Linear Discriminant Analisys (LDA)

Seminário de Estatística

Pós Graduação em Modelagem Computacional

Lorran de Araújo Durães Soares

November 25, 2024

Laboratório Nacional de Computação Científica

Table of Contents

- O LDA é um método estatístico utilizado para redução de dimensionalidade e classificação. Seu objetivo principal é encontrar um espaço de características de menor dimensão, de modo que as classes de dados sejam mais separáveis.
- Isso é feito maximizando a variabilidade entre as classes e minimizando a distância entre as médias das classes;
- Reduz para no máximo $C - 1$ dimensões, onde C é o número de classes.

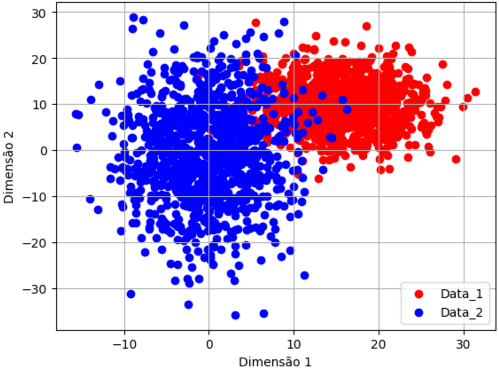
- Também pode ser usado como classificador, projetando os dados e aplicando, por exemplo (no caso 2D), um treshold para determinar a separação.
- É útil em situações onde as classes são linearmente separáveis, como em problemas de reconhecimento de faces, processamento de texto, e bioinformática.

Table of Contents

Projeta os dados de entrada para uma dimensão:

$$y = \mathcal{W}^T x$$

Como escolher o melhor $\mathcal{W}$?



[a]

[a] Fonte: Acervo Pessoal

1 - Maximizando a distância entre as classes:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n \in C_1} x_n, \quad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n \in C_2} x_n.$$

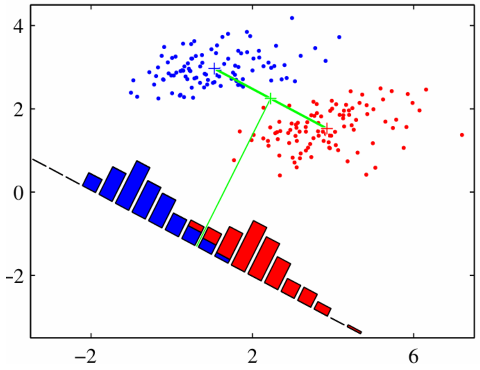
$$m_2 - m_1 = w^T(\mathbf{m}_2 - \mathbf{m}_1)$$

onde

$$m_k = w^T \mathbf{m}_k$$

Problemas:

- O vetor $\mathcal{W}$ pode se tornar arbitrariamente grande;
- Pode não escolher a melhor direção que separe os dados:



[a]

[a] Fonte: Bishop - Patter Recognition

2 - Minimizar variância dentro de cada classe:

$$\mathbf{s}_k{}^2 = \sum_{n \in C_K} (y_n - m_k)^2$$

Chegamos então ao critério de Fisher:

Critério de Fisher

$$\mathcal{J}(\mathcal{W}) = \frac{m_2 - m_1}{s_1{}^2 + s_2{}^2}$$

Em função de W fica:

$$\mathcal{J}(W) = \frac{W^T S_B W}{W^T S_W W}$$

Onde $\mathbf{S}_B$ é a matriz de covariância entre classes e é dada por:

$$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^\top$$

E $\mathbf{S}_W$ é a matriz de covariância total dentro das classes, dada por:

$$\mathbf{S}_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^\top + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^\top$$

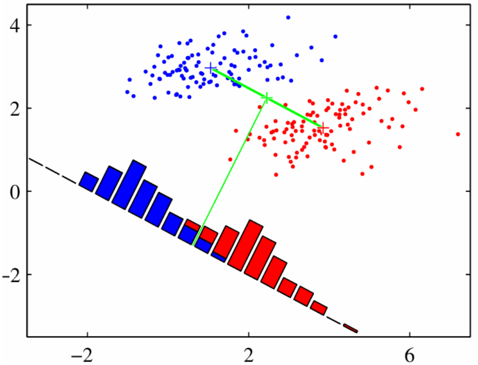
Diferenciando com respeito a $\mathbf{w}$ e igualando a zero, encontramos que $J(\mathbf{w})$ é maximizado por:

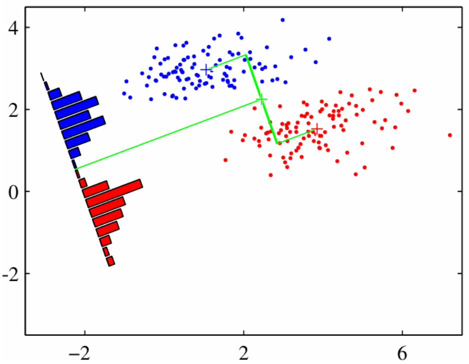
$$(\mathbf{w}^\top \mathbf{S}_B \mathbf{w})\mathbf{S}_W \mathbf{w} = (\mathbf{w}^\top \mathbf{S}_W \mathbf{w})\mathbf{S}_B \mathbf{w}$$

Que implica que:

$$\mathbf{w} \propto \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

Generalização para mais classes segue o mesmo príncipio.





[1]

[1] Fonte: Bishop - Patter Recognition

Classificação

- Treshold = 0;
- Teorema de Bayes:

$$p(\mathbf{x} \mid C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\}.$$

$$p(C_1 \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_1)p(C_1)}{p(\mathbf{x} \mid C_1)p(C_1) + p(\mathbf{x} \mid C_2)p(C_2)}$$

Classificação

$$p(C_1 \mid \mathbf{x}) = \sigma\left(\mathbf{w}^\top \mathbf{x} + w_0\right) = \frac{1}{1 + \exp(\mathbf{w}^\top \mathbf{x} + w_0)}$$

Classifica a amostra na classe com a maior probabilidade a posteriori.

Table of Contents

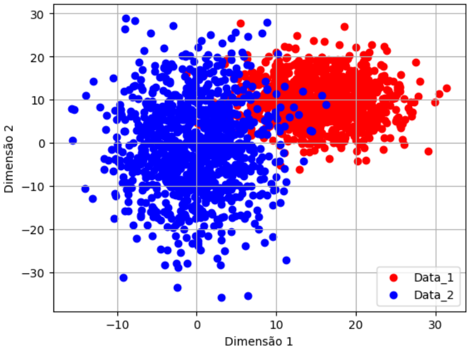
3 Hipóteses Implícitas

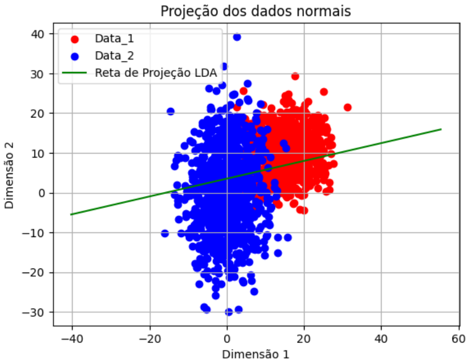
- Dados seguem uma distribuição normal multivariada;
- Classes com mesma matriz de covariância;
- Variáveis e características devem ser linearmente independentes;
- Funciona melhor com mais dados do que variáveis (N suficientemente grande).

Table of Contents
4 Aplicações

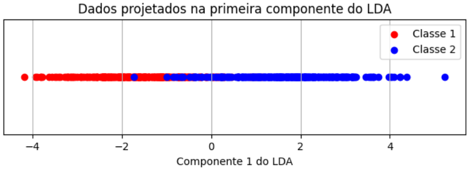
Dados Sintéticos Normais

- 1000 dados
- test size = 0.2
- 95 % de acurácia no conjunto de treinamento
- 93 % de acurácia no conjunto de teste





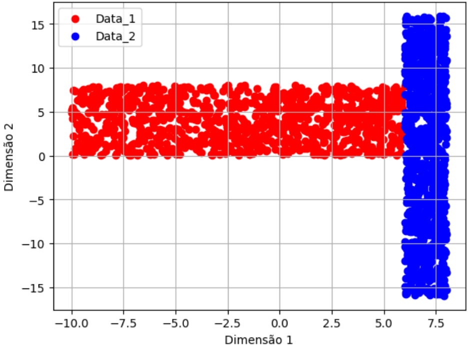
Projeção dos dados normais
Data_1
Data_2
Reta de Projeção LDA
Dimensão 1
Dimensão 2
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40
60

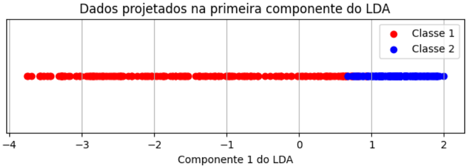


Dados projetados na primeira componente do LDA
Classe 1
Classe 2
Componente 1 do LDA
−4
−2
0
2
4

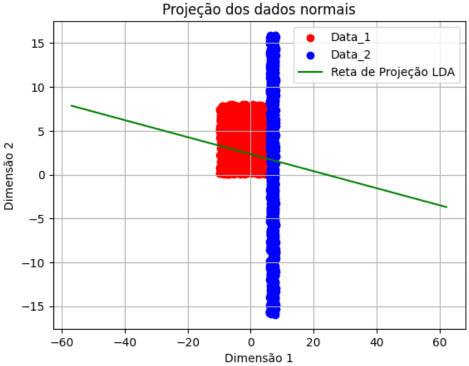
Dados Sintéticos não Normais

- Teste de não normalidade (Shapiro-Wilk com $\alpha = 0.5$):
  - Classe 1 - Dimensão 1: Estatística = 0.954, p = 0.000 Dimensão 2: Estatística = 0.954, p = 0.000
  - Classe 2 - Dimensão 1: Estatística = 0.959, p = 0.000 Dimensão 2: Estatística = 0.954, p = 0.000
- 1000 dados
- test size = 0.2
- 89 % de acurácia no conjunto de teste





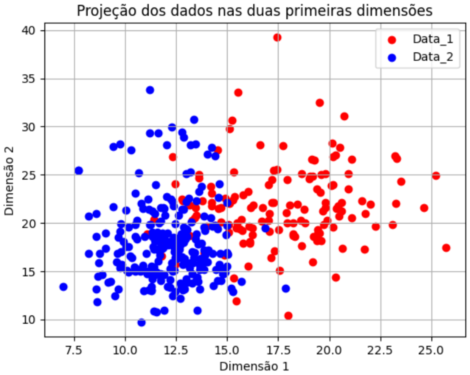
Dados projetados na primeira componente do LDA
Classe 1
Classe 2
−4
−3
−2
−1
0
1
2
Componente 1 do LDA

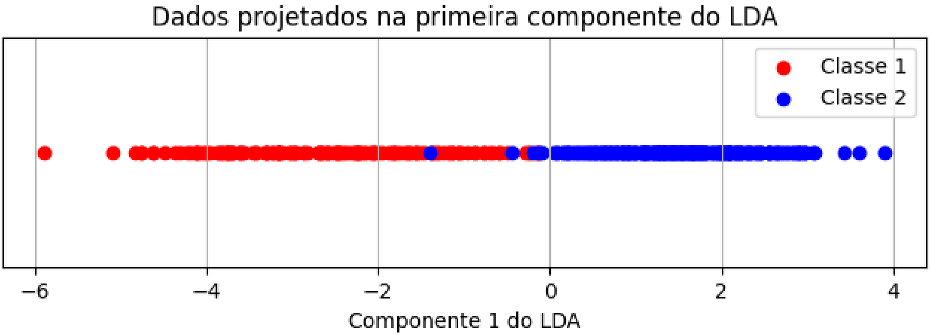


Projeção dos dados normais
Data_1
Data_2
Reta de Projeção LDA
15
10
5
0
−5
−10
−15
Dimensão 2
−60
−40
−20
0
20
40
60
Dimensão 1

Breast Cancer Database

- Teste de não normalidade (Shapiro-Wilk com $\alpha = 0.5$):
  - Classe 1: 28 dimensões não normais
  - Classe 2: 24 dimensões não normais
- 569 dados (212 malignos e 357 benignos)
- test size = 0.2
- 96 % de acurácia no conjunto de teste





Dados projetados na primeira componente do LDA
Classe 1
Classe 2
−6
−4
−2
0
2
4
Componente 1 do LDA

Table of Contents

- Técnica poderosa para dados linearmente separáveis;
- Tomar cuidado ao usar esses métodos devido às hipoteses estatísticas;
- Compreendendo as hipóteses, é possível usar de forma eficaz.

Table of Contents

- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Borges, M. R. (n.d.). GA-030 Estatística. Laboratório Nacional de Computação Científica. https://lncc.br/ mrborges/.
- Giraldi, G. A. (2024). Fundamentals of neural networks and statistical learning. Lecture Notes - Graduate Course, Graduate Program in Computational Modeling, PG-LNCC.

Linear Discriminant Analisys (LDA) *Thank*

you for listening!
Any questions?