure 2: Projeção dos dados na componente com maior variabilidade

Note que o método *PCA* projeta alguns elementos de classes distintas sob pontos bem próximos, como por exemplo o ponto $(1, 2) \in \mathcal{C}_1$ e $(2, 1) \in \mathcal{C}_2$. Consequentemente, vemos que a direção de maior variância não será útil para o problema de classificação.

Discriminante Linear de Fisher:

O objetivo é encontrar a direção em que devemos projetar os dados de modo que os elementos de classes distintas estejam melhor separados. Vamos inicialmente apresentar o procedimento para o caso mais simples em que temos apenas 2 classes e os dados x_i estão em $\mathbb{R}^2$. Obs.: no que se segue, todos os vetores unidimensionais são vetores coluna.

Vamos então supor um $w \in \mathbb{R}^2$ tal que a direção dada por esse vetor seja a qual os dados estarão melhor projetados. Em efeito, se escolhermos w normalizado, a projeção dos dados sob o subespaço gerado por w será dada por $w^T \cdot x \in [w]$, $\forall x \in X$. Note que as projeções serão números reais.

Se buscamos a projeção na qual classes distintas estarão melhor separadas, então é de se esperar que a projeção da média μ_1 associada a classe $\mathcal{C}_1$ esteja o mais distante possível da média μ_2 da classe C_2. Isto é, sejam $\tilde{\mu}_1$ e $\tilde{\mu}_2$ as projeções dessas médias, dadas por:

$$\tilde{\mu}_1 = \frac{1}{n_1} \sum_{x_i \in \mathcal{C}_1}^{n_1} w^T \cdot x_i = w^T \cdot \mu_1 \tag{1}$$

$$\tilde{\mu}_2 = \frac{1}{n_2} \sum_{x_j \in \mathcal{C}_2}^{n_2} w^T \cdot x_j = w^T \cdot \mu_2 \tag{2}$$

onde n_1 e n_2 são os números de representantes das classes $\mathcal{C}_1$ e $\mathcal{C}_2$, respectivamente.

Num primeiro momento, podemos desejar que a magnitude da diferença entre a projeção das médias ($|\tilde{\mu}_1 - \tilde{\mu}_2|$) seja a maior possível (equivalentemente podemos tomar $(\tilde{\mu}_1 - \tilde{\mu}_2)^2$).

Entretanto, uma grande distância entre as posições médias de cada classe não nos garante por si só que não existirão representantes de classes distintas próximos um do outro. Pode ser que a dispersão dos dados em torno das médias de cada classe seja suficientemente grande para que existam representantes de uma classe perto da média de outra classe, fazendo com que continue havendo uma sobreposição expressiva entre os dados projetados.

Dessa forma, vemos que é importante ter uma medida para o espalhamento dos representantes de uma determinada classe e então buscar minimizar essa quantidade no espaço projetado. Uma abordagem é definir a dispersão de uma amostra como a soma dos quadrados das distâncias entre um elemento da amostra à sua respectiva média. Para elementos da classe C_i teremos então

$$s_i = \sum_{x \in \mathcal{C}_i}^{n_i} \|x - \mu_i\|^2 = \sum_{x \in \mathcal{C}_i}^{n_i} (x - \mu_i)^T \cdot (x - \mu_i), \tag{3}$$

onde novamente n_i é o nº de elementos na classe C_i.

Veja que (3) diz respeito a dispersão dos dados em seu espaço de atributos original. Entretanto, assim como a média de uma classe, estamos interessados na dispersão associada a projeção dos dados **no subespaço gerado por** w. Seja então $y_i = w^T \cdot x_i \in \mathbb{R}$; a dispersão dos elementos de uma classe (a 1 por exemplo) nesse subespaço será

$$\tilde{s}_1 = \sum_{y_i \in \tilde{\mathcal{C}}_1}^{n_1} (y_i - \tilde{\mu}_1)^2 = \sum_{y_i \in \tilde{\mathcal{C}}_1}^{n_1} (w^T \cdot x_i - w^T \cdot \mu_1)^2, \tag{4}$$

onde usamos que $\tilde{\mu_1} = w^T \cdot \mu_1$. Desenvolvendo a expressão anterior, encontramos

$$\tilde{s}_1 = \sum_{x_i \in \mathcal{C}_1}^{n_1} = (y_i - \tilde{\mu}_1)^2 = \sum_{y_i \in \tilde{\mathcal{C}}_1}^{n_1} (w^T \cdot x_i - w^T \cdot \mu_1)^2 \tag{5}$$

$$[w^T \cdot (x_i - \mu_1)] \cdot [w^T \cdot (x_i - \mu_1)]^T = w^T \cdot (x_i - \mu_1) \cdot (x_i - \mu_1)^T \cdot w \tag{6}$$

$$\Rightarrow \tilde{s}_1 = w^T \cdot \left(\sum_{x_i \in \mathcal{C}_1}^{n_1} (x_i - \mu_1) \cdot (x_i - \mu_1)^T \right) \cdot w = w^T \cdot S_{w_1} \cdot w \tag{7}$$

onde S_{w_1} é uma matriz (note que temos o produto exterior entre um vetor coluna e um vetor linha) $m \times m$ (m é a dimensão do espaço dos dados) que representa a dispersão dos elementos da classe C_1, sendo proporcional a matriz de covariância desta amostra.

Como já apontamos, é desejável que a dispersão de cada classe seja pequena. Uma forma de considerar isso simultaneamente com o afastamento das médias é maximizar uma fração, colocando a dispersão das classes no denominador. Por exemplo, no presente caso de 2 classes, é definido o quociente

$$J(w) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1 + \tilde{s}_2}, \tag{8}$$

chamado de discriminante linear de Fisher. Veja que maximizando esta grandeza garantimos que:

1. $(\tilde{\mu}_1 - \tilde{\mu}_2)$ é grande: as médias estarão afastadas;
2. $\tilde{s}_1$: os membros de $\mathcal{C}_1$ estejam agrupados em torno de $\tilde{\mu}_1$, com a menor dispersão possível;
3. $\tilde{s}_2$: os membros de $\mathcal{C}_2$ estejam agrupados em torno de $\tilde{\mu}_2$, com a menor dispersão possível;

Note que apesar de não aparecer explicitamente em (8), o discriminante de Fisher de fato depende de w. Vejamos como expressá-lo explicitamente em função dele. Vejamos primeiramente o numerador:

$$\begin{aligned}(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = [w^T \cdot \mu_1 - w^T \cdot \mu_2]^2 = [w^T \cdot (\mu_1 - \mu_2)]^2 = [w^T \cdot (\mu_1 - \mu_2)][w^T \cdot (\mu_1 - \mu_2)]^T = \\ [w^T \cdot (\mu_1 - \mu_2)][(\mu_1 - \mu_2)^T \cdot w] = w^T \cdot [(\mu_1 - \mu_2) \cdot (\mu_1 - \mu_2)^T] \cdot w = w^T \cdot S_B \cdot w\end{aligned} \tag{9}$$

onde S_B é uma matriz que indica neste caso o quanto as médias estão separadas. Ela recebe o nome de *Scatter Matrix Between Classes*.

Agora voltemos a atenção para o denominador. Note que, pelo mesmo desenvolvimento que fizemos para $\tilde{s}_1$, vale que $\tilde{s}_2 = w^T S_{w_2} w$. Assim,

$$\tilde{s}_1 + \tilde{s}_2 = w^T S_{w_1} w + w^T S_{w_2} w = w^T S_W w \tag{10}$$

onde $S_W = S_{w_1} + S_{w_2}$ recebe o nome de *Scatter Within Classes*. Dessa forma, o discriminante (8) pode ser escrito como

$$J(w) = \frac{w^T \cdot S_B \cdot w}{w^T \cdot S_W \cdot w} \tag{11}$$

Se encontrarmos o vetor w que torna $J(w)$ o maior possível, então garantimos uma boa separabilidade entre as classes (note que apesar da notação vetorial $J(w)$ é uma função escalar). Um aspecto muito importante a ser notado é que apenas a direção de w importa no valor de (11), devido a presença de w^T e w tanto no numerador como no denominador. Ou seja, $J(w) = J(\alpha w)$ para qualquer escalar $\alpha \neq 0$. Dessa forma, é suficiente procurar w no conjunto $A = \{w \in \mathbb{R}^m \,|w^T \cdot S_W \cdot w = 1\}$, visto que a condição $w^T \cdot S_W \cdot w = 1$ restringe apenas o tamanho dos vetores candidatos w e não sua direção. Portanto, basta resolvermos o problema

$$\begin{aligned}&\text{Maximizar } w^T \cdot S_B \cdot w \\ &\text{sujeito a } w^T \cdot S_W \cdot w = 1\end{aligned} \tag{12}$$

.

O formato do problema nos lembra o quociente de Rayleigh. Na verdade, podemos reescrevê-lo dessa forma sob certas hipóteses. Primeiramente, como a matriz S_W é simétrica, o teorema espectral garante que podemos decompô-la na forma $S_W = UDU^T$, onde D é uma matriz diagonal e $U \cdot U^T = I$. Além disso, como ela é semi-definida positiva, todos os autovalores (presentes na diagonal de D) serão não-negativos e terão raiz quadrada. Então, vemos que $S_W^{\frac{1}{2}}$ existe neste caso ($S_W^{\frac{1}{2}} = UD^{\frac{1}{2}}U^T$). Note também que $S_W^{\frac{1}{2}}$ é simétrica. Então, a restrição pode ser reescrita como

$$w^T \cdot S_W \cdot w = w^T \cdot S_W^{\frac{1}{2}} \cdot S_W^{\frac{1}{2}} \cdot w = (S_W^{\frac{1}{2}} \cdot w)^T \cdot S_W^{\frac{1}{2}} \cdot w = 1.$$

Assim, isso sugere a transformação $x = S_W^{\frac{1}{2}} w$. Se S_W for invertível (todos os autovalores positivos), então $S_W^{\frac{1}{2}}$ também será, e então $w = S_W^{-\frac{1}{2}} \cdot x$ (com $S_W^{-\frac{1}{2}}$ também simétrica). Logo, o problema se resumirá a

$$\begin{aligned}&\text{Maximizar } (S_W^{-\frac{1}{2}} \cdot x)^T \cdot S_B \cdot S_W^{-\frac{1}{2}} \cdot x = x^T \cdot (S_W^{-\frac{1}{2}} \cdot S_B \cdot S_W^{-\frac{1}{2}}) \cdot x \\ &\text{sujeito a } x^T \cdot x = 1\end{aligned} \tag{13}$$

.

Agora, como a matriz $B = S_W^{-\frac{1}{2}} \cdot S_B \cdot S_W^{-\frac{1}{2}}$ é simétrica e semi-definida positiva, ela possui uma decomposição espectral e autovalores não negativos. Pela teoria que vimos sobre o quociente de Rayleigh, sabemos que a

solução do problema (13) é o autovetor x_1 associado ao maior autovalor λ_1 da matriz B. Então, a solução para o problema original (12) é dado por

$$w_1 = S_W^{-\frac{1}{2}} \cdot x_1. \tag{14}$$

Pode ser útil a seguinte reformulação do problema. Vimos no parágrafo anterior que estamos interessados no problema de autovalores

$$(S_W^{-\frac{1}{2}} \cdot S_B \cdot S_W^{-\frac{1}{2}} \cdot x_1 = \lambda_1 x_1 \Rightarrow S_B \cdot S_W^{-\frac{1}{2}}) \cdot x_1 = \lambda_1 S_W^{-\frac{1}{2}} \cdot x_1 \Rightarrow S_B \cdot w_1 = \lambda_1 S_W \cdot w_1, \tag{15}$$

onde na última passagem usamos que $x_1 = S_W^{\frac{1}{2}} w_1$. A última igualdade é o que se chama de problema de autovalores generalizados, uma vez que em geral $S_W \neq I$.

No caso em que temos duas classes podemos obter uma expressão ainda mais simples para o w ótimo. A matriz S_B é dada simplesmente pelo produto exterior $(\mu_1 - \mu_2) \cdot (\mu_1 - \mu_2)^T$ (tendo posto 1 portanto), de modo que para um vetor v teremos

$$S_B \cdot v = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \cdot v = \beta(\mu_1 - \mu_2) \tag{16}$$

visto que o produto $(\mu_1 - \mu_2)^T \cdot v$ sempre dá um escalar β, sendo que $\beta = 0$ se v for ortogonal a $\mu_1 - \mu_2$. Em outras palavras, $S_B \cdot v$ sempre fornece um múltiplo de $\mu_1 - \mu_2$, em particular para o caso de w_1. Como $\lambda_1 > 0$ (a menos que S_B seja a matriz nula - o que não é um caso interessante)), (15) fornece

$$\beta(\mu_1 - \mu_2) = \lambda_1 S_W \cdot w_1 \Rightarrow w_1 = \frac{\beta}{\lambda_1} S_W^{-1} \cdot (\mu_1 - \mu_2). \tag{17}$$

Como estamos interessados apenas na direção de w, podemos tomar simplesmente

$$w_1 = S_W^{-1}(\mu_1 - \mu_2). \tag{18}$$

Também é interessante observar que podemos chegar no problema de autovalores generalizados (15) de outra forma. Ela consiste em procurar por pontos críticos da função (11) tomando a "derivada" na direção w (fornecerá um vetor) e igualando ao vetor nulo (condição necessária para termos pontos de máximo):

$$\frac{d}{d\,w} J(w) = \frac{\left(\frac{d}{d\,w}(w^T \cdot S_B \cdot w)\right) \cdot w^T \cdot S_W \cdot w - \left(\frac{d}{d\,w}(w^T \cdot S_W \cdot w)\right) \cdot w^T \cdot S_B \cdot w}{(w^T \cdot S_w \cdot w)^2} = 0 \tag{19}$$

$$\frac{d}{d\,w} J(w) = \frac{(2S_B \cdot w) \cdot w^T \cdot S_W \cdot w - (2S_W \cdot w) \cdot w^T \cdot S_B \cdot w}{(w^T \cdot S_W \cdot w)^2} = 0 \tag{20}$$

Solucionaremos a equação considerando a seguinte relação de proporcionalidade

$$\frac{d}{d\,w} J(w) \propto (S_B \cdot w) \cdot w^T \cdot S_W \cdot w - (S_W \cdot w) \cdot w^T \cdot S_B \cdot w = 0 \tag{21}$$

$$\frac{w^T \cdot S_W \cdot w \cdot (S_B \cdot w)}{w^T \cdot S_W \cdot w} - \frac{w^T \cdot S_B \cdot w \cdot (S_W \cdot w)}{w^T \cdot S_W \cdot w} = 0 \tag{22}$$

$$\text{Definindo } \lambda = \frac{w^T \cdot S_B \cdot w}{w^T \cdot S_w \cdot w} \Rightarrow S_B \cdot w = \lambda \cdot (S_W \cdot w). \tag{23}$$

Generalizações

A primeira generalização que podemos pensar em fazer é quando temos elementos de mais de 2 classes distintas, digamos n classes. A ideia continua a mesma: na direção que procuramos projetar os dados, as médias de cada classe devem estar bem separadas umas das outras e a dispersão dos elementos de cada classe deve ser pequena. Quanto ao último requisito, a definição de S_W (*Scatter Within Classes*) continua a mesma, com a única diferença que teremos que somar mais termos. O escalar $w^T S_{w_i} w$ irá representar a dispersão da i-ésima classe no espaço gerado por w, e queremos minimizar a soma $\sum_i w^T S_{w_i} w$, que pode ser reescrita de maneira mais compacta (aplicando propriedade distributiva) como $w^T S_W w$, onde $S_W = \sum_i S_{w_i}$. Em termos dos dados originais x_i, temos

$$S_W = \sum_{i=1}^{n} S_{w_i} = \sum_{i=1}^{n} \sum_{j \in C_i} (x_i - \mu_i)(x_i - \mu_i)^T, \tag{24}$$

onde μ_i é a média da i-ésima classe.

Agora vejamos o que precisamos mudar na definição de S_B. Quando tínhamos apenas duas classes, para garantir que as médias estariam bem separadas bastou considerar direta ou indiretamente a distância entre elas. Se temos mais de 2 classes, precisamos considerar as posições das n médias de cada classe ao mesmo tempo para evitar que uma fique próxima da outra. Uma medida que podemos considerar é, assim como fizemos para cada classe, a dispersão. Só que agora tal dispersão, que queremos maximizar, é das médias de cada classe (daí o nome *Scatter Matrix Between Classes*). Portanto,

$$S_B = \sum_{i=1}^{n} (\mu_i - \mu)(\mu_i - \mu)^T, \tag{25}$$

onde μ é a **média das médias de cada classe**. Note que, uma vez feitas essas generalizações nas definições de S_W e S_B, toda a dedução que fizemos anteriormente permanece válida, visto que não alteramos as propriedades daquelas matrizes (simetria e o fato de serem semi-definidas positivas) além de exigir que S_W tenha inversa. Obs.: existem outras definições possíveis para S_B. A utilizada pela biblioteca *Sklearn* por exemplo é

$$S_B = \sum_{i=1}^{n} N_i(\mu_i - \bar{x})(\mu_i - \bar{x})^T, \tag{26}$$

onde N_i é o número de elementos da i-ésima classe e $\bar{x}$ é a média geral dos dados, sem considerar o agrupamento entre classes. Essa definição surge quando se exige que $S_B + S_W = S_T$, onde S_T é a matriz de dispersão total dos dados:

$$S_T = \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T \tag{27}$$

Até agora, buscamos a melhor direção para se projetar os dados, de modo que eles são projetados num subespaço de dimensão 1. Entretanto, frequentemente apenas uma dimensão não é o suficiente para garantir uma boa separação dos dados (possibilitando uma classificação apropriada), ainda mais quando o espaço original dos dados possui dimensão alta. Assim, pode ser útil projetar em espaços com maior dimensão. O problema de autovalores a que chegamos pode admitir mais que um autovalor não nulo como solução, de modo que, assim como no PCA, podemos escolher os autovetores relacionados aos maiores k autovalores. Tais autovetores serão então as k melhores direções para se projetar os dados. Observamos que, como o posto máximo da matriz S_B é $n-1$ (por construção), o número de classes limita a dimensão k do subespaço em que intencionamos realizar a projeção dos dados.

Resultados Obtidos com LDA

Abaixo apresentamos uma imagem que mostra o resultado de aplicarmos o método de Fisher na amostra que usamos no começo deste relatório (figura (1)):

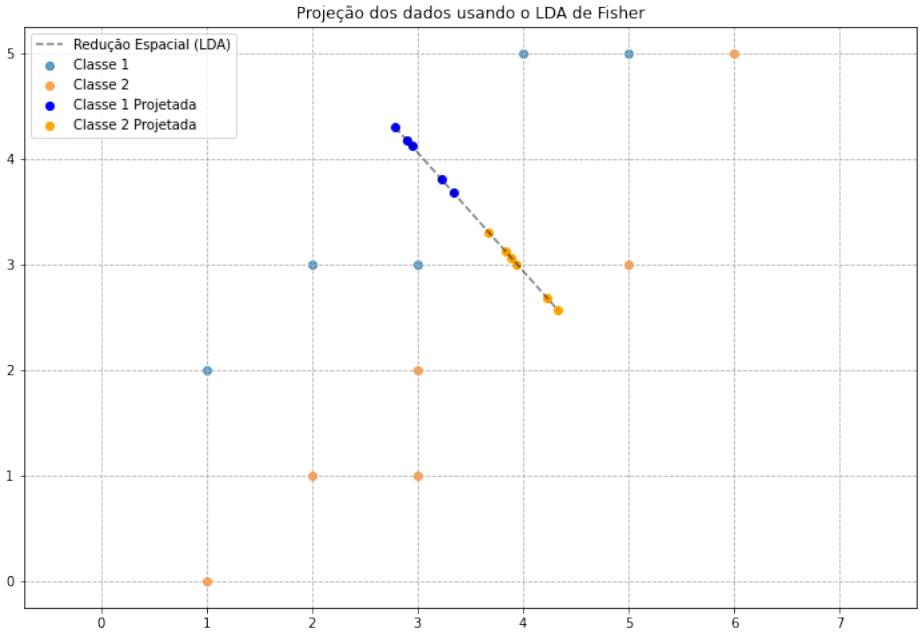

Figure 3: Projeção dos dados na direção fornecida pelo discriminante de Fisher

Como se pode ver, os elementos das classes foram projetados na direção indicada pela linha tracejada, fornecida pelo método LDA. Vemos que as classes estão separadas de uma maneira bem razoável, ao contrário do que observamos na projeção realizada no caso do PCA.