Relatório 03

ANÁLISE DISCRIMINANTE LINEAR
“Linear Discriminant Analysis - LDA”

&

Most Discriminant Features - MDF

Cristiano Lopes Moreira

Matrícula: 119103-0

Sumário

1. Introdução

Discriminar, classificar, arranjar, arrumar, dispor, ordenar, organizar, qualificar, denominar, dividir em grupos ou classes que possuam características parecidas, determinar a classe de alguma coisa dentro de determinado grupo ou conjunto; desenvolvia por Ronaldo A. Fisher (1936) a metodologia de análise discriminante linear (LDA) destina-se primariamente em segmentar amostras em grupos com características semelhantes. Se assemelha à técnica do PCA (principal compoment Analisys) pelo uso da rotação dos eixos de referência, porém, enquanto o PCA busca o eixo de maior variação dos dados, o LDA busca o eixo de maior distinção entre os dados de forma a maximizar a variação entre as classes.

2. Desenvolvimento teórico

O LDA é uma técnica da estatística multivariada que estuda a separação de objetos de uma população em duas ou mais classes, que utiliza de conceitos de estatística: variância, desvio padrão, covariância, autovetores e autovalores; tem por finalidade básica a análise dos dados pela escolha das formas mais representativas de dados a partir de combinações lineares das variáveis originais.

Por este método é possível identificar uma componente principal ou eixo que melhor representa a distinção entre os agrupamentos dos dados, (Linha pontilhada na Figura 1).

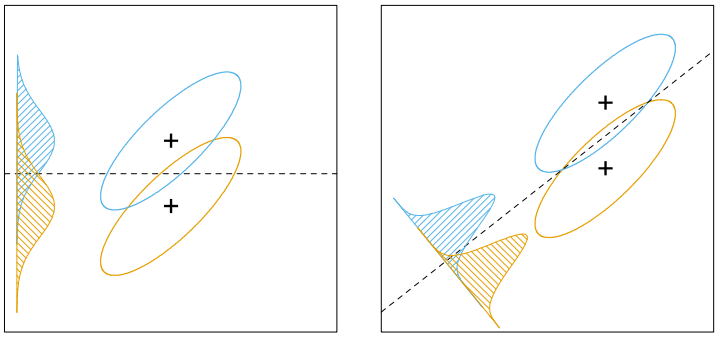

Fig1: Elements of Statistical Leaning (2nd Ed.) Hastile, Tibshirani & Friedman 2009 Cap4

2.1. Método para obtenção do LDA

6 etapas são necessárias para realizar a Análise Discrimina-te Linear - LDA.

1 Etapa: Obter conjunto de dados.

2 Etapa: Calcular os vetores de média d-dimensional para as diferentes classes do conjunto de dados.

3 Etapa: Calcular as matrizes de dispersão (matriz de dispersão intra classe e entre classe) e a matriz de projeção.

4 Etapa: Calcular os autovalores e autovetores.

5 Etapa: Classificar os autovetores diminuindo os autovalores e escolher “k” autovetores com os maiores autovalores para formar uma matriz dimensional$_{(d\times k)}$ (onde cada coluna representa um autovetor).

6 Etapa: Usar a matriz de autovetores para transformar as amostras no novo subespaço (rotacionando o eixo cartesiano).

O autovetor com o maior autovalor associado, corresponde à componente principal do conjunto de dados usados, essa componente é a mais significativa na dimensão dos dados.

2.1.1. Etapa 2

Média aritmética, ou simplesmente média, é a soma do total de valores de uma variável dividida pelo número total de observações

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

em que n é o número total de observações no conjunto de dados X_i, para i=1, ...n, representando cada um dos valores de x (FÁVERO et al., 2009).

Variância é a medida de dispersão dos dados em torno da média (FÁVERO et al., 2009).

Para população

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \quad (2)$$

Para amostras

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \quad (3)$$

Na segunda etapa ser calculado a médias das amostras e realizado um novo conjunto de dados normalizado pela média:

$$X_{\ i} = \frac{\sum_{i=1}^{n}(X_{i,j} - \bar{X})(X_{i,j} - \bar{X})^T}{N_i - 1} \quad (4)$$

2.1.2. Etapa 3

Matriz de dispersão intra-classes Definido por:

$$S_b = \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \quad (7)$$

Matriz de dispersão inter-classes Definido por:

$$S_w = \sum_{i=1}^{g}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x})(x_{i,j} - \bar{x})^T \quad (8)$$

Na etapa 3 deve ser calculada a **matriz de projeção** P_{lda} que maximiza a razão entre as matrizes de dispersão Sb e Sw.

$$\mathrm{P}_{lda} = argmax \frac{|P^T S_b P|}{|P^T S_w P|} \tag{9}$$

A matriz de projeção pode ser vista também com uma solução de autovalores e autovetores.

$$S_b \mathrm{P} - \mathrm{S}_w \mathrm{P}\Lambda = 0 \tag{10}$$

Multiplicando ambos os lados pela matriz inversa de S_w tempos:

$$\left(\mathrm{S}_w{}^{-1}\, \mathrm{S}_b\right) = \mathrm{P}\Lambda \tag{11}$$

2.1.3. Etapa 4

Autovalores, eigenvalues Λ, mostra a variância total explicada por cada dimensão, quanto maior o valor do eigenvalue maior é a variância explicada por sua dimensão (FÁVERO et al., 2009).

$$Det|(\Sigma - \Lambda I)\,| = 0 \tag{12}$$

Na etapa 4 deve ser calculado os autovalores, que são as raízes do determinante da diferença da matriz de covariância Σ e a matriz identidade multiplicada pelo eigenvalues.

Para calcular o determinante, um conjunto de equações lineares deve ser desenvolvido. Os casos de equações de segunda ordem podem ser calculados pela fórmula de Bhaskara

$$x = \frac{-b \pm \sqrt{\Delta}}{2a} \tag{13}$$

$$\Delta = b^2 - 4ac \quad (14)$$

Para equações polinomiais de ordens superiores deve ser utilizado a transformação de Jacobi, o método consiste em uma sequência de transformação de similaridade ortogonal da forma da equação

$$A^T A = A\,A^T = 1 \quad (15)$$

cada transformação (uma rotação jacobi) é apenas uma rotação plana projetada para aniquilar um dos elementos da matriz fora da diagonal. A transformação sucessiva até os elementos fora da diagonal ficarem cada vez menores, até a matriz diagonal ter a precisão desejada. A acumulação de produtos de transformações fornece a matriz de autovetores, enquanto os elementos da matriz diagonal final são os valores próprios (WILLIAM, 2007).

2.1.4. Etapa 5

O autovetor com o maior autovalor associado, corresponde à componente principal do conjunto de dados usados, essa componente é a mais significativa na dimensão dos dados. Através dela os autovetores devem ser ordenados na ordem de maior significância (de maior autovalor).

Este processo irá possibilitar analisar a componente mais importante, como também filtrar as componentes de baixa relevância.

2.1.5. Novo conjunto de dados

Concluído o método LDA, basca recompor os dados pela equação

$$DadoEixoClassificado = (Feature\ Vector^T \times DadosGrupo) \quad (16)$$

3. Proposta de implementação

3.1. Algoritmo de Análise Discriminante Linear LDA:

O algoritmo para realizar a análise discriminante linear irá utilizar, além das rotinas básicas do PCA, as rotinas: feature_sw, rotina para cálculo da dispersão intragrupo, recebe a matriz com os dados segmentados, quantidade de grupos e variáveis/dimensões e retorna a matriz de dispersão de cada grupo; e feature_sb, rotina para cálculo da dispersão intergrupos, recebe a matriz com os dados segmentados e quantidade de variaveis/dimensoes e retorna a matriz de dispersao entre grupos

3.2. Pseudocódigos

feature_sw(matrizD, medias, grupos, variaveis)
recebe matrizes dados$_{(n,m)}$ com dados separados por grupos
Media$_{(m)}$ ← calcula a média de cada coluna m da matrizD
Matrizsw$_{(i,j)}$ ← somatória (D$_{(i,j)}$ - Media$_{(j)}$)* (D$_{(i,j)}$ - Media$_{(j)}$)
retorna Matrizsw

feature_sb(A, medias, grupos, variaveis)
recebe matrizes dados$_{(n,m)}$ com dados separados por grupos
N ← número de elementos
MGlobal$_{(m)}$ ← calcula a média de todas colunas m da matriz
MatrizSb$_{(i,j)}$ ← somatória N* (A$_{(i,j)}$- MediaGlobal$_{(j)}$)*(A$_{(i,j)}$-MGlobal$_{(j)}$)
retorna MatrizSb

4. Experimentação e Resultados

Para verificar o funcionamento do algoritmo de análise discriminante linear, foi realizada a implementação em Python confrontando os resultados entre LDA, PCA e MDP (LDA+PCA) com a base de dados de Iris de Fisher:

- Dados da classificação de 3 espécies de flore, setosa, versicolor e virginica; com 150 amostras segmentadas pela largura e comprimento da sépala e da pétala, (http://en.wikipedia.org/wiki/Iris_flower_data_set).

Ambiente:

PyCharm 2019.2.2 (Professional Edition) Build#PY-192.6603.34

Python 3.7.5 (tags/v3.7.5:5c02a39a0b, Oct 15 2019, 01:31:54) on win32

Bibliotecas:

matplotlib-3.1.1	(utilizado para plotagem de gráficos)
pandas-0.25.2	(suporte à plotagem de gráficos)
xlrd-1.2.0	(leitura de arquivos do Excel - base de dados)
numpy-1.17.4	(gestão de matrizes)

Base de Dados:

LDAdb.xlsx	(Base íris fisher)

4.1. Iris Fisher

Base de dados:

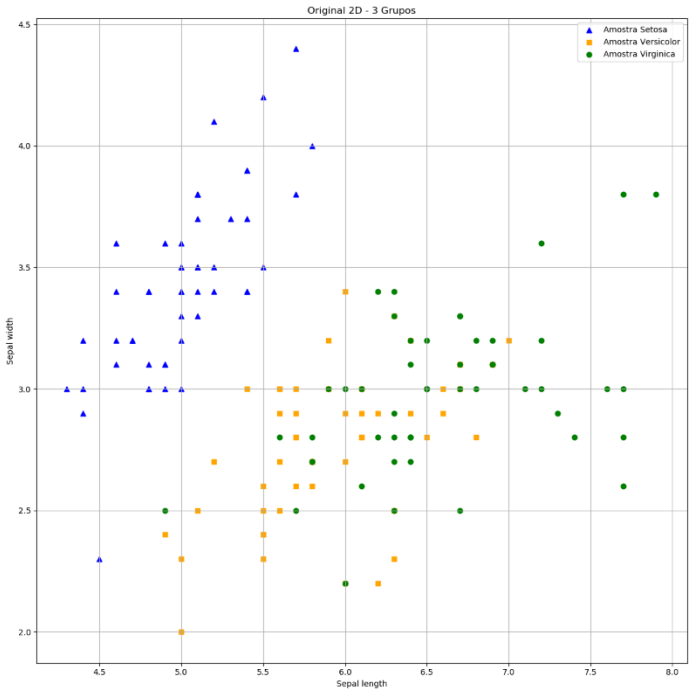

Fig2: Distribuição em 2D da base de dados Iris-Fisher

Iris-Fisher				
Tipo	Sepal length	Sepal width	Petal length	Petal width
Setosa	5.1	3.5	1.4	0.2
Setosa	4.9	3	1.4	0.2
Setosa	4.7	3.2	1.3	0.2
Setosa	4.6	3.1	1.5	0.2
		...		
		...		
Versicolor	7	3.2	4.7	1.4
Versicolor	6.4	3.2	4.5	1.5
Versicolor	6.9	3.1	4.9	1.5
Versicolor	5.5	2.3	4	1.3
		...		
		...		
Virginica	6.3	3.3	6	2.5
Virginica	5.8	2.7	5.1	1.9
Virginica	7.1	3	5.9	2.1
Virginica	6.3	2.9	5.6	1.8

Resultados:

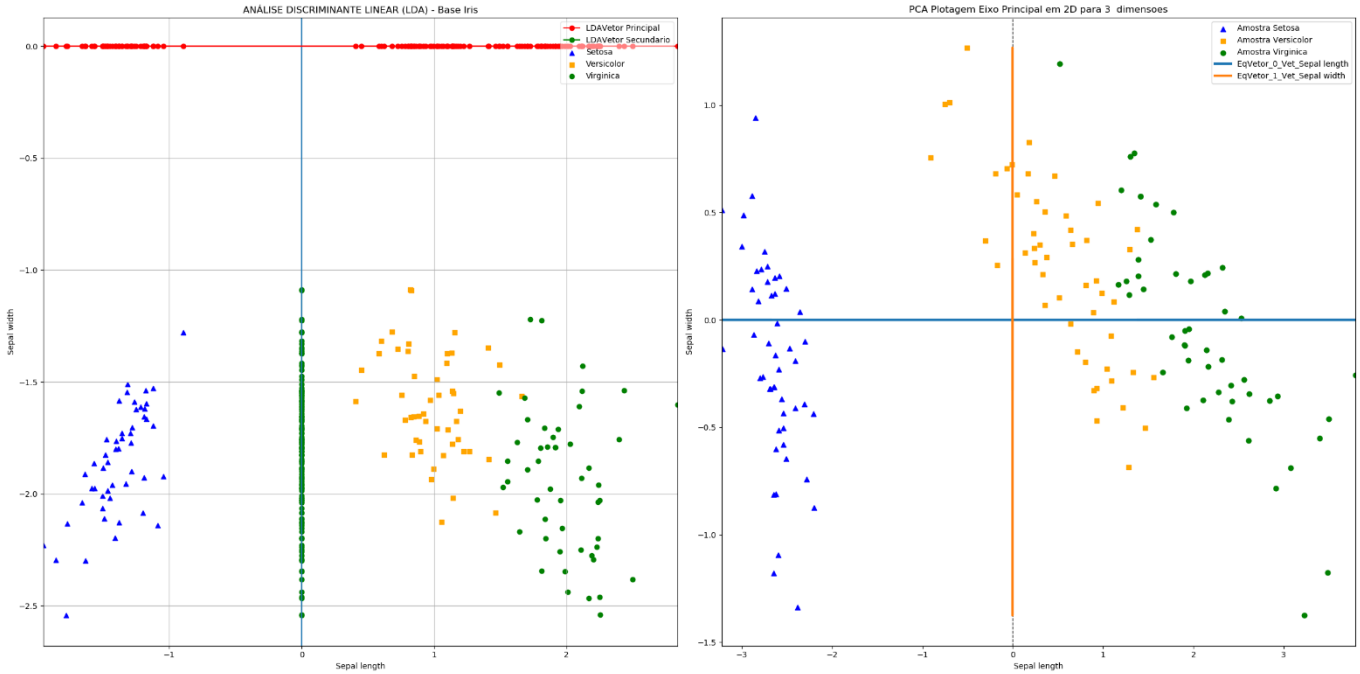

Fig3: Distribuição em 2D - Iris-Fisher [LDA x PCA]

Ambos os métodos alteram, de forma distinta, pela rotação vetorial dos dados, os eixos cartesianos de referência observados na figura 2.

A representação do LDA e do PCA nos gráficos das figuras 3 mostram no primeiro, pelo o eixo vermelho, a maior segmentação entre os grupos de flores e, no segundo, a direção da componente de maior variação dos dados, componente principal azul, comprimento da sépala, na relação com a largura da sépala.

É possível observar que tanto o LDA quanto o PCA possibilitam segmentar facilmente o grupo da setoda das demais flores, porém, o LDA possibilita uma distinção maior entre a vesicolor e virginica através do eixo longitudinal ao de maior distinção entre os grupos, o eixo vermelho.

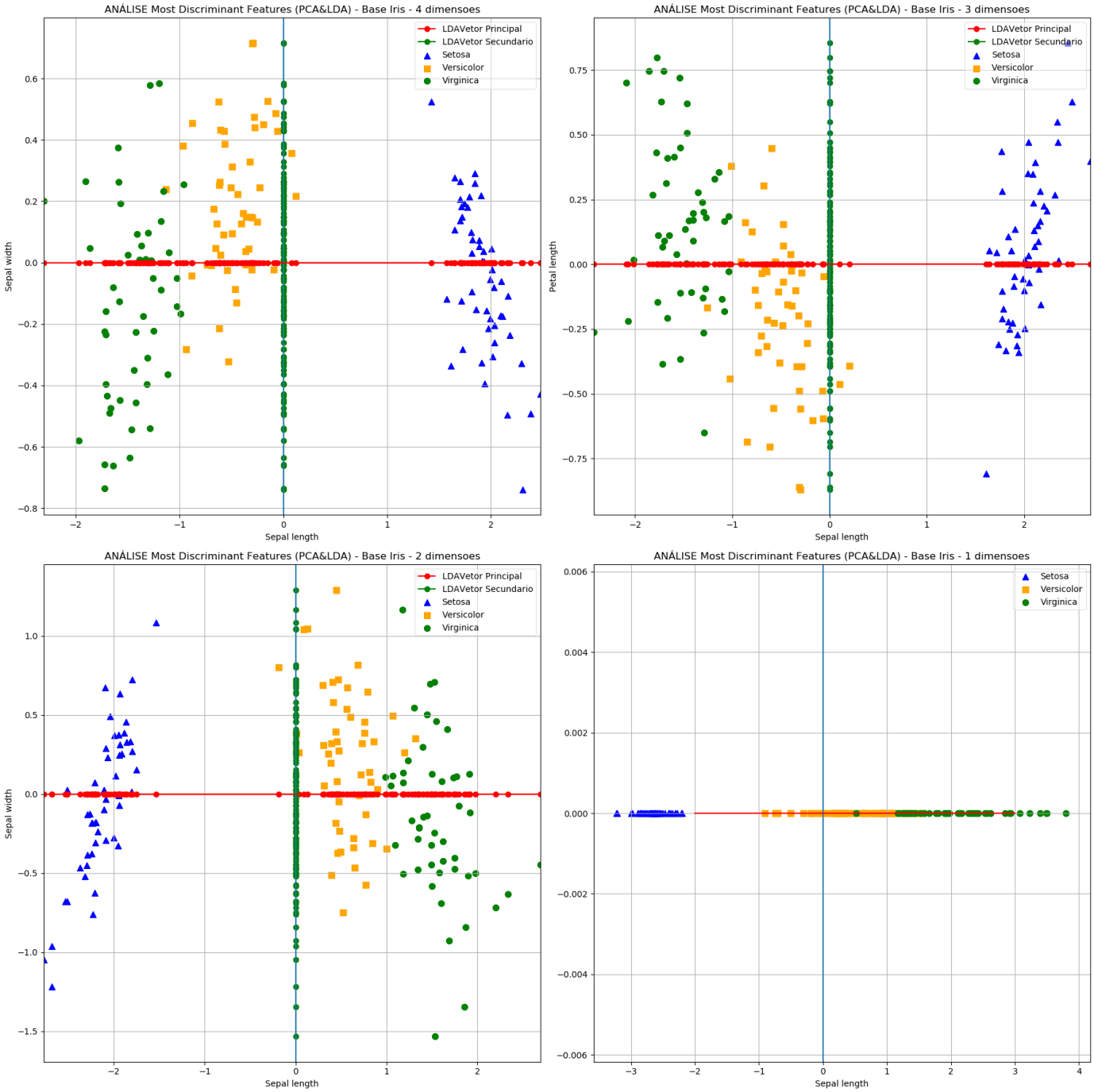

Fig4: Distribuição em 2D, interferência de múltiplas dimensões - Iris-Fisher [MDF - LDA + PCA]

A figura 4 mostra a utilização cruzada dos métodos PCA e LDA, denominado Most Discriminant Feature, sendo o primeiro responsável por preparar os dados de maior variação antes de serem classificados pelo LDA, aumentando com a distinção entre as classes pelo LDA.

Nos gráficos da figura 4 foram reduzidas as dimensões da base de dados Iris pelas técnicas do PCA e reprocessadas pelo LDA, é possível observar uma mudança dos eixos secundários, e um menor espalhamento dos dados até concluir por 1 dimensão na qual a distinção entre as classes é a maior observada.

Iris-Fisher		
PCA		
Covariância	Σ	$\begin{pmatrix} 0.685 & -0.0424 & 1.2743 & 0.515 \\ -0.042 & 0.19 & -0.3297 & -0.1213] \\ 1.2743 & -0.3297 & 3.1163 & 1.294 \\ 0.5157 & -0.1213 & 1.294 & 0.5797 \end{pmatrix}$
Autovalores	Λ	$(4.2269 \quad 0.2426 \quad 0.0783 \quad 0.0238)$
Autovetores	Φ	$\begin{pmatrix} 0.3615 & -0.6564 & -0.5821 & 0.3156 \\ -0.0845 & 0.1737 & 0.0761 & -0.4795 \\ 0.3579 & 0.0746 & 0.5467 & 0.7533 \\ 0.3579 & 0.0746 & 0.5467 & 0.7533 \end{pmatrix}$
LDA		
Dispersão intra Grupos - Sw	S_w	$\begin{pmatrix} 44.2384 & -26.0009 & -6.109 & -1.7931 \\ -26.0009 & 31.164 & 1.8706 & 0.3692 \\ -6.109 & 1.8706 & 10.4059 & -0.2694 \\ -1.7931 & 0.3692 & -0.2694 & 3.4897] \end{pmatrix}$
Dispersão inter Grupos - Sb	S_b	$\begin{pmatrix} 58.55 & 26.00 & 6.109 & 1.793 \\ 26.00 & 4.989 & -1.876 & -0.3692 \\ 6.109 & -1.876 & 1.26 & 0.2694 \\ 1.793 & -0.3692 & 0.2694 & 0.05892 \end{pmatrix}$

5. Conclusão

O método da análise dos componentes principais (LDA) oferece uma oportunidade para a criação de modelos estatísticos com a segmentação de elementos de uma massa de dados difusa, na ótica de eixos/componentes de maior relevância, neste método a separação inter-classes é enfatizada através da substituição da matriz de covariância total do PCA por uma medida de separabilidade com o critério Fisher.

Em geral, a abordagem LDA possibilita a obtenção de resultados de discriminação de dados melhores que o PCA com redução de dimensionalidade. Porém a modalidade MDF, que utiliza as técnicas PCA de redução de dimensionalidade, e em seguida a LDA para a classificação das amostras, mostra um resultado ainda melhor para a segmentação de elementos em uma amostra.

6. Referências

[1] FISHER, R.A. **The use of multiple measurements in taxonomic problems**. Annals of human Eugenics, v.7, p.179-188, 1936.

[2] FÁVERO, Luiz Paulo et al. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

[3] HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning**: Data Mining, Inference, and Prediction. 2. ed. Stanford: Springer, 2008.

[4] WILBUR, Anderson Theodore. **An introduction to Multivariate Statistical Analysis**. 2. ed. Stanford: Wiley, 1971.

[5] PRESS, William H. et al. **Numerical Recipes**: The Art of Scientific Computing. 3. ed. Cambridge, Massachusetts: Cambridge University Press, 2007.