

# Análise Discriminante Linear e Quadrática

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)  
Engenharia de Computação  
Departamento Acadêmico de Informática (Dainf)  
Universidade Tecnológica Federal do Paraná (UTFPR)  
Campus Pato Branco

- Análise Discriminante Linear e Quadrática
- Exemplo de Classificação
- Redução de Dimensionalidade Utilizando Análise Discriminante Linear

- A análise discriminante é um método estatístico de análise multivariada utilizado para identificar diferenças entre grupos
  - Relacionamento entre uma variável dependente (e.g. classe/alvo) e variáveis independentes
  - Obtenção da combinação linear de variáveis independentes com maior discriminação entre grupos
  - Introduzida por Fisher em 1936
  - Exemplos de métodos
    - Análise discriminante linear (LDA – *linear discriminant analysis*)
    - Análise discriminante quadrática (QDA – *quadratic discriminant analysis*)

- LDA e QDA são derivados de modelos probabilísticos simples
- Probabilidades
  - $P(A)$ : probabilidade de ocorrência do evento  $A$
  - $p(x)$ : função de densidade de probabilidade (pdf) para uma variável  $x$
  - $p(X)$ : pdf para um vetor de variáveis aleatórias  $X$
- Probabilidades condicionais:
  - $P(A|B)$ : probabilidade condicional de  $A$  dado  $B$
  - $P(x|B)$  e  $P(X|B)$

## Análise Discriminante Linear e Quadrática

# Análise Discriminante Linear e Quadrática

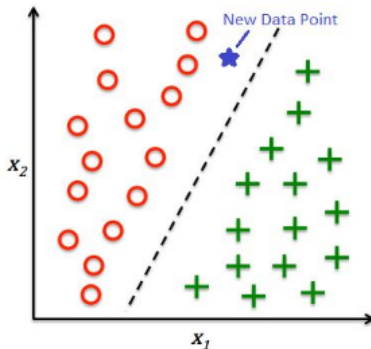
- LDA é uma técnica de aprendizado de máquina supervisionado que tem o propósito de separar grupos ou classes de dados com base em combinações lineares de características
  - Generalização do discriminante linear de Fisher
- Dadas características (atributos) de um grande conjunto de treinamento para a classe  $\omega_i$
- Cada um desses padrões de treinamento tem um valor  $x$  diferente para as características
  - Probabilidade condicional da classe:  $p(x|\omega_i)$

- Com que frequência os exemplos de classe  $\omega_i$  apresentam a característica  $x$ ?

Outlook			Temperature			Humidity			Windy			Play		
	Yes	No	Yes	No		Yes	No		Yes	No		Yes	No	
Sunny	2	3	64, 68,	65, 71,		65, 70,	70, 85,		False	6	2	9	5	
Overcast	4	0	69, 70,	72, 80,		70, 75,	90, 91,		True	3	3			
Rainy	3	2	72, ...	85, ...		80, ...	95, ...							
Sunny	2/9	3/5	$\mu = 73$	$\mu = 75$		$\mu = 79$	$\mu = 86$		False	6/9	2/5	9/14	5/14	
Overcast	4/9	0/5	$\sigma = 6.2$	$\sigma = 7.9$		$\sigma = 10.2$	$\sigma = 9.7$		True	3/9	3/5			
Rainy	3/9	2/5												

- Classificação
  - Dado um vetor de características  $X$ , qual a probabilidade do mesmo pertencer a uma classe  $\omega_i$ ?

$$P(\omega_i, X)$$





- Durante o treinamento, é dada  $p(X|\omega_i)$  (*a priori*), as o que é desejável seria  $p(\omega_i|X)$  (*a posteriori*)
- Teorema de Bayes
  - Forma geralmente apresentada

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Contexto deste material

$$P(\omega_i|X) = \frac{P(X|\omega_i)P(\omega_i)}{P(X)}$$

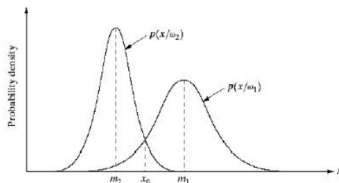
$$P(\omega_i|X) = \frac{P(X|\omega_i)P(\omega_i)}{P(X)}$$

- $P(X|\omega_i)$ : probabilidade ou verossimilhança condicionada à classe
- $P(\omega_i)$ : probabilidade *a priori*
- $P(X)$ : evidência (geralmente ignorada)
- $P(\omega_i|X)$ : probabilidade *posteriori*

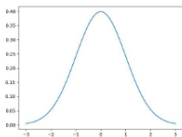
- Estrutura de classificadores baseados em análise discriminante (linear e quadrática)
  - Treinamento: estimar  $p(X|\omega_i)$  de cada classe
  - Conhecimento *a priori*: estimar  $p(\omega_i)$  da população em geral
- Classificação
  - Extração de características ( $X$ ) para o novo padrão
  - Calcular probabilidades *a posteriori*  $P(\omega_i|X)$  para cada classe
  - Atribuir uma classe ao novo padrão para o que obteve maior valor de  $P(\omega_i|X)$

- Suposição de normalidade
  - Para as análises discriminantes linear e quadrática, assume-se que  $p(x|\omega_i)$  tenha sido modelada como uma distribuição Gaussiana multivariada
    - Probabilidades condicionais de classe normalmente distribuídas (1D)

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2}$$



- De probabilidades para discriminantes: caso 1-D
  - Desejável maximizar:  $P(\omega_i|X) = \frac{p(X|\omega_i)P(\omega_i)}{p(x)}$
  - O mesmo que maximizar:  $p(X|\omega_i)P(\omega_i)$
  - Que para uma distribuição normal é:  $\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(X-\mu_i)^2/\sigma_i^2} P(\omega_i)$
  - Aplicação do logaritmo na base 2:  
 $\log_2 \frac{1}{\sqrt{2\pi}} - \log_2 \sigma_i - \frac{1}{2}(X - \mu_i)^2/\sigma_i^2 + \log_2 P(\omega_i)$
  - Remoção de constantes:  $\log_2 P(\omega_i) - \log_2 \sigma_i - \frac{1}{2}(X - \mu_i)^2/\sigma_i^2$

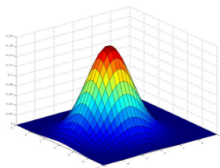


$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2}$$

- De probabilidades para discriminantes: múltiplas características
  - O termo-chave para uma distribuição normal 1-D é a distância ao quadrado da média em desvios-padrão:  $(X - \mu)^2 / \sigma_i^2$
  - A modelagem acima pode ser estendida para múltiplas características por meio da normalização da distância de cada atributo pelo respectivo desvio-padrão
    - Em seguida, utilizar a classificação pela distância mínima
  - Essa normalização é também denominada como naïve Bayes por ignorar relações entre características

- Distribuição Gaussiana multivariada

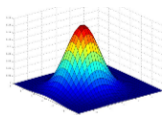
$$\begin{aligned} p(\mathbf{x}) &= \left( \frac{1}{\sqrt{2\pi}} \right)^d \frac{1}{|\mathbf{C}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \\ &= (2\pi)^{-d/2} |\mathbf{C}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \end{aligned}$$



- Para a classificação multiclasse, cada classe  $\omega_i$  tem um vetor de médias ( $m_i$ ) e uma matriz de covariância ( $C_i$ )
  - Dessa forma, as probabilidades condicionais de classe são dadas pela equação abaixo:

$$p(X|\omega_i) = (2\pi)^{-\frac{d}{2}} |C_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(X-m_i)^T C_i^{-1}(X-m_i)}$$

- De probabilidades para discriminantes: caso N-D
  - Desejável maximizar:  $P(\omega_i|X) = \frac{p(X|\omega_i)P(\omega_i)}{p(x)}$
  - O mesmo que maximizar:  $p(X|\omega_i)P(\omega_i)$
  - O mesmo que maximizar:  $\log_2 p(X|\omega_i) + \log_2 P(\omega_i)$
  - Que para uma distribuição normal é:  
 $-\frac{d}{2} \log_2 2\pi - \frac{1}{2} \log_2 |C_i| - \frac{1}{2}(X - m_i)^T C_i^{-1}(X - m_i) + \log_2 P(\omega_i)$
  - Maximize:  $\log_2 P(\omega_i) - \frac{1}{2} \log_2 |C_i| - \frac{1}{2}(X - m_i)^T C_i^{-1}(X - m_i)$



$$p(\mathbf{x}|\omega_i) = (2\pi)^{-d/2} |C_i|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T C_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$



- Distância de Mahalanobis
  - A expressão  $(X - m_i)^T C_i^{-1} (x - m_i)$  pode ser também definido como  $\|x - m_i\|_{C^{-1}}^2$ 
    - Por mais que pareça uma distância quadrática (como a Euclidiana), a inversa da matriz de covariância  $C^{-1}$  atua como uma métrica
    - O reconhecimento de padrões usando distribuições normais multivariadas é simplesmente um classificador de distância mínima (de Mahalanobis)!
    - Temos 3 casos de matriz de covariância a serem considerados

- Caso 1: matriz de identidade (*naïve* Bayes)
  - Suponha que a matriz de covariância para todas as classes seja uma matriz identidade:  $C_i = I$  ou  $C_i = \sigma^2 I$
  - Se os dados estão normalizados por meio do método z-score e não estão correlacionados, a matriz de correlação é a matriz identidade com desvio padrão unitário

$$g_i(X) = -\frac{1}{2}(X - m_i)^T(X - m_i) + \log_2 P(\omega_i)$$

- Supondo que todas as classes sejam igualmente prováveis *a priori*:  $g_i(X) = -\frac{1}{2}(X - m_i)^T(X - m_i)$
- Ao ignorarmos a constante  $\frac{1}{2}$ , temos:  
 $g_i(X) = -(X - m_i)^T(X - m_i)$

- Caso 2: mesma matriz de covariância (análise discriminante linear)
  - Caso cada classe possua a mesma matriz de covariância:
$$g_i(X) = -\frac{1}{2}(X - m_i)^T C (X - m_i) + \log_2 P(\omega_i)$$
  - Os loci de probabilidade constante são hiper-elipses orientados com os autovetores de  $C$ 
    - Direções dos autovetores dos eixos da elipse
    - Variância dos autovalores (comprimento do eixo ao quadrado) na direção do eixo
  - Os limites de decisão ainda são hiperplanos, embora possam não ser mais normais às linhas entre as respectivas médias de classe

- Caso 3: diferentes matrizes de covariância para cada classe (análise discriminante quadrática)
  - Suponha que cada classe tenha sua própria matriz de covariância arbitrária (o caso mais geral):  $C_i \neq C_j$

$$g_i(X) = \log_2 P(\omega_i) - \frac{1}{2} \log_2 |C_i| - \frac{1}{2} (X - m_i)^T C_i^{-1} (X - m_i)$$

- Os loci de probabilidade constante para cada classe são orientados por hiper-elipses com os autovetores de  $C_i$  para essa classe
- Os limites de decisão são quadráticos, especificamente, hiper-elipses ou hiper-hiperboloides.

## Exemplo de Classificação

# Exemplo de Classificação

- Treinamento: determinar médias e matriz de covariâncias

$$\mathbf{X} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \\ 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

$$\mathbf{C} = \begin{pmatrix} 0.2403 & -0.2694 \\ -0.2694 & 1.9742 \end{pmatrix}$$

$$\mathbf{C}^{-1} = \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix}$$

$$\boldsymbol{\mu} = [2.88 \quad 5.6771] \quad g = 2$$

# Exemplo de Classificação

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}(\mathbf{x} - \mathbf{m}_i) + \log P(\omega_i)$$

Classificação:  $\mathbf{x} = [3 \quad 7]$   $P(i | \mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{C}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - 2 \ln(P(i))$

- Classe 1:  $P(1 | \mathbf{x}) = ([3 \quad 7] - [3.05 \quad 6.38])^T \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix} ([3 \quad 7] - [3.05 \quad 6.38]) + 1.1192$   
 $\boldsymbol{\mu}_1 = [3.05 \quad 6.38]$   $P(1) = 4/7$
- Classe2:  $P(2 | \mathbf{x}) = ([3 \quad 7] - [2.67 \quad 4.73])^T \begin{pmatrix} 4.9129 & 0.6705 \\ 0.6705 & 0.5980 \end{pmatrix} ([3 \quad 7] - [2.67 \quad 4.73]) + 1.6946$   
 $\boldsymbol{\mu}_2 = [2.67 \quad 4.73]$   $P(2) = 3/7$

# Exemplo de Classificação

$$g_i(\mathbf{x}) = \log P(\omega_i) - \frac{1}{2} \log |\mathbf{C}_i| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i)$$

$$\mathbf{X}_{classe1} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix}$$

$$\mathbf{C}_1 = \begin{pmatrix} 0.1876 & -0.4127 \\ -0.4127 & 1.1372 \end{pmatrix}$$

$$\mathbf{C}_1^{-1} = \begin{pmatrix} 26.3961 & 9.5785 \\ 9.5785 & 4.3552 \end{pmatrix}$$

$$\boldsymbol{\mu}_1 = [3.05 \quad 6.38] \quad P(1) = 4/7$$

$$\mathbf{X}_{classe2} = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}$$

$$\mathbf{C}_2 = \begin{pmatrix} 0.3141 & -0.7308 \\ -0.7308 & 1.8785 \end{pmatrix}$$

$$\mathbf{C}_2^{-1} = \begin{pmatrix} 33.5580 & 13.0550 \\ 13.0550 & 5.6111 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = [2.67 \quad 4.73] \quad P(2) = 3/7$$



# Exemplo de Classificação

- QDA:  $P(1|\mathbf{x})=-0.8791$ ;  $P(2|\mathbf{x})=50.9385$

- $x$  é da **classe 2**

$$g_i(\mathbf{x}) = \log P(\omega_i) - \frac{1}{2} \log |\mathbf{C}_i| - \frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x} - \mathbf{m}_i)$$

- Fisher:  $P(1|\mathbf{x})=1.3198$ ;  $P(2|\mathbf{x})=6.3157$

- $x$  é da **classe 2**

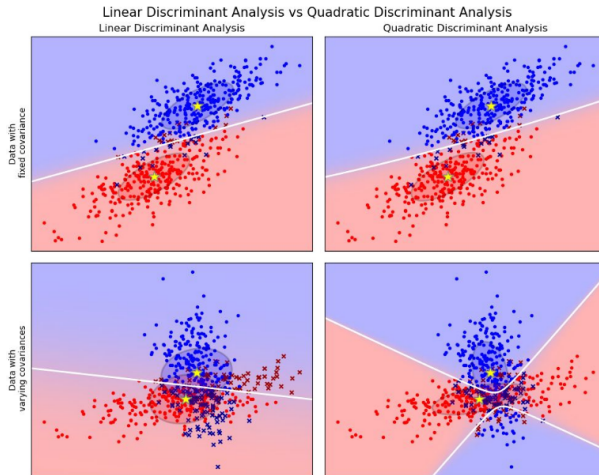
$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}(\mathbf{x} - \mathbf{m}_i) + \log P(\omega_i)$$

- Bayes:  $P(1|\mathbf{x})=1.5061$ ;  $P(2|\mathbf{x})=6.9564$

- $x$  é da **classe 2**

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^T (\mathbf{x} - \mathbf{m}_i)$$

# Exemplo de Classificação



## **Redução de Dimensionalidade Utilizando Análise Discriminante Linear**

# Redução de Dimensionalidade Utilizando Análise Discriminante Linear

- A redução de dimensionalidade tem a finalidade de facilitar a visualização e o processamento de conjuntos de exemplos com várias características (multidimensional)
- A LDA busca maximizar a separabilidade entre as classes
- Passo-a-passo para a redução de dimensionalidade usando LDA:
  - 1 Cálculo das médias para cada classe
  - 2 Obtenção da matriz de dispersão intra-classe
  - 3 Obtenção da matriz de dispersão entre-classe
  - 4 Geração de autovalores e autovetores
  - 5 Seleção dos autovetores com maiores autovalores
  - 6 Projeção dos dados em novo espaço

# Redução de Dimensionalidade Utilizando Análise Discriminante Linear

- Passo 1: cálculo das médias para cada classe
  - Para cada classe  $\omega_i$ , obter o vetor de médias:  $m_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$
  - Cálculo da média global:  $m = \frac{1}{N} \sum_{i=1}^N x_i$
- Passo 2: obtenção da matriz de dispersão intra-classe ( $S_W$ )
  - Determinar o quanto os dados estão dispersos dentro de cada classe, onde  $t$  é o número total de classes:

$$S_W = \sum_{i=1}^t \sum_{x \in \omega_i} (x - m_i)(x - m_i)^T$$

# Redução de Dimensionalidade Utilizando Análise Discriminante Linear

- Passo 3: obtenção da matriz de dispersão entre-classe ( $S_B$ )
  - Determinação de quanto as médias das classes diferem da média global

$$S_B = \sum_{i=1}^t N_i (m_i - m)(m_i - m)^T$$

- Passo 4: geração de autovalores e autovetores
  - Para a obtenção da matriz de projeção  $W$ , maximizar a razão:

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

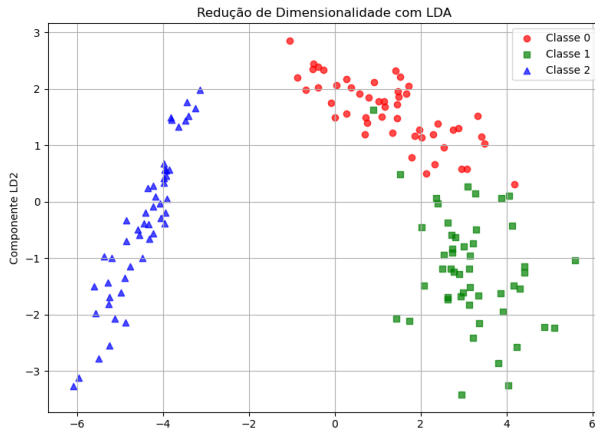
- Restrição de ortogonalidade:  $W^T W = I$
- Problema de autovalores generalizados, resultando em  $S_W^{-1} S_B W = \lambda W$ , onde:
  - $W$  são autovetores
  - $\lambda$  são autovalores

# Redução de Dimensionalidade Utilizando Análise Discriminante Linear

- Passo 5: seleção dos autovetores com maiores autovalores
  - Seleção dos  $k$  maiores autovalores, onde  $k < d$
  - A matriz  $W_k$  terá dimensão  $d \times k$
- Passo 6: projeção dos dados em novo espaço

$$X_{\text{reduzido}} = X \cdot W_k$$

# Redução de Dimensionalidade Utilizando Análise Discriminante Linear





# Considerações Finais

- Atributos
  - Numéricos e simétricos
  - Suportam probabilidades *a priori*
  - Assume que atributos são igualmente importantes
  - Seleção de atributos
- Capacidade de classificar padrões com valores ausentes
- Hipótese de dependência entre atributos
- Pode ter melhor desempenho em comparação com o Naïve Bayes, especialmente caso sejam utilizados atributos correlacionados
- Complexidade computacional entre  $O(n)$  e  $O(n^3)$



BISHOP, C. M.

*Pattern Recognition and Machine Learning.*

Springer, 2006.



CASANOVA, D.

LDA and QDA. Aprendizado de Máquina.

*Slides.* Engenharia de Computação. Dainf/UTFPR, 2020.



DUDA R., Hart P., STORK D.

Pattern Classification.

Wiley Interscience, 2002.



MENOTTI D.

Classificação. Aprendizado de Máquinas.

*Slides.* Especialização em Engenharia Industrial 4.0. UFPR, 2020.



MITCHELL T.

Machine Learning.

WCB McGraw-Hill, 1997.



RASCHKA, S.; MIRJALILI, V.

*Python Machine Learning.*

*Packt, 2017.*