

Introdução à Classificação de Padrões

Guilherme de Alencar Barreto

`gbarreto@ufc.br`

Grupo de Aprendizado de Máquinas – GRAMA
Departamento de Engenharia de Teleinformática
Universidade Federal do Ceará – UFC
<http://lattes.cnpq.br/8902002461422112>

Introdução à Classificação de Padrões

Pré-Requisitos

- ① Noções de Álgebra Linear
- ② Noções de Geometria Analítica
- ③ Noções de Funções
- ④ Noções de Limites

Introdução à Classificação de Padrões

Objetivo Geral

Apresentar vários classificadores baseados em medidas de dissimilaridade e seu amplo uso em reconhecimento de padrões, bem como relações com medidas de similaridade e o conceito de métrica.

Introdução à Classificação de Padrões

Conteúdo dos Slides

- ① O Que é Ser Inteligente?
- ② Inteligência e Reconhecimento de Padrões
- ③ Conceituação Intuitiva do Problema
- ④ Um Computador pode Reconhecer Padrões?
- ⑤ Componentes de um Problema de Classificação
- ⑥ Classificadores Elementares (NN e Centróide)
- ⑦ Medidas de (Dis-)similaridade
- ⑧ Técnicas de Normalização dos Dados
- ⑨ Correlação e Matriz de Covariância
- ⑩ Classificador Quadrático

Introdução à Classificação de Padrões

Relação com Inteligência Natural

O Que é Ser Inteligente?

- Seria ter habilidade específicas em certos domínios do conhecimento?
(e.g. ser um neurocirurgião)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

O Que é Ser Inteligente?

- Seria ter habilidade específicas em certos domínios do conhecimento?
(e.g. ser um neurocirurgião)
- Ou resolver problemas genéricos de modo aproximado, porém satisfatório?
(e.g. achar vaga em estacionamento)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

O Que é Ser Inteligente?

- Seria ter habilidade específicas em certos domínios do conhecimento?
(e.g. ser um neurocirurgião)
- Ou resolver problemas genéricos de modo aproximado, porém satisfatório?
(e.g. achar vaga em estacionamento)
- Ou ainda: ter conhecimento enciclopédico?
(e.g. ser um erudito)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

O Que é Ser Inteligente?

- Seria ter habilidade específicas em certos domínios do conhecimento?
(e.g. ser um neurocirurgião)
- Ou resolver problemas genéricos de modo aproximado, porém satisfatório?
(e.g. achar vaga em estacionamento)
- Ou ainda: ter conhecimento enciclopédico?
(e.g. ser um erudito)
- Tocar um instrumento? Falar outras línguas? Jogar bola bem?

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ➊ Comportamentais (fulano age sempre assim!)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ❶ Comportamentais (fulano age sempre assim!)
 - ❷ Sonoros (Este barulho não é normal!)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ❶ Comportamentais (fulano age sempre assim!)
 - ❷ Sonoros (Este barulho não é normal!)
 - ❸ Táteis (Estes tecidos têm texturas distintas!)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ❶ Comportamentais (fulano age sempre assim!)
 - ❷ Sonoros (Este barulho não é normal!)
 - ❸ Táteis (Estes tecidos têm texturas distintas!)
 - ❹ Visuais (Acho que vai chover hoje!)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ❶ Comportamentais (fulano age sempre assim!)
 - ❷ Sonoros (Este barulho não é normal!)
 - ❸ Táteis (Estes tecidos têm texturas distintas!)
 - ❹ Visuais (Acho que vai chover hoje!)
 - ❺ Olfativos (De quem é esse perfume?)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ❶ Comportamentais (fulano age sempre assim!)
 - ❷ Sonoros (Este barulho não é normal!)
 - ❸ Táteis (Estes tecidos têm texturas distintas!)
 - ❹ Visuais (Acho que vai chover hoje!)
 - ❺ Olfativos (De quem é esse perfume?)
 - ❻ Numéricos (Quem pegou um dos doces?)

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
 - ➊ Comportamentais (fulano age sempre assim!)
 - ➋ Sonoros (Este barulho não é normal!)
 - ➌ Táteis (Estes tecidos têm texturas distintas!)
 - ➍ Visuais (Acho que vai chover hoje!)
 - ➎ Olfativos (De quem é esse perfume?)
 - ➏ Numéricos (Quem pegou um dos doces?)

Definição Informal de Reconhecimento de Padrões

Reconhecer padrões equivale a categorizar determinado objeto físico ou situação como pertencente (ou não) a um certo número de categorias previamente estabelecidas.

Introdução à Classificação de Padrões

Relação com Inteligência Natural

Inteligência e Reconhecimento de Padrões



(a) Grupo 1



(b) Grupo 2



(c) Grupo 3

Em qual dos grupos acima você colocaria o objeto abaixo:



Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Certamente, a sua decisão é tomada com base no **grau de similaridade** entre a fruta desconhecida e as frutas conhecidas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Certamente, a sua decisão é tomada com base no **grau de similaridade** entre a fruta desconhecida e as frutas conhecidas.
- Que mecanismo seu cérebro usa para realizar esta tarefa?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Certamente, a sua decisão é tomada com base no **grau de similaridade** entre a fruta desconhecida e as frutas conhecidas.
- Que mecanismo seu cérebro usa para realizar esta tarefa?
- Será que implementa uma comparação entre o objeto novo e objetos armazenados?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Certamente, a sua decisão é tomada com base no **grau de similaridade** entre a fruta desconhecida e as frutas conhecidas.
- Que mecanismo seu cérebro usa para realizar esta tarefa?
- Será que implementa uma comparação entre o objeto novo e objetos armazenados?
- É possível “replicar” este mecanismo em uma máquina???

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

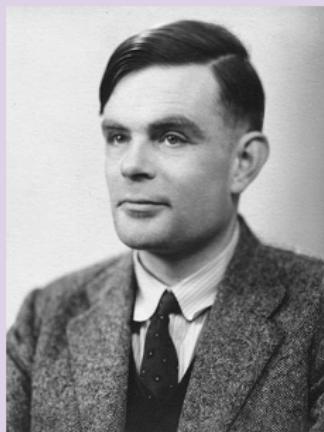


Figura: Alan Turing (23/06/1912 – 07/06/1954)^a

^aA. Turing (1950). “Computing Machinery and Intelligence”, *Mind*, vol. LIX, no. 236, p. 433–460. Disponível em <https://academic.oup.com/mind/article/LIX/236/433/986238>

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de
 - Uma **representação** dos atributos físicos das frutas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de
 - Uma **representação** dos atributos físicos das frutas.
 - Uma **memória** para armazenar as frutas já aprendidas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de
 - Uma **representação** dos atributos físicos das frutas.
 - Uma **memória** para armazenar as frutas já aprendidas.
 - Uma **regra de decisão** para classificar frutas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de
 - Uma **representação** dos atributos físicos das frutas.
 - Uma **memória** para armazenar as frutas já aprendidas.
 - Uma **regra de decisão** para classificar frutas.
 - Um **aprendizado** para introduzir novas frutas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?
 - ➌ Qual a sua cor?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?
 - ➌ Qual a sua cor?
 - ➍ Qual a textura da casca?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?
 - ➌ Qual a sua cor?
 - ➍ Qual a textura da casca?
 - ➎ Seu cheiro é ativo ou não?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?
 - ➌ Qual a sua cor?
 - ➍ Qual a textura da casca?
 - ➎ Seu cheiro é ativo ou não?
- Todos esses atributos são igualmente importantes?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
 - ➊ Qual seu formato?
 - ➋ É uma fruta cítrica?
 - ➌ Qual a sua cor?
 - ➍ Qual a textura da casca?
 - ➎ Seu cheiro é ativo ou não?
- Todos esses atributos são igualmente importantes?
- Quão difícil é a tarefa de definir os atributos de um objeto?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.
 - ➌ Alaranjada.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.
 - ➌ Alaranjada.
 - ➍ Rugosa.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.
 - ➌ Alaranjada.
 - ➍ Rugosa.
 - ➎ Ativo.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.
 - ➌ Alaranjada.
 - ➍ Rugosa.
 - ➎ Ativo.
- Provavelmente não!

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
 - ➊ Oval.
 - ➋ Cítrico.
 - ➌ Alaranjada.
 - ➍ Rugosa.
 - ➎ Ativo.
- Provavelmente não!
- Horrivelmente árdua! :-(

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!
- Como transformar os atributos físicos da tangerina em números?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!
- Como transformar os atributos físicos da tangerina em números?
- Temos que representar cada objeto como um vetor de p atributos numéricos!

$$\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_p]^T$$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!
- Como transformar os atributos físicos da tangerina em números?
- Temos que representar cada objeto como um vetor de p atributos numéricos!

$$\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_p]^T$$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!
- Como transformar os atributos físicos da tangerina em números?
- Temos que representar cada objeto como um vetor de p atributos numéricos!

$$\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_p]^T$$

- Assim, o objeto (tangerina) será representado como

$$\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 1]^T$$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

x_2 : {não, sim} = {0, 1}

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

x_2 : {não, sim} = {0, 1}

x_3 : {amarelo, vermelho, alaranjado, verde, marrom} = {0, 1, 2, 3, 4}

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

x_2 : {não, sim} = {0, 1}

x_3 : {amarelo, vermelho, alaranjado, verde, marrom} = {0, 1, 2, 3, 4}

x_4 : {lisa, rugosa, espinhosa} = {0, 1, 2}

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

x_2 : {não, sim} = {0, 1}

x_3 : {amarelo, vermelho, alaranjado, verde, marrom} = {0, 1, 2, 3, 4}

x_4 : {lisa, rugosa, espinhosa} = {0, 1, 2}

x_5 : {não, sim} = {0, 1}

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Para isso, foi definida e utilizada a seguinte convenção:

x_1 : {esférico, oval, alongado} = {0, 1, 2}.

x_2 : {não, sim} = {0, 1}

x_3 : {amarelo, vermelho, alaranjado, verde, marrom} = {0, 1, 2, 3, 4}

x_4 : {lisa, rugosa, espinhosa} = {0, 1, 2}

x_5 : {não, sim} = {0, 1}

- Note que os atributos devem descrever também outras frutas.

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $x = [0 \ 1 \ 2 \ 1 \ 0]$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $x = [0 \ 1 \ 2 \ 1 \ 0]$
 - **Objeto Maçã:** $y = [0 \ 0 \ 1 \ 0 \ 0]$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $x = [0 \ 1 \ 2 \ 1 \ 0]$
 - **Objeto Maçã:** $y = [0 \ 0 \ 1 \ 0 \ 0]$
 - **Objeto Tangerina:** $z = [0 \ 1 \ 2 \ 1 \ 1]$

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $x = [0 \ 1 \ 2 \ 1 \ 0]$
 - **Objeto Maçã:** $y = [0 \ 0 \ 1 \ 0 \ 0]$
 - **Objeto Tangerina:** $z = [0 \ 1 \ 2 \ 1 \ 1]$
- O objeto z se assemelha mais a x ou a y ? Ou seja,

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 0]$
 - **Objeto Maçã:** $\mathbf{y} = [0 \ 0 \ 1 \ 0 \ 0]$
 - **Objeto Tangerina:** $\mathbf{z} = [0 \ 1 \ 2 \ 1 \ 1]$
- O objeto \mathbf{z} se assemelha mais a \mathbf{x} ou a \mathbf{y} ? Ou seja,
 - $\mathbf{z} = [0 \ 1 \ 2 \ 1 \ 1]$ está mais próximo de $\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 0]$?

Introdução à Classificação de Padrões

Relação com Inteligência Artificial

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:
 - **Objeto Laranja:** $\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 0]$
 - **Objeto Maçã:** $\mathbf{y} = [0 \ 0 \ 1 \ 0 \ 0]$
 - **Objeto Tangerina:** $\mathbf{z} = [0 \ 1 \ 2 \ 1 \ 1]$
- O objeto \mathbf{z} se assemelha mais a \mathbf{x} ou a \mathbf{y} ? Ou seja,
 - $\mathbf{z} = [0 \ 1 \ 2 \ 1 \ 1]$ está mais próximo de $\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 0]$?
 - Ou o objeto \mathbf{z} está mais próximo de $\mathbf{y} = [0 \ 0 \ 1 \ 0 \ 0]$?

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Componentes de um Problema de Classificação

Considere um conjunto de dados \mathcal{X} formado por vetores de atributos e seus respectivos rótulos

$$\mathcal{X} = \{(\mathbf{x}_n, \omega_n)\}_{n=1}^N \subset \mathbb{R}^p \times \Omega \quad (1)$$

em que

- p é a dimensão do vetor de atributos.
- C é o número de classes.
- $\Omega = \{\omega_1, \omega_2, \dots, \omega_C\}$ é um conjunto finito de C rótulos (não necessariamente numéricos) associados às C classes do problema.
- N é o número de exemplos de treinamento, ou seja, a cardinalidade de \mathcal{X} : $card(\mathcal{X}) = \#\mathcal{X} = N$

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Pausa pra revisão!! Vetores no \mathbb{R}^p

- Um vetor é um segmento de reta orientado que tem comprimento (norma) e um ângulo em relação a outro vetor ou alguma linha de referência, em geral, a linha horizontal.
- Para um vetor $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_p]^T$, sua norma é dada por

$$r = \|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2} \quad (2)$$

- O ângulo θ entre 2 vetores $\mathbf{x}, \mathbf{v} \in \mathbb{R}^p$ é determinado pelo produto escalar, também chamado de produto interno:

$$\mathbf{x} \cdot \mathbf{v} = \|\mathbf{x}\| \times \|\mathbf{v}\| \times \cos(\theta) \quad (3)$$

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Pausa pra revisão!! Vetores no \mathbb{R}^p

- De modo que chegamos em

$$\cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|} \quad \Rightarrow \quad \theta = \arccos\left(\frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|}\right) \quad (4)$$

em que $\mathbf{v} = [v_1 \ v_2 \ \cdots \ v_p]^T$ e $\mathbf{x} \cdot \mathbf{v} = \sum_{i=1}^p x_i v_i$.

- Por exemplo, o ângulo entre o vetor $\mathbf{x} = [1 \ 1]^T$ e o eixo horizontal pode ser calculado como

$$\theta = \arccos\left(\frac{(1).(1) + (1).(0)}{\sqrt{2} \times (1)}\right) = \arccos\left(\frac{1}{\sqrt{2}}\right) = 45^\circ \quad (5)$$

em que $\mathbf{v} = [1 \ 0]^T$ é o vetor de referência escolhido.

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Pausa pra revisão!! Vetores no \mathbb{R}^p

- Qual o ângulo entre o vetor $\mathbf{x} = [1 \ 2 \ 3]^T$ e o plano horizontal?
- Solução: Neste caso, o vetor de referência é $\mathbf{v} = [1 \ 2 \ 0]^T$. Assim, temos que

$$\theta = \arccos \left(\frac{(1).(1) + (2).(2) + (3).(0)}{\sqrt{14} \times \sqrt{5}} \right) \quad (6)$$

$$= \arccos \left(\frac{5}{\sqrt{70}} \right) \quad (7)$$

$$\approx 53^\circ \quad (8)$$

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Classificador Vizinho Mais Próximo (NN, sigla em Inglês)

Passo 1 - Armazenar em disco ou memória todo o conjunto \mathcal{X} , no formato adequado.

Passo 2 - Para cada novo vetor de atributos \mathbf{x}_{new} ainda não-classificado, realizar uma busca em \mathcal{X} pelo índice do vetor de atributos mais próximo de \mathbf{x} :

$$i^* = \arg \min_{i=1,\dots,N} dist(\mathbf{x}_{new}, \mathbf{x}_i) \quad (9)$$

em que $dist(\mathbf{x}_{new}, \mathbf{x}_i)$ é uma função que mede a distância entre os dois vetores \mathbf{x}_{new} e \mathbf{x}_i .

Passo 3 - Atribuir ao vetor \mathbf{x}_{new} à mesma classe que \mathbf{x}_{i^*} .

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Classificador Vizinho Mais Próximo (NN, sigla em Inglês)

- Alternativamente, os Passos 2 e 3 pode ser unificados e reescritos como um único passo da seguinte maneira:

Passo 2 Atribuir ao vetor \mathbf{x}_{new} a mesma classe que \mathbf{x}_{i^*} , se

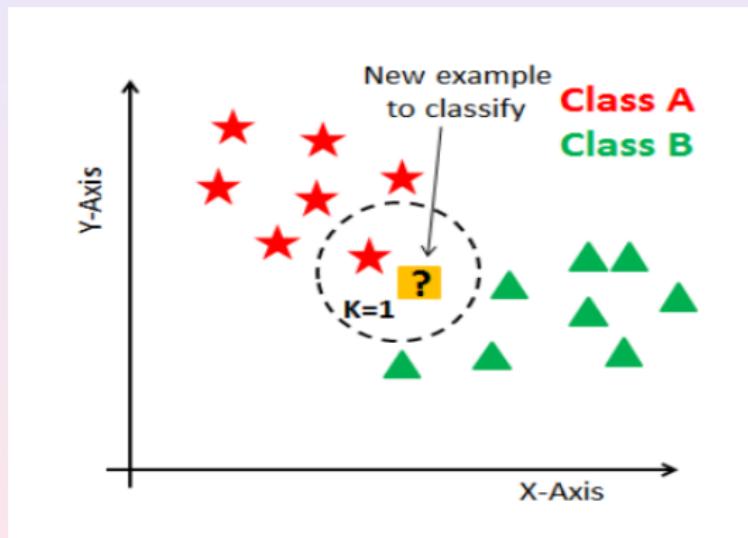
$$dist(\mathbf{x}_{new}, \mathbf{x}_{i^*}) < dist(\mathbf{x}_{new}, \mathbf{x}_i), \quad \forall i \neq i^*$$

(10)

- Nas Expressões (9) e (10) pode-se usar várias funções distância, tais como a distância quarteirão ou distância euclidiana.

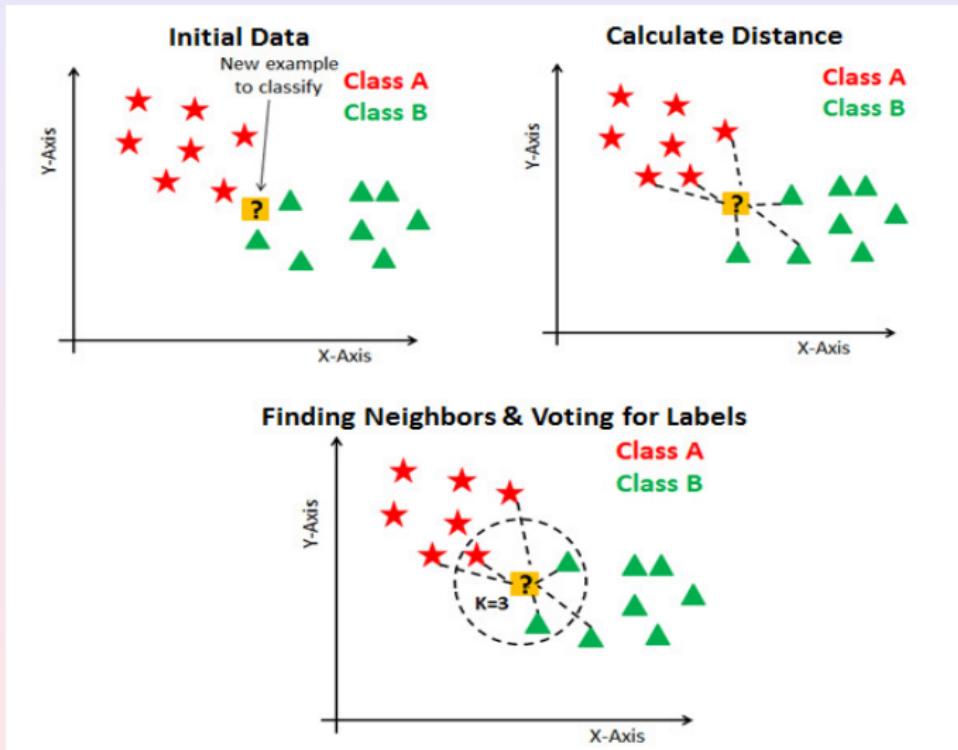
Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima



Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima



Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

- Considere dois vetores definidos como

$$\mathbf{x} = [0 \ 1 \ 2 \ 1 \ 0] \quad \text{e} \quad \mathbf{y} = [0 \ 0 \ 1 \ 0 \ 0] \quad (11)$$

- A distância quarteirão entre \mathbf{x} e \mathbf{y} é definida como

$$dist(\mathbf{x}, \mathbf{y}) = |0 - 0| + |1 - 0| + |2 - 1| + |1 - 0| + |0 - 0| = 3$$

- Já a distância euclidiana entre \mathbf{x} e \mathbf{y} é definida como

$$dist(\mathbf{x}, \mathbf{y}) = \sqrt{(0 - 0)^2 + (1 - 0)^2 + (2 - 1)^2 + (1 - 0)^2 + (0 - 0)^2} = \sqrt{3}$$

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

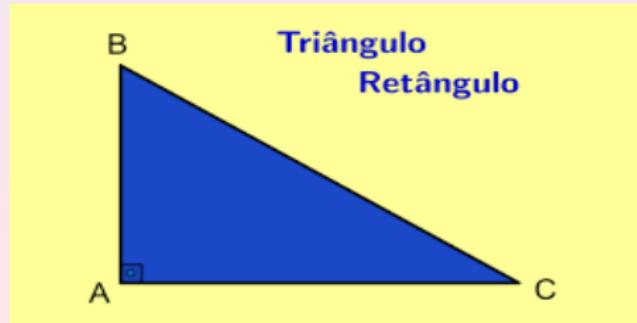
Distâncias Quarteirão e Euclidiana

- Embora tenhamos usado a operação de raiz quadrada na definição da distância euclidiana, ela não é necessária em classificação de padrões.
- Um primeiro motivo para não fazer isso é reduzir o custo computacional da operação de busca pela menor distância.
- Outro motivo é que a função distância euclidiana (com ou sem raiz quadrada) leva ao mesmo objeto mais próximo; ou seja, o mínimo da função $f(x) = |x|$ é o mesmo que o da função $g(x) = [f(x)]^2$.

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

- Relação Geométrica entre as Distâncias Quarteirão e Euclidiana.
- Distância eucliana entre **B** e **C** = comprimento do segmento \overline{BC} = comprimento da hipotenusa.
- Distância quarteirão entre **B** e **C** = soma dos comprimentos dos segmentos \overline{BA} e \overline{AC} = soma dos comprimentos dos catetos.



Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Classificador Centróide Mais Próximo (MDC, sigla em Inglês)

Passo 1 - Encontrar o vetor centróide de cada classe:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x} \quad (12)$$

em que N_i é o número de exemplos da i -ésima classe (cujo rótulo é ω_i), $i = 1, \dots, C$.

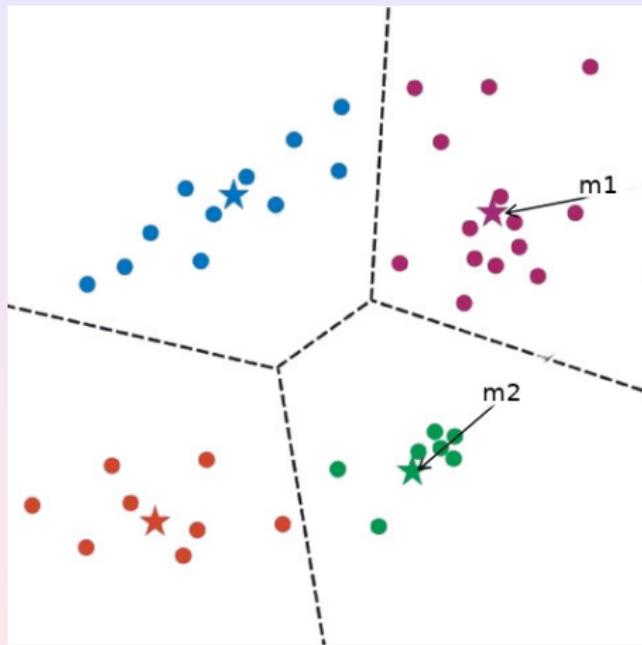
Passo 2 - Atribuir ao novo vetor de atributos \mathbf{x}_{new} o rótulo da classe de \mathbf{m}_{i^*} , se

$$dist(\mathbf{x}_{new}, \mathbf{m}_{i^*}) < dist(\mathbf{x}_{new}, \mathbf{m}_i), \quad \forall i \neq i^* \quad (13)$$

em que $dist(\mathbf{x}_{new}, \mathbf{m}_i)$ é uma função que mede a distância entre os dois vetores \mathbf{x}_{new} e \mathbf{m}_i .

Introdução à Classificação de Padrões

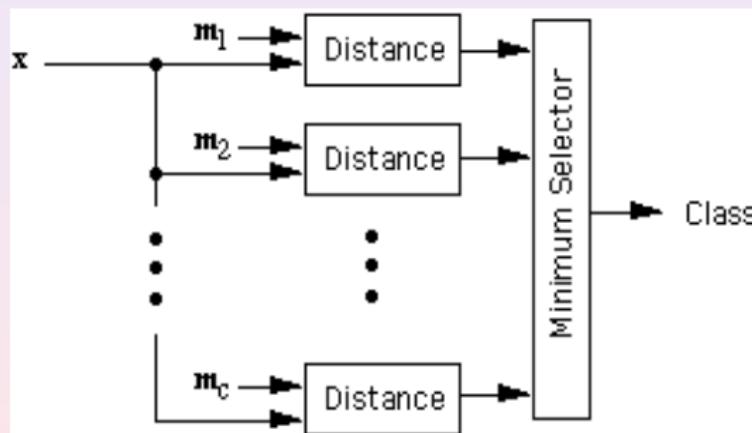
Classificadores Baseados em Distância Mínima



Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

- Diagrama de fluxo de sinais para o classificador MDC para um problema de C classes.



Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

- É possível notar que o cálculo de distâncias é muito importante para os dois classificadores mencionados anteriormente (NN e MDC).
- Em geral, utiliza-se norma euclidiana em seus algoritmos, mas existem várias outras normas que podem ser utilizadas.
- O conhecimento destas normas é de suma importância em reconhecimento de padrões, bem como as suas interpretações geométricas.
- A seguir vamos primeiro definir formalmente uma medida de dissimilaridade (i.e. distância) e suas propriedades, para depois listar algumas que são comumente usadas em classificação e clusterização de dados.

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Definição: Medidas de Dissimilaridade

- Se d_{rs} é uma medida de dissimilaridade entre os objetos^a r e s , então deve satisfazer as seguintes condições:
 - ① $d_{rs} \geq 0$, para todo r, s .
 - ② $d_{rr} = 0$, para todo r .
 - ③ $d_{rs} = d_{sr}$, para todo r, s .
- A condição 3 (simetria) nem sempre é satisfeita por certas medidas de dissimilaridade usuais (e.g. divergente de Kullback-Liebler).

^aTais objetos podem ser vetores, por exemplo. Ou nuvens (distribuições) de pontos.

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

- **Exemplo:** se a dissimilaridade entre dois lugares em uma cidade é medida pela distância percorrida por carro, então por causa de sistemas de mão única, a distância em um sentido pode ser mais longo do que a percorrida no outro.
- **Medidas de dissimilaridade** são aquelas que quanto mais próximos os objetos, menor são seus valores. Pra encurtar, dizemos que são medidas do tipo quanto menor, mais similar!
- **Medidas de similaridade** são aquelas que quanto mais próximos os objetos, maior são seus valores. Pra encurtar, dizemos que são medidas do tipo quanto maior, mais similar!

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Dissimilaridade vs. Similaridade

- Medidas de dissimilaridade podem ser transformadas em medidas de similaridade usando várias transformações.
- Se d_{rs} é uma medida de dissimilaridade entre os objetos r e s , uma medida de similaridade correspondente (s_{rs}) pode ser definida como

$$s_{rs} = \frac{1}{1 + d_{rs}} \quad (14)$$

ou como

$$s_{rs} = \max\{c - d_{rs}, 0\} \quad (15)$$

onde c é uma constante positiva.

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Dissimilaridade vs. Similaridade

- Distâncias assumem valores no intervalo $[0, \infty)$. Ou seja, possuem um limite inferior, mas não um limite superior. Com medidas de similaridade, verifica-se a existência de um intervalo para os valores de s_{rs} , normalmente entre 0 e 1.
- O limite inferior igual a 0, indica similaridade mínima. O limite superior igual a 1 indica similaridade máxima. Por exemplo, para a medida de similaridade

$$s_{rs} = \frac{1}{1 + d_{rs}} \quad (16)$$

temos as seguintes propriedades:

$$d_{rs} \rightarrow 0, \quad s_{rs} \rightarrow 1 \quad (17)$$

$$d_{rs} \rightarrow \infty, \quad s_{rs} \rightarrow 0 \quad (18)$$

Introdução à Classificação de Padrões

Classificadores Baseados em Distância Mínima

Definição: Métrica

- Se além das 3 condições discutidas anteriormente, a medida de dissimilaridade também satisfaz a desigualdade triangular

$$d_{rs} \leq d_{rt} + d_{ts}, \quad \forall r, s, t \quad (19)$$

então a medida de dissimilaridade é uma *métrica* e o termo *distância* é normalmente adotado.

- Em Matemática, métrica é um conceito que generaliza a ideia geométrica de distância.
- Um conjunto em que há uma métrica definida recebe o nome de espaço métrico.

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Definição: Distância de Minkowski de Ordem m

- A *distância de Minkowski* é uma medida de dissimilaridade bem geral, a partir da qual podemos obter algumas das medidas de dissimilaridade mais comuns em RP.
- Sejam dois vetores reais $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_p]^T$ e $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_p]^T$; ou seja, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$.
- A distância de Minkowski de ordem m ($m \geq 0$) entre \mathbf{x} e \mathbf{y} é definida pela seguinte expressão:

$$d_M^{(m)}(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^p |x_j - y_j|^m \right)^{\frac{1}{m}} \quad (20)$$

onde x_j e y_j são as j -ésimas componentes de \mathbf{x} e \mathbf{y} , respectivamente.



Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Se $m = 1 \Rightarrow$ Distância Quarteirão

- Se fizermos $m = 1$ na Equação (20) obtemos a expressão da distância quarteirão (*city block*, em Inglês):

$$d_{cb}(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| \quad (21)$$

$$= \sum_{j=1}^p |x_j - y_j| \quad (22)$$

$$= |x_1 - y_1| + \cdots + |x_p - y_p| \quad (23)$$

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Se $m = 2 \Rightarrow$ Distância Euclidiana

- Se fizermos $m = 2$ na Equação (20) obtemos a conhecida expressão da distância euclidiana:

$$d_e(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| \quad (24)$$

$$= \left(\sum_{j=1}^p |x_j - y_j|^2 \right)^{\frac{1}{2}} \quad (25)$$

$$= \sqrt{\sum_{j=1}^p |x_j - y_j|^2} \quad (26)$$

$$= \sqrt{|x_1 - y_1|^2 + \cdots + |x_p - y_p|^2} \quad (27)$$

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Se $m \rightarrow \infty \Rightarrow$ Distância de Chebyshev

- Se fizermos $m \rightarrow \infty$ na Equação (20) obtemos a expressão da distância de Chebyshev:

$$d_{ch}(\mathbf{x}, \mathbf{y}) = \lim_{m \rightarrow \infty} \left(\sum_{j=1}^p |x_j - y_j|^m \right)^{\frac{1}{m}} \quad (28)$$

$$= \max\{|x_1 - y_1|, \dots, |x_p - y_p|\} \quad (29)$$

$$= \max_j |x_j - y_j| \quad (30)$$

Desafio!

Demonstrar a passagem da expressão mostrada na Eq. (15) para a da Eq. (16), computacional e matematicamente.

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

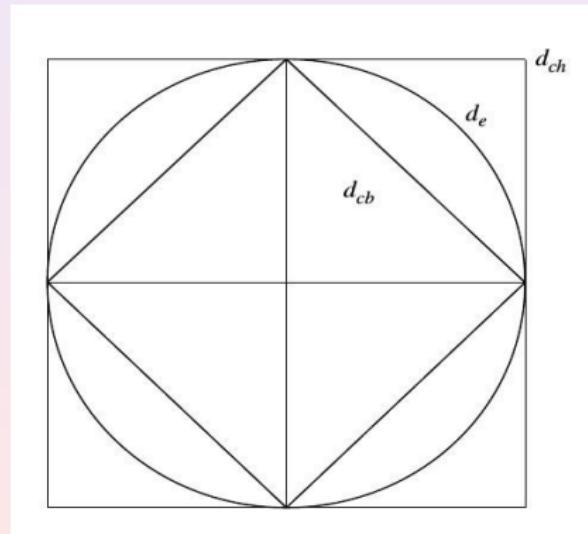
Por que tantas medidas de dissimilaridade?

- Como vimos, existem muitas medidas de dissimilaridade e continuam sendo propostas novas a cada momento.
- A escolha de uma métrica particular depende da aplicação.
- Para fins de seleção/extracão de atributos deve-se escolher a métrica que leva ao melhor desempenho do classificador.
- Caso as métricas conduzam a desempenhos semelhantes, a escolha de qual usar pode ser feita com base no custo computacional para sua implementação.

Introdução à Classificação de Padrões

Lugares Geométricos (curvas de contorno)

- A escolha da métrica deve levar em conta também a organização espacial (i.e. distribuição) dos dados.
- A figura abaixo ilustra as curvas de contorno geradas para as distâncias quarteirão, euclidiana e de Chebyshev.



Introdução à Classificação de Padrões

Lugares Geométricos (curvas de contorno)

Lugar geométrico: Distância Euclidiana

- Para os desenvolvimentos a seguir vamos supor $\mathbf{x} = [x_1 \ x_2]^T \equiv (x_1, x_2)$ e $\mathbf{y} = [0 \ 0]^T \equiv (0, 0)$.
- Lugar geométrico é uma curva ou conjunto dos pontos (x_1, x_2) que estão a uma mesma distância r do ponto $(0, 0)$.
- Assim, podemos escrever

$$\|\mathbf{x} - \mathbf{y}\| = r \Rightarrow \|\mathbf{x} - \mathbf{y}\|^2 = r^2 \quad (31)$$

$$(x_1 - 0)^2 + (x_2 - 0)^2 = r^2 \Rightarrow x_1^2 + x_2^2 = r^2. \quad (32)$$

- Sem perda de generalidade, podemos fazer $r = 1$ e escrever $x_1^2 + x_2^2 = 1$, que é a expressão do círculo de raio unitário com centro na coordenada $(0, 0)$.

Introdução à Classificação de Padrões

Lugares Geométricos (curvas de contorno)

Lugar geométrico: Distância Quarteirão

- Qual seria o lugar geométrico dos pontos (x_1, x_2) que estão a uma mesma distância quarteirão r do ponto $(0, 0)$?
- De um modo geral, tais pontos devem satisfazer a seguinte expressão:

$$|\mathbf{x} - \mathbf{y}| = r \Rightarrow |x_1 - 0| + |x_2 - 0| = r \quad (33)$$

$$|x_1| + |x_2| = r \Rightarrow |x_1| + |x_2| = 1. \quad (34)$$

- A depender do quadrante, chegamos a 4 equações de reta.
 - ① Quadrante 1 ($x_1 > 0, x_2 > 0$): $x_1 + x_2 = 1 \Rightarrow x_2 = -x_1 + 1$.
 - ② Quadrante 2 ($x_1 < 0, x_2 > 0$): $-x_1 + x_2 = 1 \Rightarrow x_2 = x_1 + 1$.
 - ③ Quadrante 3 ($x_1 < 0, x_2 < 0$): $-x_1 - x_2 = 1 \Rightarrow x_2 = -x_1 - 1$.
 - ④ Quadrante 4 ($x_1 > 0, x_2 < 0$): $x_1 - x_2 = 1 \Rightarrow x_2 = x_1 - 1$.

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Distância Quadrática ou Distância Ponderada

- A distância quadrática, também chamada de distância ponderada (*weighted distance*), é muito utilizada em RP e mineração de dados.
- Sua expressão geral é dada por

$$d_q(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \begin{bmatrix} \mathbf{Q} \\ [1 \times p] & [p \times p] & [p \times 1] \end{bmatrix} (\mathbf{x} - \mathbf{y})} \quad (35)$$

$$= \sqrt{\sum_{i=1}^p \sum_{j=1}^p (x_i - y_i) Q_{ij} (x_j - y_j)} \quad (36)$$

onde $\mathbf{Q} = [Q_{ij}]_{p \times p}$ é uma matriz $p \times p$ positiva semidefinida.

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Distância Quadrática: Algumas considerações

- A matriz \mathbf{Q} é positiva semidefinida se, para qualquer vetor \mathbf{v} , obtivermos sempre

$$\begin{matrix} \mathbf{v}^T \\ [1 \times p] \end{matrix} \mathbf{Q} \begin{matrix} \mathbf{v} \\ [p \times 1] \end{matrix} \geq 0 \quad (37)$$

- Assim, a matriz \mathbf{Q} deve ser positiva semidefinida para que o termo $(\mathbf{x} - \mathbf{y})^T \mathbf{Q} (\mathbf{x} - \mathbf{y})$ seja positivo ou nulo, garantindo que não haja argumentos negativos na função raiz quadrada na Eq. (35).
- Uma forma fácil de verificar se uma matriz é positiva semidefinida se dá por meio dos seus autovalores, que devem ser positivos ou nulos (i.e., $\lambda_i \geq 0$, $i = 1, \dots, p$).

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Por que a distância quadrática tem esse nome?

- Vamos considerar que os vetores \mathbf{x} e \mathbf{y} tem dimensão $p = 2$, podendo ser representados como

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{e} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad (38)$$

- Vamos também precisar definir a matriz \mathbf{Q} como

$$\mathbf{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad (39)$$

em que a, b, c e d são números reais.

Introdução à Classificação de Padrões

Medidas de Dissimilaridade

Por que a distância quadrática tem esse nome? (cont.)

Assim, substituindo na Eq. (35), temos

$$\begin{aligned} d_q(\mathbf{x}, \mathbf{y}) &= \sqrt{[x_1 - y_1 \ x_2 - y_2] \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}}, \quad (40) \\ &= \sqrt{[x_1 - y_1 \ x_2 - y_2] \begin{bmatrix} a(x_1 - y_1) + b(x_2 - y_2) \\ c(x_1 - y_1) + d(x_2 - y_2) \end{bmatrix}}, \end{aligned}$$

que resulta na seguinte expressão:

$$d_q(\mathbf{x}, \mathbf{y}) = \sqrt{a(x_1 - y_1)^2 + (b + c)(x_1 - y_1)(x_2 - y_2) + d(x_2 - y_2)^2}$$

onde se observa a presença de termos elevados ao quadrado.

Introdução à Classificação de Padrões

Relação entre as distâncias quadrática e euclidiana

- Se \mathbf{Q} for igual à matriz identidade, tem-se

$$\mathbf{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (41)$$

em que $a = d = 1$ e $b = c = 0$. Assim, a distância quadrática reduz-se a

$$\begin{aligned} d_q(\mathbf{x}, \mathbf{y}) &= \sqrt{a(x_1 - y_1)^2 + (b + c)(x_1 - y_1)(x_2 - y_2) + d(x_2 - y_2)^2}, \\ &= \sqrt{(1)(x_1 - y_1)^2 + (0 + 0)(x_1 - y_1)(x_2 - y_2) + (1)(x_2 - y_2)^2}, \\ &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}, \\ &= d_e(\mathbf{x}, \mathbf{y}). \end{aligned}$$

- De onde se conclui que a distância euclidiana é um caso particular da quadrática.

Introdução à Classificação de Padrões

Relação entre as distâncias quadrática e euclidiana

Versão matricial do exemplo anterior

- Se fizermos $\mathbf{Q} = \mathbf{I}_p$ na Eq. (35), obtemos a seguinte expressão:

$$d_q(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{I}_p (\mathbf{x} - \mathbf{y})} \quad (42)$$

$$= \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} \quad (43)$$

$$= \sqrt{\|\mathbf{x} - \mathbf{y}\|^2} \quad (44)$$

$$= \|\mathbf{x} - \mathbf{y}\|. \quad (45)$$

- De onde podemos concluir que, neste caso particular, a distância quadrática reduz-se à distância euclidiana.

Introdução à Classificação de Padrões

Outras escolhas comuns para a matriz \mathbf{Q}

- Se \mathbf{Q} for igual à seguinte matriz:

$$\mathbf{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}, \quad (46)$$

em que $a = 1/\sigma_1^2$, $b = 1/\sigma_2^2$ e $b = c = 0$, chegamos a

$$\begin{aligned} d_q(\mathbf{x}, \mathbf{y}) &= \sqrt{a(x_1 - y_1)^2 + (b + c)(x_1 - y_1)(x_2 - y_2) + d(x_2 - y_2)^2}, \\ &= \sqrt{\frac{1}{\sigma_1^2}(x_1 - y_1)^2 + (0 + 0)(x_1 - y_1)(x_2 - y_2) + \frac{1}{\sigma_2^2}(x_2 - y_2)^2}, \\ &= \sqrt{\left(\frac{x_1 - y_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - y_2}{\sigma_2}\right)^2}. \end{aligned}$$

- De onde se nota que esta escolha de \mathbf{Q} normaliza automaticamente cada atributo pelo seu desvio padrão.

Parte I

Normalização dos Dados

Introdução à Classificação de Padrões

Normalização dos Dados

Objetivos

Objetivo: Entender a necessidade de equalizar as ordens de grandeza dos atributos usados em um problema de classificação/clusterização.

- **Método 1:** Manter constante a norma dos vetores.
- **Método 2:** Mudança da escala original para os intervalos [0, 1] ou [-1,+1].
- **Método 3:** Padronização z -score (i.e. média=0, variância=1).
- **Método 4:** Padronização Robusta (i.e. mediana=0, iqr=1).

Introdução à Classificação de Padrões

Normalização dos Dados

Método 1: Norma Constante

- Uma das técnicas mais simples de normalização consiste em manter constantes e iguais a 1 as normas dos vetores de atributos \mathbf{x} e dos centróides \mathbf{m}_i .
- Este procedimento deve ser aplicado a todos os vetores de atributos e todas os centróides.
- Para isso, basta dividir cada vetor por sua respectiva norma euclidiana:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \quad \text{e} \quad \tilde{\mathbf{m}}_i = \frac{\mathbf{m}_i}{\|\mathbf{m}_i\|} \quad (47)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Método 1: Norma Constante

- Por exemplo, considere o seguinte vetor, que não possui norma unitária:

$$\mathbf{x} = \begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} \quad (48)$$

- A norma deste vetor é calculada como

$$\|\mathbf{x}\| = \sqrt{(\sqrt{3})^2 + (3)^2 + (-2)^2} = \sqrt{16} = 4. \quad (49)$$

- Assim, a versão normalizada do vetor \mathbf{x} é dada por

$$\tilde{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{1}{4} \begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} \sqrt{3}/4 \\ 3/4 \\ -1/2 \end{bmatrix} \quad (50)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Propriedades do Método 1: Norma Constante

- A normalização descrita no slide anterior não altera a direção do vetor, apenas muda seu comprimento.
- Em outras palavras, o vetor resultante é um múltiplo do vetor original conforme pode ser visto na operação a seguir.

$$\tilde{\mathbf{x}} = \frac{1}{\|\mathbf{x}\|} \mathbf{x} = \alpha \mathbf{x}, \quad (51)$$

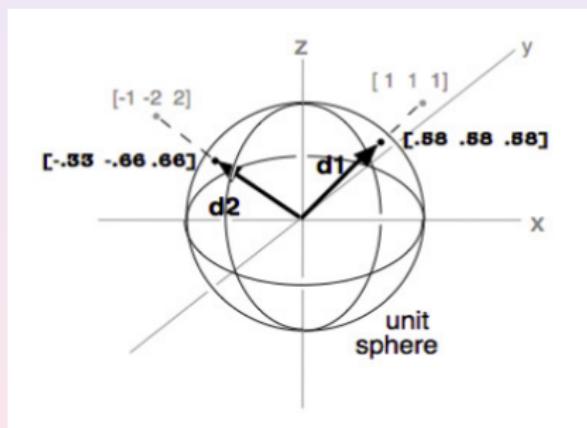
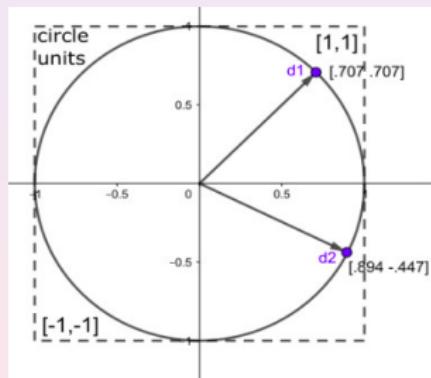
em que $\alpha = 1/\|\mathbf{x}\|$ é uma constante positiva.

- Note que a normalização assim realizada depende apenas dos valores das componentes do vetor sendo normalizado.
- Assim, chamaremos este tipo de procedimento de normalização local.

Introdução à Classificação de Padrões

Normalização dos Dados

- Método 1: Interpretação Geométrica



Introdução à Classificação de Padrões

Normalização dos Dados

Propriedades do Método 1: Norma Constante

- A similaridade entre 2 vetores $\mathbf{x}, \mathbf{v} \in \mathbb{R}^p$ de norma unitária pode ser calculada pelo cosseno do ângulo entre eles.
- A partir da fórmula do produto escalar,
$$\mathbf{x} \cdot \mathbf{v} = \|\mathbf{x}\| \times \|\mathbf{v}\| \times \cos(\theta)$$
, chega-se à seguinte expressão:

$$s(\mathbf{x}, \mathbf{v}) = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{v}}{\|\mathbf{x}\| \times \|\mathbf{v}\|} \quad (52)$$

$$= \frac{\mathbf{x} \cdot \mathbf{v}}{1 \times 1} = \mathbf{x}^T \mathbf{v} = \sum_{j=1}^p x_j v_j. \quad (53)$$

- Resumo: A similaridade entre dois vetores de norma unitária é computada pelo produto escalar entre eles.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 1: Norma constante

- A normalização pelo Método 1 é particularmente útil para o classificador de máxima correlação (MC).
- O classificador MC nada mais é do que uma implementação dos classificadores de distância mínima (NN ou DMC) em que a medida de dissimilaridade é substituída por uma medida de similaridade, no caso, o produto escalar.
- O algoritmo do classificador MC é apresentado no próximo slide.

Introdução à Classificação de Padrões

Normalização dos Dados

Classificador de Máxima Correlação

Passo 1 - Encontrar o vetor centróide de cada classe:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x} \quad (54)$$

em que N_i é o número de exemplos da i -ésima classe (cujo rótulo é ω_i), $i = 1, \dots, C$.

Passo 2 - Atribuir um novo vetor de atributos \mathbf{x}_{new} à mesma classe que \mathbf{m}_{i^*} , se

$$\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} > \tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new}, \quad \forall i \neq i^* \quad (55)$$

em que $\tilde{\mathbf{m}}_i = \mathbf{m}_i / \|\mathbf{m}_i\|$ e $\tilde{\mathbf{x}}_{new} = \mathbf{x}_{new} / \|\mathbf{x}_{new}\|$ são as versões de norma unitária de \mathbf{m}_i e \mathbf{x}_{new} , respectivamente.

Introdução à Classificação de Padrões

Normalização dos Dados

Sobre Equivalência entre os classificadores DMC e MC

- O classificador MC nada mais é do que uma implementação dos classificadores de distância mínima (NN ou DMC) em que a medida de dissimilaridade é substituída por uma medida de similaridade, no caso, o produto escalar.
- Esta equivalência é fácil de mostrar a partir de um resultado muito conhecida da álgebra linear: $\|\mathbf{v}\|^2 = \mathbf{v}^T \mathbf{v}$.
- Assim, considere a desigualdade da Eq. (13), em que a função genérica $dist(\cdot, \cdot)$ é instanciada pela distância euclidiana quadrática.

Introdução à Classificação de Padrões

Normalização dos Dados

Sobre Equivalência entre os classificadores DMC e MC

Assim, tem-se que

$$dist(\mathbf{x}_{new}, \mathbf{m}_i) = \|\mathbf{x}_{new} - \mathbf{m}_i\|^2, \quad (56)$$

$$= (\mathbf{x}_{new} - \mathbf{m}_i)^T (\mathbf{x}_{new} - \mathbf{m}_i), \quad (57)$$

$$= \mathbf{x}_{new}^T \mathbf{x}_{new} - \mathbf{x}_{new}^T \mathbf{m}_i - \mathbf{m}_i^T \mathbf{x}_{new} + \mathbf{m}_i^T \mathbf{m}_i, \quad (58)$$

$$= \mathbf{x}_{new}^T \mathbf{x}_{new} - 2\mathbf{m}_i^T \mathbf{x}_{new} + \mathbf{m}_i^T \mathbf{m}_i, \quad (59)$$

$$= \|\mathbf{x}_{new}\|^2 - 2\mathbf{m}_i^T \mathbf{x}_{new} + \|\mathbf{m}_i\|^2, \quad (60)$$

tal que, para $\|\mathbf{x}_{new}\| = \|\mathbf{m}_i\| = 1$, resulta em

$$dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_i) = 2 - 2\tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new}. \quad (61)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Sobre Equivalência entre os classificadores DMC e MC

Substituindo a Eq. (61) na Eq. (13), tem-se que

$$\begin{aligned} dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_{i^*}) &< dist(\tilde{\mathbf{x}}_{new}, \tilde{\mathbf{m}}_i) \\ 2 - 2\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &< 2 - 2\tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new} \\ -2\tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &< -2\tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new} \\ \tilde{\mathbf{m}}_{i^*}^T \tilde{\mathbf{x}}_{new} &> \tilde{\mathbf{m}}_i^T \tilde{\mathbf{x}}_{new} \end{aligned}$$

sendo a última desigualdade a regra de decisão do classificador de máxima correlação.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 2: Mudança de escala

- Para classificadores baseados em distância euclidiana, uma normalização que promove uma mudança na escala das variáveis, é mais comum.
- Este procedimento é realizado variável a variável e requer a determinação do valor mínimo (x_{min}) e do valor máximo (x_{max}) da variável sendo normalizada.
- Por isso, chamaremos este tipo de procedimento de normalização global.
- Este tipo de normalização torna a variável adimensional.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 2: Mudança de escala para o intervalo $[0,1]$

$$\tilde{x}_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}, \quad j = 1, \dots, p \quad (62)$$

com $\max(x_j)$ e $\min(x_j)$ sendo os valores máximo e mínimo, respectivamente, do atributo x_j no conjunto de dados.

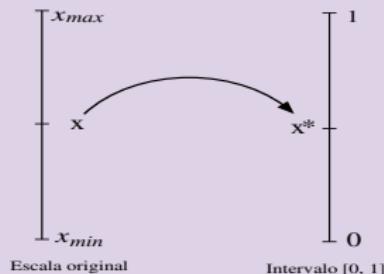


Figura: Mudança da escala do atributo x_j para o intervalo $[0,1]$.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 2: Exemplo 1

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\min(x_1)=11,03$ e $\max(x_1)=14,83$.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} = \frac{x_1 - 11,03}{14,83 - 11,03} = \frac{x_1 - 11,03}{3,80} \quad (63)$$

- Assim, a observação $x_1=13,50$ na escala original, terá o seguinte valor no intervalo $[0, 1]$:

$$\tilde{x}_1 = \frac{13,50 - 11,03}{3,80} = 0,65. \quad (64)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Método 2: Mudança de escala para o intervalo $[-1, +1]$

$$\tilde{x}_j = 2 \left(\frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)} \right) - 1 \quad (65)$$

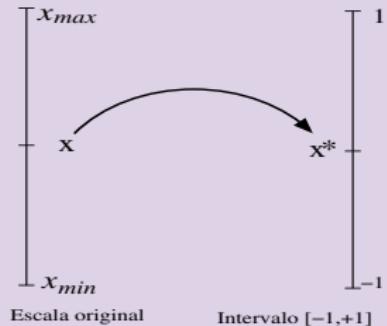


Figura: Mudança da escala original de x_j para o intervalo $[-1, +1]$.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 2: Exemplo 2

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\min(x_1)=11,03$ e $\max(x_1)=14,83$.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = 2 \left(\frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)} \right) - 1 = 2 \left(\frac{x_1 - 11,03}{3,80} \right) - 1 \quad (66)$$

- Assim, a observação $x_1=13,50$ na escala original, terá o seguinte valor no intervalo $[-1, +1]$:

$$\tilde{x}_1 = 2 \left(\frac{13,50 - 11,03}{3,80} \right) - 1 = 0,30. \quad (67)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Método 3: Padronização da variável ($\text{média}=0$, $\text{variância}=1$)

- Assim como as normalizações descritas no Método 2, devemos aplicar a padronização às variáveis do problema, uma a uma.
- Este tipo de normalização requer o cálculo da média ($\hat{\mu}_j$) e do desvio-padrão ($\hat{\sigma}_j$) da variável x_j .
- Por isso, a padronização também pode ser chamada de normalização estatística, normalização pelo desvio-padrão, ou ainda normalização *z-score*.
- Este procedimento também é um tipo de normalização global.
- Este tipo de normalização torna a variável adimensional.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 3: Padronização z -score (média=0, variância=1)

- A normalização estatística é dada por

$$\tilde{x}_j = \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j} \quad (68)$$

com a média e o desvio-padrão amostrais de x_j calculados como

$$\hat{\mu}_j = \frac{\sum_{n=1}^N x_j(n)}{N} \quad \text{e} \quad \hat{\sigma}_j = \sqrt{\left(\frac{\sum_{n=1}^N (x_j(n) - \hat{\mu}_j)^2}{N - 1} \right)} \quad (69)$$

tal que $x_j(n)$ é a n -ésima observação de x_j e N é o número total de observações de x .

Introdução à Classificação de Padrões

Normalização dos Dados

Método 3: Exemplo numérico

- Usando o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\hat{\mu}_1=13,00$ e $\hat{\sigma}_1=0,81$.
- Assim, a função de normalização é dada por

$$\tilde{x}_1 = \frac{x_1 - 13,00}{0,81} \quad (70)$$

- Assim, a observação $x_1=13,50$ na escala original, terá o seguinte valor padronizado:

$$\tilde{x}_1 = \frac{13,50 - 13,00}{0,81} = 0,617. \quad (71)$$

Introdução à Classificação de Padrões

Normalização dos Dados

Método 4: Padronização robusta (mediana=0, iqr=1)

- Esta normalização usa estatísticas robustas, como mediana e o intervalo interquartil^a (IQR):

$$\tilde{x}_j = \frac{x_j - \text{mediana}(x_j)}{\text{IQR}(x_j)} \quad (72)$$

em que a mediana é uma estatística robusta de tendência central, enquanto o IQR é uma medida robusta de dispersão das medidas.

- Maiores detalhes sobre o IIQ em
https://pt.wikipedia.org/wiki/Amplitude_interquartil

^aPor vezes chamado de *amplitude* ou *faixa interquartil*.

Introdução à Classificação de Padrões

Normalização dos Dados

Método 4: Padronização robusta (mediana=0, iqr=1)

- No Octave/Matlab, a mediana de um conjunto de medidas de um atributo x_j pode ser estimada de 2 diferentes maneiras por meio dos seguintes comandos:
» `med1=median(x);`
» `med2=prctile(x,50);`
- De modo similar, o iqr para um conjunto de medidas de x_j pode ser estimado pelos seguintes comandos:
» `iqr1=iqr(x);`
» `iqr2=prctile(x,75)-prctile(x,25);`

Introdução à Classificação de Padrões

Normalização dos Dados

Normalização como Transformação Linear

- As técnicas para normalização de variáveis descritas anteriormente podem ser vistas como uma transformação linear aplicada à variável original.
- Por exemplo, a normalização via Método 2 pode ser escrita da seguinte forma:

$$\begin{aligned}\tilde{x}_j &= \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)} \\ &= \left(\frac{1}{\max(x_j) - \min(x_j)} \right) x_j - \left(\frac{\min(x_j)}{\max(x_j) - \min(x_j)} \right) \\ &= ax_j + b\end{aligned}$$

em que $a = \frac{1}{\max(x_j) - \min(x_j)}$ e $b = -\frac{\min(x_j)}{\max(x_j) - \min(x_j)}$.

Introdução à Classificação de Padrões

Normalização dos Dados

Propriedade 1 dos Métodos de Normalização

- Por serem transformações lineares, as normalizações descritas anteriormente não alteram a distribuição da variável normalizada em relação à variável original não-normalizada.
- Em outras palavras, o tipo de distribuição da variável permanece o mesmo. Por exemplo, se for gaussiana, continua gaussiana após a transformação.
- Os parâmetros da distribuição podem mudar, mas a forma dela não.
- Este resultado é suportado por um resultado teórico muito importante, que discutiremos a seguir.

Introdução à Classificação de Padrões

Normalização dos Dados

Propriedade 1 dos Métodos de Normalização

- Seja $x \in \mathbb{R}$ uma variável aleatória contínua, de média μ_x e variância σ_x^2 , cuja densidade de probabilidade é $f_X(x)$.
- Seja $y \in \mathbb{R}$ a variável aleatória resultante de uma operação matemática sobre x : $y = g(x)$.
- Demonstra-se que densidade de probabilidade de y é dada por

$$f_Y(y) = \frac{f_X(x)}{\left| \frac{dy}{dx} \right|} \quad (73)$$

- Assim, para $y = ax + b$, com a e b constantes reais, tem-se $|dy/dx| = |a|$ e $f_Y(y) = |a|f_X(x)$.
- Além disso, temos que $\mu_y = E[y] = E[ax + b] = aE[x] + b = a\mu_x + b$. E também $\sigma_y^2 = a^2\sigma_x^2$ (demonstrar!)

Introdução à Classificação de Padrões

Normalização dos Dados

Propriedade 2 dos Métodos de Normalização

- Como estas técnicas de normalização só utilizam estatísticas descritivas (min, max, média e desvio-padrão) das variáveis, tomadas individualmente, a correlação entre duas variáveis quaisquer permanece a mesma antes e depois da normalização.
- Variáveis normalizadas pelos métodos descritos anteriormente serão adimensionais.

Introdução à Classificação de Padrões

Normalização dos Dados

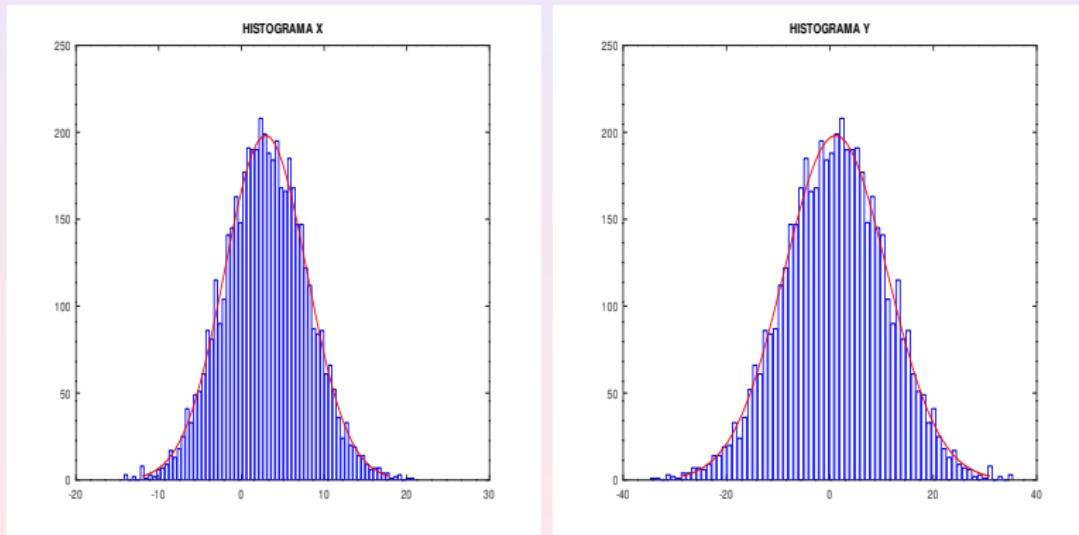
Verificação da Propriedade 1: Código Octave/Matlab

```
> mtx=3; stx=5; % estatisticas teoricas de x
> x=normrnd(mtx,stx,5000,1); % gera 5000 observacoes N(mtx,stx2)
> STATSx=[mean(x) std(x)] % estatisticas amostrais de x
STATSx = 2.9447    4.9757
> a=-2; b=7; % parametros da transformacao linear
> y=a*x+b; % aplica transf. linear a x
> figure; histfit(x); % histograma de x
> figure; histfit(y); % histograma de y
> mty=a*mtx+b, sty=abs(a)*stx % estatisticas teoricas de y
mty= 1
sty= 10
> STATSy=[mean(y) std(y)] % estatisticas amostrais de y
STATSy = 1.1105    9.9514
```

Introdução à Classificação de Padrões

Normalização dos Dados

- Propriedade 1: Histogramas de X e Y .



Introdução à Classificação de Padrões

Normalização dos Dados

Verificação da Propriedade 2: Código Octave/Matlab

```
» Cd=[4 2.8;2.8 9]; % matriz de covar desejada
» x=normrnd(0,1,5000,2); % gera 5000 observacoes 2 VA's
» A=chol(Cd); % gera matriz de mistura
» z=x*A; % gera VA's correlacionadas
» z1=z(:,1); z2=z(:,2);
» r12=corr(z1,z2) % correlacao entre z1 e z2
r12 = 0.45630
» z1n=(z1-mean(z1))/std(z1); % padroniza z1
» z2n=(z2-mean(z2))/std(z2); % padroniza z2
» STATSz1n=[mean(z1n) std(z1n)] % estatisticas de z1
STATSz1n = 1.6742e-17    1.0000e+00
» STATSz2n=[mean(z2n) std(z2n)] % estatisticas de z1
STATSz2n = -3.9257e-17    1.0000e+00
» r12n=corr(z1n,z2n) % correlacao entre z1 e z2
r12n = 0.45630
```

Introdução à Classificação de Padrões

Normalização dos Dados

Implementação do Método 3 no Excel e LibreOffice Calc

- Dadas N observações conjuntas de um atributo qualquer, a normalização estatística (ou padronização) podem ser facilmente implementada em planilhas numéricas.
 - No Excel, usar os comandos PADRONIZAR ou NORMALIZAR.
 - No LibreOffice Calc, usar o comando PADRONIZAR.

Parte II

Distribuição Gaussiana Multivariada

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Multivariada

FDP Gaussiana Multivariada

- A função densidade de probabilidade gaussiana (ou normal) multivariada é definida pela seguinte expressão:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\mathbf{C}_{\mathbf{x}}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

em que $|\mathbf{C}_{\mathbf{x}}|$ e $\mathbf{C}_{\mathbf{x}}^{-1}$ denotam, respectivamente, o determinante e a inversa da matriz de covariância.

- Escrevemos que um vetor aleatório $\mathbf{x} \in \mathbb{R}^p$ está distribuído segundo uma normal multivariada da seguinte forma:

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{C}_{\mathbf{x}})$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Multivariada

FDP Gaussiana Multivariada

- A Eq. (83) faz uso de vetores e matriz para simplificar a escrita e, por isso, parece um tanto difícil de ler. Porém, esta expressão é muito similar à da normal univariada. É tudo uma questão de como escrever as duas.
- **FDP Gaussiana uni-variada:**

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \\ &= \frac{1}{(2\pi)^{1/2}(\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right\} \end{aligned} \quad (74)$$

- **FDP Gaussiana p -variada:**

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{C}_{\mathbf{x}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$



FDP Gaussiana Multivariada (cont.-1)

- Alternativamente pode-se escrever a FDP gaussiana multivariada como

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\mathbf{C}_{\mathbf{x}}|^{1/2}} \exp \left\{ -\frac{1}{2} Q(\mathbf{x}) \right\}$$

em que

$$Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

em que a raiz quadrada de $Q(\mathbf{x})$ é chamada de *Distância de Mahalanobis*.

Variáveis Aleatórias Multidimensionais

Matriz de Covariância

- Costuma-se dispor as covariâncias entre todas as variáveis, tomadas duas a duas, σ_{ij} , $i, j = 1, \dots, p$, em uma *matriz de covariância*, $\mathbf{C}_x = [\sigma_{ij}]_{p \times p}$:

$$\mathbf{C}_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \cdots & \sigma_p^2 \end{bmatrix}$$

- Os elementos da diagonal principal desta matriz são as variâncias dos p atributos envolvidas no problema.
- Enquanto fora da diagonal principal tem-se as covariâncias das variáveis X_i e X_j , para $i \neq j$.

Variáveis Aleatórias Multidimensionais

Matriz de Covariância

- A matriz de covariância também pode ser escrita como

$$\mathbf{C}_x = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1}\sigma_p\sigma_1 & \cdots & \cdots & \sigma_p^2 \end{bmatrix}$$

se lembrarmos que o coeficiente de correlação entre X_i e X_j é dado por

$$\boxed{\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j}} \quad (75)$$

de onde tiramos que a covariância entre X_i e X_j pode ser expressa como

$$\boxed{\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j} \quad (76)$$

Variáveis Aleatórias Multidimensionais

Matriz de Covariância

Caso particular: $\rho_{ij} = 0$

- Quando as correlações entre os atributos são todas nulas (i.e. $\rho_{ij} = 0, \forall i, j$), a matriz de covariância correspondente é diagonal:

$$\mathbf{C_x} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}_{p \times p} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$$

Variáveis Aleatórias Multidimensionais

Matriz de Covariância

Caso particular: $\rho_{ij} = 0$

- Para este caso particular, o determinante da matriz é facilmente calculado pelo produto dos elementos da diagonal principal:

$$|\mathbf{C}_x| = \sigma_1^2 \cdot \sigma_2^2 \cdots \sigma_p^2 = \prod_{i=1}^p \sigma_i^2 \quad (77)$$

- A inversa de tal matriz também é facilmente calculada

$$\mathbf{C}_x^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_p^2} \end{bmatrix}_{p \times p} = \text{diag}\left(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots, \frac{1}{\sigma_p^2}\right)$$

Introdução à Classificação de Padrões

Matriz de Covariância

Propriedades da Matriz de Covariância

- ① A matriz \mathbf{C}_x é *simétrica*, pois $\sigma_{ij} = \sigma_{ji}$:

$$\mathbf{C}_x = \mathbf{C}_x^T$$

- ② A matriz \mathbf{C}_x é *definida positiva*:

$$\mathbf{x}^T \mathbf{C}_x \mathbf{x} > 0,$$

para qualquer vetor \mathbf{x} real e não-nulo.

- ③ Os autovalores λ_i de \mathbf{C}_x são sempre positivos.
- ④ O determinante de \mathbf{C}_x é sempre positivo.

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada

- Considerando $p = 2$, o vetor aleatório e o vetor de médias passam a ser representados como:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ e } \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (78)$$

- Logo, a matriz de covariância \mathbf{C}_x reduz-se a uma matriz quadrada de ordem 2:

$$\mathbf{C}_x = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}_{2 \times 2}$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada (cont.-1)

- Uma forma alternativa da matriz de covariância para o caso bidimensional é dada por

$$\mathbf{C}_x = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (79)$$

em que ρ é chamado de *Coeficiente de Correlação* entre x_1 e x_2 :

$$\boxed{\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}} \quad (80)$$

em que σ_1 e σ_2 são os desvios-padrões de x_1 e x_2 , respectivamente.

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada (cont.-2)

- Assim, o determinante da matriz de covariância é dado por

$$\begin{aligned} |\mathbf{C}_x| &= \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} \\ &= \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)(\rho\sigma_1\sigma_2) \\ &= (1 - \rho^2)\sigma_1^2\sigma_2^2 \end{aligned} \quad (81)$$

- E a inversa da matriz de covariância é dada por

$$\mathbf{C}_x^{-1} = \begin{bmatrix} \frac{\sigma_2^2}{|\mathbf{C}_x|} & -\frac{\rho\sigma_1\sigma_2}{|\mathbf{C}_x|} \\ -\frac{\rho\sigma_1\sigma_2}{|\mathbf{C}_x|} & \frac{\sigma_1^2}{|\mathbf{C}_x|} \end{bmatrix} = \begin{bmatrix} \frac{1}{(1-\rho^2)\sigma_1^2} & \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} \\ \frac{-\rho}{(1-\rho^2)\sigma_1\sigma_2} & \frac{1}{(1-\rho^2)\sigma_2^2} \end{bmatrix}$$

Introdução à Classificação de Padrões

Matriz de Covariância

Exercício Resolvido 1 (Matriz de Covariância)

- Dada a seguinte matriz:

$$\mathbf{A} = \begin{bmatrix} 1 & 0,8 \\ 0,8 & 4 \end{bmatrix}$$

- Pede-se determinar se ela pode ser candidata a matriz de covariância.
- **Solução:**
 - ① Logo de cara vemos que a matriz é simétrica!
 - ② Seu determinante é positivo: $\det(\mathbf{A}) = 3,36$.
 - ③ E seus autovalores são positivos: $\lambda_1 = 0,8$ e $\lambda_2 = 4,2$.
- Logo, concluímos que a matriz \mathbf{A} pode de fato ser uma matriz de covariância.

Introdução à Classificação de Padrões

Matriz de Covariância

Matriz de Covariância no Matlab/Octave e no Excel

- Dadas N observações conjuntas de p atributos, a matriz de covariância pode ser facilmente calculada em vários ambientes de programação científica (Matlab/Octave) e em planilhas numéricas.
 - No Matlab/Octave, usar o comando **COV**.
 - No Excel, faz-se necessário adicionar um plugin de análise de dados. O tutorial abaixo disponível na internet ensina como fazê-lo.

www.youtube.com/watch?v=5fEdfQ03g4c

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada (cont.-3)

- Finalmente, podemos escrever a FDP gaussiana bivariada como

$$f_{\mathbf{x}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q(x_1, x_2)\right\}, \quad (82)$$

em que

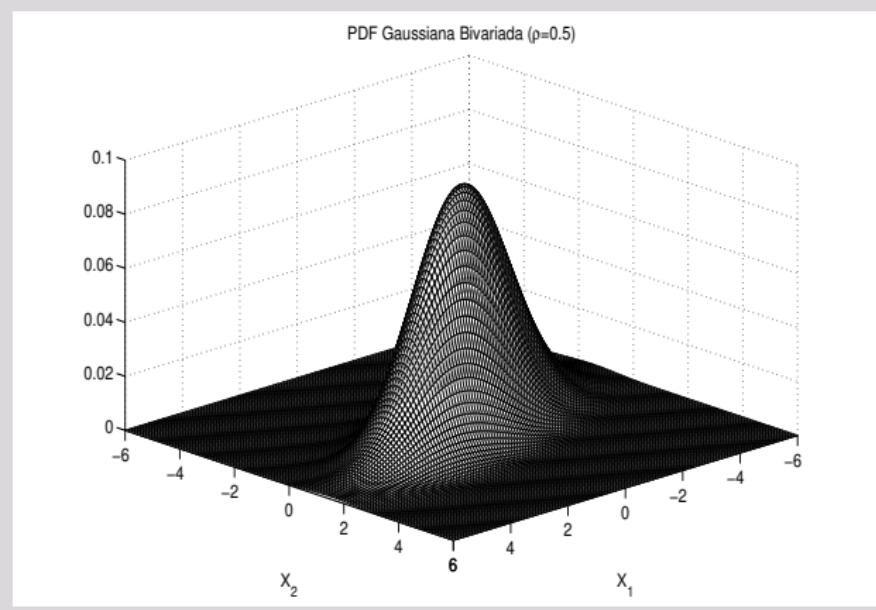
$$\begin{aligned} Q(x_1, x_2) &= \frac{1}{(1-\rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \\ &= \frac{1}{(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \end{aligned}$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada (cont.-4)

- Parâmetros: $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$ e $\rho = 0,5$.



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDPs Marginais da FDP Gaussiana Bivariada

- As PDF gaussianas marginais de X_1 e X_2 são dadas pelas seguintes expressões (Demonstrar!):

$$\begin{aligned}f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(x_1, x_2) dx_2 \\&= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right\}\end{aligned}$$

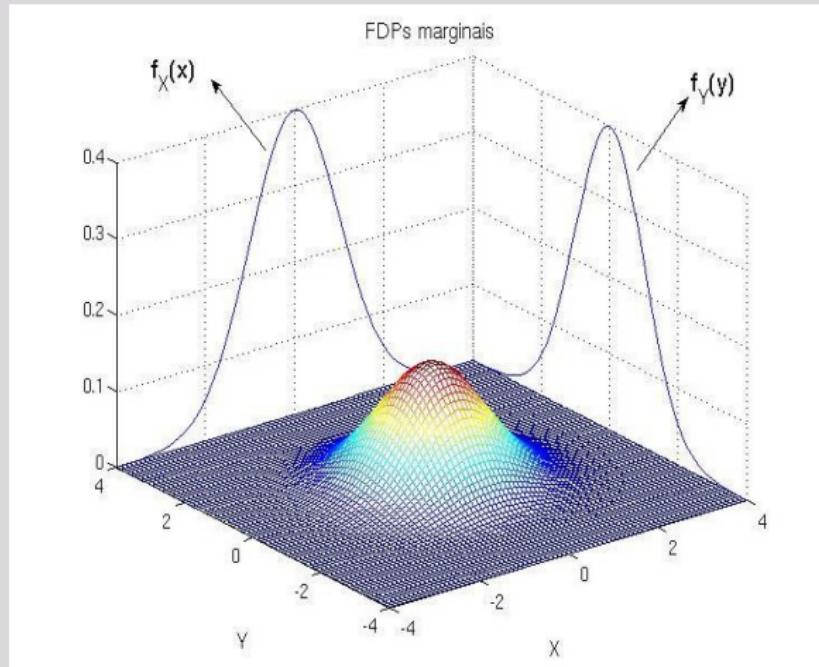
$$\begin{aligned}f_{X_2}(x_2) &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(x_1, x_2) dx_1 \\&= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right\}\end{aligned}$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDPs Marginais da FDP Gaussiana Bivariada (cont.-1)

- Parâmetros: $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ e $\rho = 0$.



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Multivariada

FDPs Condicionais a partir da FDP Gaussiana Bivariada

- PDFs gaussianas condicionais em x_1 ou x_2 podem ser obtidas a partir das seguintes expressões:

$$f_{\mathbf{x}}(x_1|x_2) = \frac{f_{\mathbf{x}}(x_1, x_2)}{f_{x_2}(x_2)} \quad \text{ou} \quad f_{\mathbf{x}}(x_2|x_1) = \frac{f_{\mathbf{x}}(x_1, x_2)}{f_{x_1}(x_1)} \quad (83)$$

- Para um valor específico de x_1 ou x_2 , digamos x^* , estas expressões podem ser escritas como:

$$f_{\mathbf{x}}(x_1|x_2 = x^*) = \frac{f_{\mathbf{x}}(x_1, x_2 = x^*)}{f_{x_2}(x_2 = x^*)} \quad \text{ou} \quad (84)$$

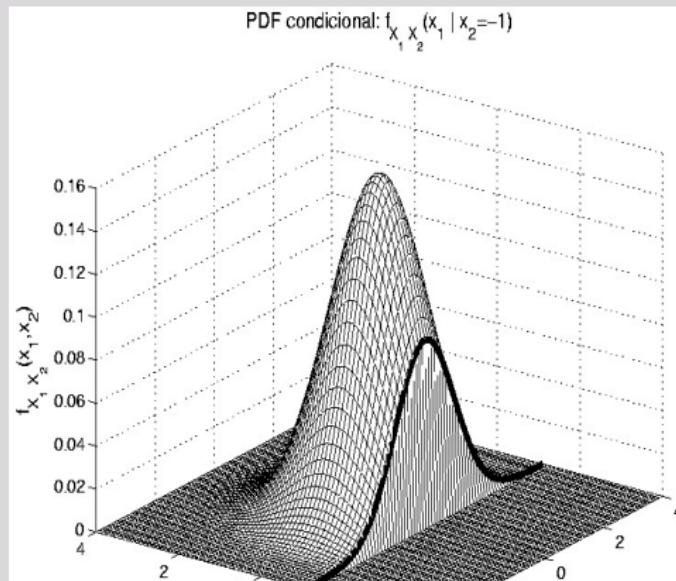
$$f_{\mathbf{x}}(x_2|x_1 = x^*) = \frac{f_{\mathbf{x}}(x_1 = x^*, x_2)}{f_{x_1}(x_1 = x^*)} \quad (85)$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Multivariada

FDPs Condicionais a partir da FDP Gaussiana Bivariada (cont.-1)

- Parâmetros: $x_2 = x^* = -1$, $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ e $\rho = 0$.



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada com $\rho = 0$.

- Quando $\rho = 0$, tem-se as seguintes consequências para a matriz de covariância, seu determinante e sua inversa:

$$\mathbf{C}_x = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \Rightarrow |\mathbf{C}_x| = \sigma_1^2 \sigma_2^2$$

$$\mathbf{C}_x^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

FDP Gaussiana Bivariada com $\rho = 0$. (cont.-1)

- Note que com $\rho = 0$, podemos escrever a FDP gaussiana bivariada como

$$f_{\mathbf{x}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\}$$

- E usando a identidade $e^{a+b} = e^a e^b$, podemos escrever

$$\begin{aligned} f_{\mathbf{x}}(x_1, x_2) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left\{ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right\}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{\left\{ -\frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right\}} \\ &= f_{X_1}(x_1) \cdot f_{X_2}(x_2) \end{aligned} \tag{86}$$

Não-Correlação e Independência

- **Regra Geral:**

Se variáveis aleatórias são independentes, então elas também são não-correlacionadas!

- O raciocínio inverso não é válido:

Se variáveis aleatórias são não-correlacionadas, isto não implica que elas sejam independentes!

- A Eq. (86) revela uma importante exceção à regra geral:

Variáveis aleatórias gaussianas não-correlacionadas são sempre independentes!!

Parte III

Contornos da Gaussiana Bivariada

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada

- Vimos que o gráfico da FDP gaussiana bivariada consiste em uma superfície tridimensional.
- Na prática, não há necessidade de desenhar esta superfície toda vez que o interesse for a análise **qualitativa** da correlação entre duas variáveis aleatórias x_i e x_j .
- Esta análise, pode ser feita numericamente através do coeficiente de correlação ρ_{ij} , ou através das **Curvas de Contorno**.

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada

- Uma curva de contorno é a curva resultante da interseção do plano $z = C$, plano este paralelo ao plano (x_1, x_2) , e a superfície de $f_{\mathbf{x}}(x_1, x_2)$, em que $C > 0$ é uma constante.
- Matematicamente, isto é o mesmo que escrever:

$$z = f_{\mathbf{x}}(x_1, x_2) = C \quad \Rightarrow \quad \exp \left\{ -\frac{1}{2} Q(x_1, x_2) \right\} = C, \quad (87)$$

em que

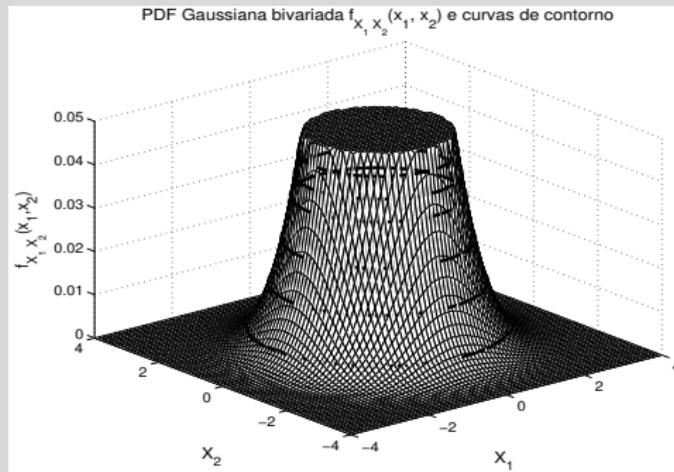
$$Q(x_1, x_2) = \frac{1}{(1 - \rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right]$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada (cont.-1)

- Geometricamente, uma curva de contorno é a curva fechada constituída pelo perfil resultante de um corte transversal da superfície $f_{\mathbf{x}}(x_1, x_2)$ a uma altura C .



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada (cont.-2)

- Após alguma manipulação algébrica chega-se à seguinte expressão:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = C \quad (88)$$

- A forma da curva da Eq. (88) depende dos valores de ρ , σ_1 e σ_2 .
- Dois casos de interesse serão estudados a seguir, a saber:

- ① **Variáveis Não-Correlacionadas ($\rho = 0$)**
- ② **Variáveis Correlacionadas ($\rho \neq 0$)**

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho = 0$)

- Neste caso, a Eq. (88) reduz-se à seguinte expressão:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = C \quad (89)$$

a partir da qual, após dividir ambos os lados por C , chega-se à *forma canônica* da elipse:

$$\boxed{\frac{(x_1 - \mu_1)^2}{(\sqrt{C}\sigma_1)^2} + \frac{(x_2 - \mu_2)^2}{(\sqrt{C}\sigma_2)^2} = 1} \quad (90)$$

- Esta elipse tem centro na coordenada (μ_1, μ_2) e os comprimentos dos semi-eixos são dados por $\sqrt{C}\sigma_1$ e $\sqrt{C}\sigma_2$.

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho = 0$) (cont.-1)

Temos 3 casos a considerar:

$\sigma_1 > \sigma_2$ - Elipse mais alongada na horizontal (eixo X1).

Exemplo de matriz de covariância: $\mathbf{C} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$

$\sigma_1 < \sigma_2$ - Elipse mais alongada na vertical (eixo X2).

Exemplo de matriz de covariância: $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$

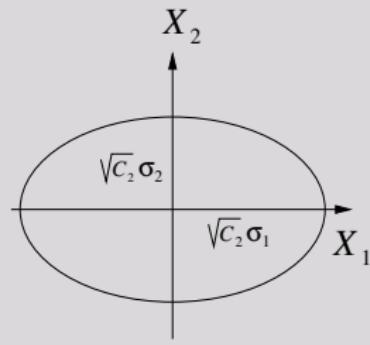
$\sigma_1 = \sigma_2$ - Elipse degenerada em uma circunferência.

Exemplo de matriz de covariância: $\mathbf{C} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$

Variáveis Aleatórias Multidimensionais

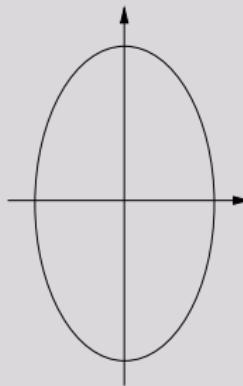
Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho = 0$) (cont.-2)



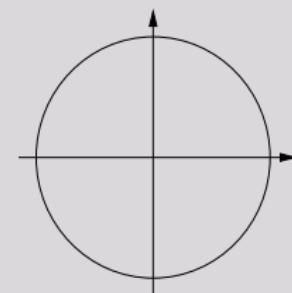
$$\sigma_1 > \sigma_2$$

$$\mathbf{C} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\sigma_1 < \sigma_2$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$



$$\sigma_1 = \sigma_2$$

$$\mathbf{C} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$)

- Neste caso, a Eq. (88), continua representando uma curva de contorno em forma de elipse.
- Contudo, torna-se importante conhecer o efeito introduzido pelo termo

$$-\frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2}, \quad (91)$$

no desenho da elipse.

- Para facilitar a análise, vamos assumir $\mu_1 = \mu_2 = 0$ e $\sigma_1 = \sigma_2 = 1$.
- Assim, a Eq. (88) passa a ser escrita da seguinte maneira:

$$x_1^2 - 2\rho x_1 x_2 + x_2^2 = C \quad (92)$$

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$) (cont.-1)

- A Eq. (92) pode ser reorganizada de tal modo a representar uma equação do segundo grau em x_2 :

$$ax_2^2 + bx_2 + c = 0 \quad (93)$$

em que $a = 1$, $b = -2\rho x_1$ e $c = x_1^2 - C_2$.

- As raízes desta equação são calculadas pela seguinte expressão:

$$x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (94)$$

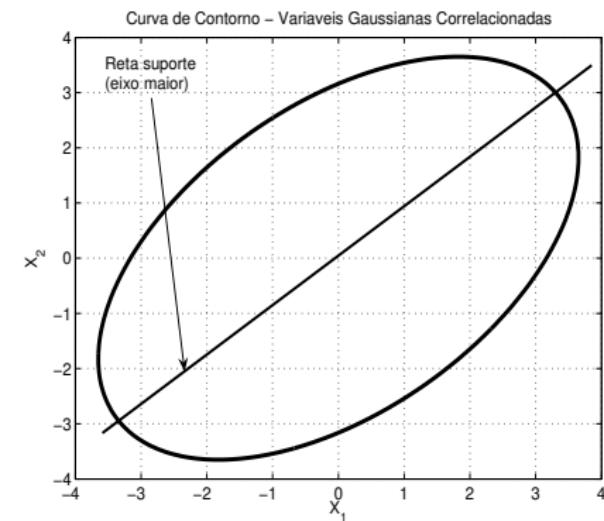
- Usando a Eq. (93), define-se primeiramente uma faixa de valores para x_1 e, em seguida, os valores de x_2 são calculados para cada valor de x_1 usando a Eq. (94).

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$) (cont.-2)

Curva de contorno para x_1 e x_2 com correlação positiva ($\rho = 0,5$). Matriz de covariância associada: $\mathbf{C} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$.

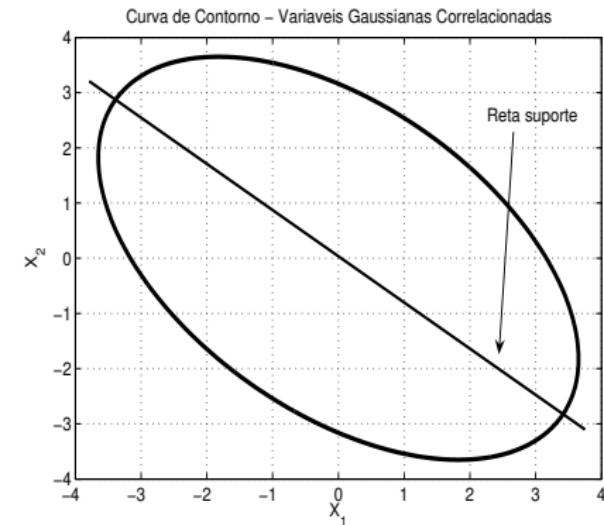


Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$) (cont.-3)

Curva de contorno para x_1 e x_2 com correlação negativa ($\rho = -0,5$). Matriz de covariância associada: $\mathbf{C} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$.

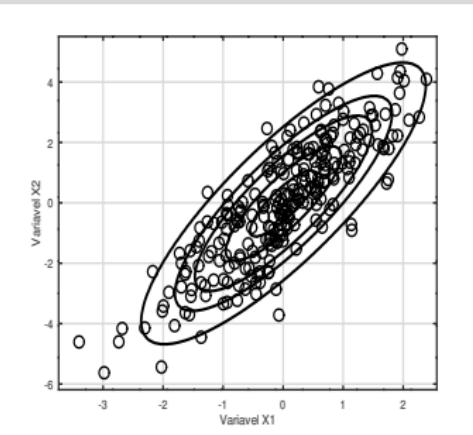
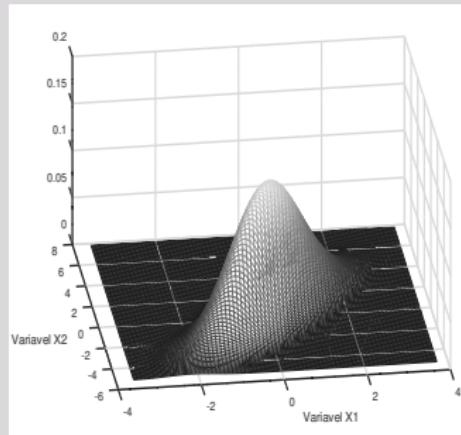


Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho = 0,8$) (cont.-4)

- Parâmetros: $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$ e $\sigma_2 = 2$.

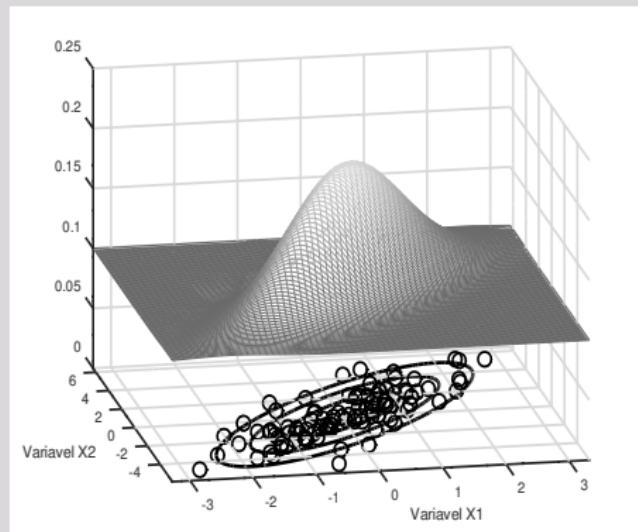


Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Contornos da FDP Gaussiana Bivariada ($\rho = 0,8$) (cont.-5)

- Parâmetros: $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$ e $\sigma_2 = 2$.



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Geração de Vetores Aleatórios Gaussianos Correlacionados

Passo 1 - Gerar vetores aleatórios gaussianos não-correlacionados:

» $X = \text{randn}(p, N);$

Passo 2 - Especificar a matriz de covariância desejada \mathbf{C}_d :

» $C_d = [1 \ 0.79; 0.79 \ 4];$

Passo 3 - Aplicar a Decomposição de Cholesky à matriz \mathbf{C}_d :

» $A = \text{chol}(C_d)';$

Passo 4 - Aplicar transformação linear aos dados originais:

» $Y = A * X;$

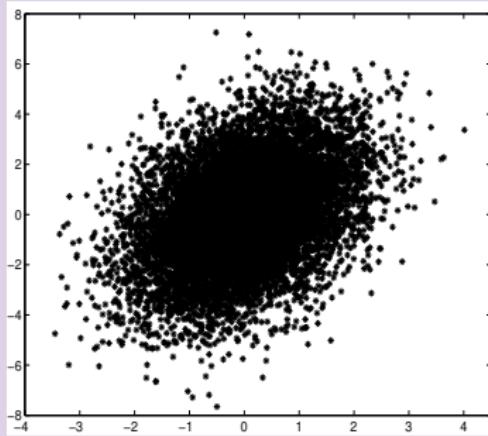
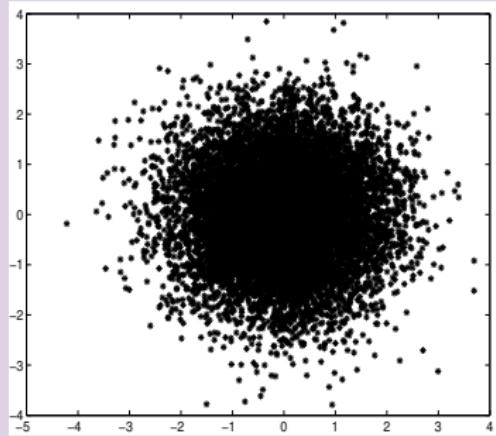
Passo 5 - Conferir se a matriz de covariância estimada $\hat{\mathbf{C}}_d$ está de acordo com a matriz teórica \mathbf{C}_d .

Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Resultados da Aplicação do Algoritmo do Slide Anterior (cont.-1)

- Não-correlacionados (esquerda) e Correlacionados (direita).



Variáveis Aleatórias Multidimensionais

Densidade Gaussiana Bivariada

Resultados da Aplicação do Algoritmo do Slide Anterior (cont.-2)

- Para este problema a matriz de covariância estimada $\hat{\mathbf{C}}_d$ resultante foi:

» $\mathbf{C}_d = \text{cov}(\mathbf{Y})$

$$\begin{matrix} \mathbf{C}_d = & 0.9983 & 0.7836 \\ & 0.7836 & 3.9609 \end{matrix}$$

- O coeficiente de correlação ρ teórico é dado por:

$$\rho\sigma_1\sigma_2 = 0,79 \quad \Rightarrow \quad \rho = \frac{0,79}{\sqrt{1}\sqrt{4}} = 0,395$$

- O coeficiente de correlação $\hat{\rho}$ estimado é dado por:

$$\hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 = 0,7836 \quad \Rightarrow \quad \hat{\rho} = \frac{0,7836}{\sqrt{0,9983}\sqrt{3,9609}} = 0,3941$$

Parte IV

Estimação da Matriz de Covariâcia

Introdução à Classificação de Padrões

Matriz de Covariância

Matriz de Covariância - Definição Teórica

- A matriz de covariância de um vetor de variáveis aleatórias $\mathbf{x} \in \mathbb{R}^p$ é representada como

$$\mathbf{C}_{\mathbf{x}} = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T], \quad (95)$$

$$= E[\mathbf{x}\mathbf{x}^T - \mathbf{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\mathbf{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T], \quad (96)$$

$$= E[\mathbf{x}\mathbf{x}^T] - E[\mathbf{x}]\boldsymbol{\mu}^T - \boldsymbol{\mu}E[\mathbf{x}]^T + E[\boldsymbol{\mu}\boldsymbol{\mu}^T], \quad (97)$$

$$= E[\mathbf{x}\mathbf{x}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T, \quad (98)$$

$$= \mathbf{R}_{\mathbf{x}} - \boldsymbol{\mu}\boldsymbol{\mu}^T, \quad (99)$$

em que o $E[\cdot]$ denota o operador valor esperado, $\boldsymbol{\mu} = E[\mathbf{x}]$ denota o valor esperado de \mathbf{x} e $\mathbf{R}_{\mathbf{x}} = E[\mathbf{x}\mathbf{x}^T]$ é a matriz de correlação.

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação da Matriz de Covariância - Forma Vetorial

- Seja um conjunto de N observações do vetor aleatório $\mathbf{x} \in \mathbb{R}^p$:

$$\{\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(N)\}$$

- A matriz de covariância pode ser estimada diretamente por meio da seguinte expressão:

$$\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N [\mathbf{x}(n) - \mathbf{m}] [\mathbf{x}(n) - \mathbf{m}]^T \quad (100)$$

- O vetor médio do conjunto de observações é estimado como

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n) \quad (101)$$

Introdução à Classificação de Padrões

Matriz de Covariância

Exemplo no Octave/Matlab: Função `mcovar1()`

- A Eq. (100) pode ser implementada pela seguinte rotina:

```
function C=mcovar1(X)

[p N]=size(X); % vetores de atributos nas colunas
soma=zeros(p);
m=mean(X,2); % vetor medio (centroide)
for j=1:N, % percorre as N colunas de X
    aux = X(:,j)-m; % vetor auxiliar
    soma = soma + aux*aux';
end
C=soma/N; % Matriz de covariancia
```

- Para executar no Octave/Matlab: » `Cx = mcovar1(X);`

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação da Matriz de Covariância - Forma Matricial

- Considere que os N vetores aleatórios formam as colunas de uma matriz $\mathbf{X} \in \mathbb{R}^{p \times N}$:

$$\mathbf{X} = [\mathbf{x}(1) | \mathbf{x}(2) | \cdots | \mathbf{x}(N)] \quad (102)$$

- Usando esta definição, a matriz de covariância pode ainda ser estimada como

$$\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{N} [\mathbf{X} - \mathbf{M}] [\mathbf{X} - \mathbf{M}]^T \quad (103)$$

em que $\mathbf{M} \in \mathbb{R}^{p \times N}$ é uma matriz cujas colunas são todas iguais \mathbf{m} :

$$\mathbf{M} = [\mathbf{m} | \mathbf{m} | \cdots | \mathbf{m}] \quad (104)$$

Introdução à Classificação de Padrões

Matriz de Covariância

Exemplo no Octave/Matlab: Função `mcovar2()`

- A Eq. (100) pode ser implementada pela seguinte rotina:

```
function C=mcovar2(X)  
  
[p N]=size(X); % dimensoes da matriz de dados  
m = mean(X,2); % vetor medio (centroide)  
aux = X-m; % matriz de dados centralizada  
C = aux*aux'/N; % Matriz de covariancia
```

- Para executar no Octave/Matlab: `» C = mcovar2(X);`
- Percebe-se o custo bem menor desta implementação em relação à anterior, não só quanto ao número de linhas do código, mas também quanto à velocidade.

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação da Matriz de Covariância - Forma Vetorial

- Pode-se usar também a seguinte expressão alternativa derivada da Eq. (99):

$$\hat{\mathbf{C}}_{\mathbf{x}} = \hat{\mathbf{R}}_{\mathbf{x}} - \mathbf{m}\mathbf{m}^T \quad (105)$$

- Com a matriz de correlação $\mathbf{R}_{\mathbf{x}}$ sendo estimada como

$$\hat{\mathbf{R}}_{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n)\mathbf{x}^T(n) \quad (106)$$

Introdução à Classificação de Padrões

Matriz de Covariância

Derivação da Equação (105) a partir da Eq. (100):

$$\begin{aligned}\hat{\mathbf{C}}_{\mathbf{x}} &= \frac{1}{N} \sum_{n=1}^N [\mathbf{x}(n) - \mathbf{m}][\mathbf{x}(n) - \mathbf{m}]^T, \\ &= \frac{1}{N} \sum_{n=1}^N \left[\mathbf{x}(n)\mathbf{x}^T(n) - \mathbf{x}(n)\mathbf{m}^T - \mathbf{m}\mathbf{x}^T(n) + \mathbf{m}\mathbf{m}^T \right], \\ &= \frac{1}{N} \left[\left(\sum_{n=1}^N \mathbf{x}(n)\mathbf{x}^T(n) \right) - \left(\sum_{n=1}^N \mathbf{x}(n) \right) \mathbf{m}^T - \mathbf{m} \left(\sum_{n=1}^N \mathbf{x}^T(n) \right) + \left(\sum_{n=1}^N \mathbf{m}\mathbf{m}^T \right) \right], \\ &= \frac{1}{N} \left[\left(\sum_{n=1}^N \mathbf{x}(n)\mathbf{x}^T(n) \right) - (N\mathbf{m})\mathbf{m}^T - \mathbf{m}(N\mathbf{m}^T) + N\mathbf{m}\mathbf{m}^T \right], \\ &= \frac{1}{N} \left[\left(\sum_{n=1}^N \mathbf{x}(n)\mathbf{x}^T(n) \right) - 2N\mathbf{m}\mathbf{m}^T + N\mathbf{m}\mathbf{m}^T \right], \\ &= \frac{1}{N} \left(\sum_{n=1}^N \mathbf{x}(n)\mathbf{x}^T(n) \right) - \mathbf{m}\mathbf{m}^T, \\ &= \hat{\mathbf{R}}_{\mathbf{x}} - \mathbf{m}\mathbf{m}^T\end{aligned}$$

Introdução à Classificação de Padrões

Matriz de Covariância

Exemplo no Octave/Matlab: Função `mcovar3()`

- A Eq. (105) pode ser implementada pela seguinte rotina:

```
function C=mcovar3(X)

[p N]=size(X); % vetores de atributos nas colunas
R=zeros(p);
m=mean(X,2); % vetor medio (centroide)
for j=1:N, % percorre as N colunas de X
    R = R + X(:,j)*X(:,j)';
end
C = R/N - m*m'; % Matriz de covariancia
```

- Para executar no Octave/Matlab: `» Cx = mcovar3(X);`
- Esta rotina evita a execução de N subtrações vetoriais $\text{aux} = \mathbf{X}(:,j) - \mathbf{m}$ dentro do laço FOR.

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação da Matriz de Correlação - Forma Matricial

- Usando a definição da matriz \mathbf{X} mostrada na Eq. (102), a matriz de correlação pode ser estimada como

$$\hat{\mathbf{R}}_{\mathbf{x}} = \frac{1}{N} \mathbf{XX}^T, \quad (107)$$

de modo que a matriz de covariância pode ser estimada via Eq. (105).

- As fórmulas matriciais são particularmente úteis para implementação em ambientes de computação numérica, tais como Matlab, Octave e Scilab.

Introdução à Classificação de Padrões

Matriz de Covariância

Exemplo no Octave/Matlab: Função `mcovar4()`

- A Eq. (105) pode ainda ser implementada de forma mais eficiente pela seguinte rotina:

```
function C=mcovar4(X)

[p N]=size(X); % dimensoes da matriz de dados
m = mean(X,2); % vetor medio (centroide)
R = X*X'/N; % Matriz de correlacao (opcional)
C = R - m*m'; % Matriz de covariancia
```

- Para executar no Octave/Matlab: `» Cx = mcovar4(X);`
- Esta rotina é a de mais rápida execução dentre todas as anteriores.

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação do Vetor de Médias - Forma Recursiva

- A Equação (101) utiliza todos os vetores aleatórios de uma só vez na estimação do vetor $\bar{\mathbf{X}}$.
- Em muitos casos, os vetores só estão disponíveis um-a-um, de forma sequencial. Assim, deve-se estimar tal vetor recursivamente. Por exemplo:

$$\begin{aligned}\mathbf{m}(n) &= \left(\frac{n-1}{n} \right) \mathbf{m}(n-1) + \frac{1}{n} \mathbf{x}(n) \quad (108) \\ &= \alpha \mathbf{m}(n-1) + (1-\alpha) \mathbf{x}(n)\end{aligned}$$

em que n é a iteração atual, $\mathbf{x}(n)$ é o vetor sendo observado e $0 < \alpha < 1$ denota uma constante.

- Em $n = 0$, faz-se $\bar{\mathbf{X}}(0) = \mathbf{0}_p$ = vetor-nulo de dimensão $p \times 1$.

Introdução à Classificação de Padrões

Matriz de Covariância

Derivação da Eq. (108)

- A Equação (101) pode ser escrita como

$$\begin{aligned}\mathbf{m}_N &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n) = \frac{1}{N} \left(\sum_{n=1}^{N-1} \mathbf{x}(n) + \mathbf{x}(N) \right), \\ &= \frac{1}{N} ((N-1)\mathbf{m}_{N-1} + \mathbf{x}(N)), \\ &= \left(\frac{N-1}{N} \right) \mathbf{m}_{N-1} + \frac{1}{N} \mathbf{x}(N)\end{aligned}$$

em que \mathbf{m}_N é a média com N vetores e \mathbf{m}_{N-1} é a média com $N - 1$ vetores da sequência.

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação da Matriz de Correlação - Forma Recursiva

- Até agora, para estimar a matriz \mathbf{R}_x , usamos todos os vetores aleatórios de uma única vez. Porém, quando os vetores só estão disponíveis um-a-um, deve-se estimar tal matriz recursivamente por uma expressão similar à Eq. (108):

$$\begin{aligned}\hat{\mathbf{R}}_x(n) &= \left(\frac{n-1}{n} \right) \hat{\mathbf{R}}_x(n-1) + \frac{1}{n} \mathbf{x}(n) \mathbf{x}^T(n) \quad (109) \\ &= \alpha \hat{\mathbf{R}}_x(n-1) + (1-\alpha) \mathbf{x}(n) \mathbf{x}^T(n)\end{aligned}$$

em que n é a iteração atual, $\mathbf{x}(n)$ é o vetor sendo observado e $0 < \alpha < 1$ denota uma constante.

- Em $n = 1$, faz-se $\hat{\mathbf{R}}_x(0) = \mathbf{I}_p$ = matriz identidade $p \times p$.

Introdução à Classificação de Padrões

Matriz de Covariância

Exemplo no Octave/Matlab: Estimação Sequencial

- A Eq. (109) pode ser implementada pela seguinte rotina:

```
X=load('datamatrix.dat');
[p N]=size(X);
m=zeros(p,1); R=eye(p);
for n=1:N, % percorre as N colunas de X
    a=((n-1)/n);
    x = X(:,n); % vetor entrada no instante n
    m = a*m + (1-a)*x;
    R = a*R + (1-a)*x*x'; % Correlacao na iteracao n
    C = R - m*m'; % Covariancia na iteracao n
end
```

Introdução à Classificação de Padrões

Matriz de Covariância

Estimação via comando COV do Octave/Matlab

- A Equação (101) pode ser escrita como

$$\begin{aligned}\mathbf{m}_N &= \frac{1}{N} \sum_{n=1}^N \mathbf{x}(n) = \frac{1}{N} \left(\sum_{n=1}^{N-1} \mathbf{x}(n) + \mathbf{x}(N) \right), \\ &= \frac{1}{N} ((N-1)\mathbf{m}_{N-1} + \mathbf{x}(N)), \\ &= \left(\frac{N-1}{N} \right) \mathbf{m}_{N-1} + \frac{1}{N} \mathbf{x}(N)\end{aligned}$$

em que \mathbf{m}_N é a média com N vetores e \mathbf{m}_{N-1} é a média com $N - 1$ vetores da sequência.

Parte V

Classificador Quadrático

Introdução à Classificação de Padrões

Classificador Quadrático

Classificação Usando a Distância Ponderada

- É um classificador baseado em distância ao centróide.
- Usa a distância ponderada (ou de Mahalanobis) em vez da Euclidiana.
- Com a matriz de ponderação sendo a inversa da matriz de covariância \mathbf{C}_i .
- Assim, a distância do vetor de atributos \mathbf{x} ao i -ésimo centróide \mathbf{m}_i é dada por

$$d(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i) = (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i). \quad (110)$$

Introdução à Classificação de Padrões

Classificador Quadrático

Passo 1 - Determinar o vetor centróide de cada classe:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x} \quad (111)$$

onde N_i é o número de exemplos da classe ω_i , $i = 1, \dots, C$.

Passo 2 - Determinar a matriz de covariância de cada classe:

$$\mathbf{C}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad (112)$$

Passo 3 - Calcular as distâncias quadráticas de cada novo vetor de atributos \mathbf{x} aos centróides de todas as classes:

$$d(\mathbf{x} | \mathbf{m}_i, \mathbf{C}_i) = (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i), \quad i = 1, \dots, C. \quad (113)$$

Passo 4 - Encontrar a classe que produz a menor distância quadrática para \mathbf{x} :

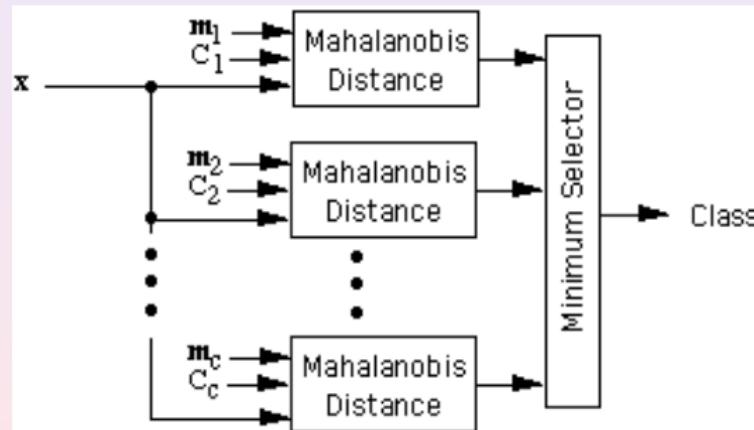
$$i^* = \arg \min_{i=1, \dots, C} \{d(\mathbf{x} | \mathbf{m}_i, \mathbf{C}_i)\} \quad (114)$$

Passo 5 - Atribuir \mathbf{x} à classe cujo rótulo é ω_{i^*} .

Introdução à Classificação de Padrões

Classificador Quadrático

- Diagrama de fluxo de sinais para o classificador quadrático para um problema de C classes.



Introdução à Classificação de Padrões

Classificador Quadrático

Limitações do Classificador Quadrático

- O desempenho do classificador quadrático depende fortemente da invertibilidade da matriz de covariância de cada classe: \mathbf{C}_i , $i = 1, \dots, C$.
- A invertibilidade de uma matriz de covariância pode ser avaliada pelo seu posto, definido como

$$\text{posto}(\mathbf{C}_i) \leq \min(p, p) = p, \quad (115)$$

em que p é a dimensão do vetor de atributos.

- A matriz é invertível se $\text{posto}(\mathbf{C}_i) = p$. Neste caso, diz-se que a matriz é de posto completo.
- No Matlab/Octave, o posto é obtido pelo comando `rank`.

Introdução à Classificação de Padrões

Classificador Quadrático

Limitações do Classificador Quadrático

- O posto de uma matriz quadrada, como a matriz de covariância, indica o número de linhas ou colunas linearmente independentes (LI).
- Se a matriz de covariância for de posto completo, isto indica que suas linhas (ou colunas) formam uma base do espaço \mathbb{R}^p , que é o espaço de atributos.
- Neste caso, o determinante da matriz de covariância será diferente de zero e esta matriz será, portanto, invertível.

Introdução à Classificação de Padrões

Classificador Quadrático

Limitações do Classificador Quadrático

- A invertibilidade da matriz de covariância \mathbf{C}_i da classe ω_i pode ser inferida a partir do posto da matriz de dados daquela classe.
- Seja $\mathbf{X}_i \in \mathbb{R}^{p \times N_i}$ a matriz de dados da classe ω_i , com N_i indicando o número de exemplos daquela classe.
- O posto desta matriz é $\text{posto}(\mathbf{X}_i) \leq \min(p, N_i)$.
- Pode-se mostrar que $\text{posto}(\mathbf{X}_i) = \text{posto}(\mathbf{C}_i)$.

Introdução à Classificação de Padrões

Classificador Quadrático

Matriz de Covariância Cheia e Normalização

- Ao usar a matriz \mathbf{C}_i na distância quadrática, não é necessário normalizar os dados previamente.
- A título de exemplo, considere 2 atributos. Neste caso, a matriz \mathbf{C}_i é dada por

$$\mathbf{C}_i = \begin{bmatrix} \sigma_{i,1}^2 & \rho_i \sigma_{i,1} \sigma_{i,2} \\ \rho_i \sigma_{i,1} \sigma_{i,2} & \sigma_{i,2}^2 \end{bmatrix}. \quad (116)$$

- Deste modo, a matriz inversa é dada por como

$$\mathbf{C}_i^{-1} = \begin{bmatrix} \frac{\sigma_{i,2}^2}{|\mathbf{C}_i|} & -\frac{\rho_i \sigma_{i,1} \sigma_{i,2}}{|\mathbf{C}_i|} \\ -\frac{\rho_i \sigma_{i,1} \sigma_{i,2}}{|\mathbf{C}_i|} & \frac{\sigma_{i,1}^2}{|\mathbf{C}_i|} \end{bmatrix} \quad (117)$$

$$= \begin{bmatrix} \frac{1}{(1-\rho_i^2)\sigma_{i,1}^2} & \frac{-\rho_i}{(1-\rho_i^2)\sigma_{i,1}\sigma_{i,2}} \\ \frac{-\rho_i}{(1-\rho_i^2)\sigma_{i,1}\sigma_{i,2}} & \frac{1}{(1-\rho_i^2)\sigma_{i,2}^2} \end{bmatrix} \quad (118)$$

Introdução à Classificação de Padrões

Matriz de Covariância Cheia e Normalização

Podemos então determinar a distância quadrática de um vetor de atributos \mathbf{x} para o centróide da i -ésima classe:

$$d_i(\mathbf{x}, \mathbf{m}_i) = [(x_1 - m_{i1}) \quad (x_2 - m_{i2})] \begin{bmatrix} \frac{1}{(1-\rho_i^2)\sigma_{i,1}^2} & \frac{-\rho_i}{(1-\rho_i^2)\sigma_{i,1}\sigma_{i,2}} \\ \frac{-\rho_i}{(1-\rho_i^2)\sigma_{i,1}\sigma_{i,2}} & \frac{1}{(1-\rho_i^2)\sigma_{i,2}^2} \end{bmatrix} \begin{bmatrix} (x_1 - m_{i1}) \\ (x_2 - m_{i2}) \end{bmatrix} \quad (119)$$

Que resulta na seguinte expressão:

$$d_i(x_1, x_2) = \frac{(x_1 - m_{i,1})^2}{\sigma_{i,1}^2(1 - \rho_i^2)} - \frac{2\rho_i(x_1 - m_{i,1})(x_2 - m_{i,2})}{(1 - \rho_i^2)\sigma_{i,1}\sigma_{i,2}} + \frac{(x_2 - m_{i,2})^2}{\sigma_{i,2}^2(1 - \rho_i^2)} \quad (120)$$

De onde concluímos que os atributos x_1 e x_2 acabam sendo automaticamente normalizados.

Parte VI

Variantes do Classificador Quadrático

Introdução à Classificação de Padrões

Classificador Quadrático

Variantes do Classificador Quadrático

- Vimos que a invertibilidade da matriz de covariância é um ponto crítico da implementação adequada do classificador quadrático.
- E que a invertibilidade da matriz de covariância está intimamente associada ao seu posto ser completo.
- Descreveremos a seguir maneiras de tornar a matriz de covariância invertível, tornando-a de posto completo.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- **Variante 1:** Usa a matriz de covariância regularizada, calculada como

$$\mathbf{C}_i(\lambda) = \mathbf{C}_i + \lambda \mathbf{I}_p, \quad (121)$$

em que $0 \leq \lambda \ll 1$ é a constante de regularização, de valor pequeno, e \mathbf{I}_p é a matriz identidade de dimensão $p \times p$.

- O efeito prático da regularização mostrada na Eq. (121) é adicionar um pequeno valor λ aos elementos da diagonal principal de \mathbf{C}_i , tornando-a diagonalmente dominante.
- Pode-se mostrar que a regularização via Eq. (121) equivale à adição de ruído branco gaussiano à matriz de dados \mathbf{X}_i da classe ω_i .

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Considere o vetor aleatório $\mathbf{r} \in \mathbb{R}^p$, que obedece à seguinte lei de distribuição de probabilidades: $\mathbf{r} \sim N(\mathbf{0}_p, \sigma_r^2 \mathbf{I}_p)$.
- Em outras palavras, \mathbf{r} está distribuído segundo uma distribuição gaussiana, de vetor médio $\mathbf{0}_p$ e matriz de covariância $\sigma_r^2 \mathbf{I}_p$.
- Note que a matriz $\sigma_r^2 \mathbf{I}_p$ é diagonal, logo as componentes do vetor \mathbf{r} são descorrelacionadas.
- Assim, os vetores de atributos da classe ω_i , adicionados de um vetor aleatório \mathbf{r} , passam a ser denotados por $\mathbf{x}^\star = \mathbf{x} + \mathbf{r}$, e têm a seguinte matriz de covariância:

$$\mathbf{C}_i^\star = E[(\mathbf{x}^\star - \mathbf{m}_i^\star)(\mathbf{x}^\star - \mathbf{m}_i^\star)^T] = \mathbf{R}_i^\star - \mathbf{m}_i^\star \mathbf{m}_i^{\star T}. \quad (122)$$

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Desenvolvendo o vetor médio \mathbf{m}_i^* , chega-se a

$$\mathbf{m}_i^* = E[\mathbf{x}^*], \quad (123)$$

$$= E[\mathbf{x} + \mathbf{r}], \quad (124)$$

$$= E[\mathbf{x}] + E[\mathbf{r}], \quad (125)$$

$$= \mathbf{m}_i + \mathbf{0}_p, \quad (126)$$

$$= \mathbf{m}_i, \quad (127)$$

ou seja, o vetor médio dos dados da classe ω_i não é alterado pela adição de ruído branco.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Desenvolvendo a matriz de correlação \mathbf{R}_i^* , chega-se a

$$\mathbf{R}_i^* = E[\mathbf{x}^* \mathbf{x}^{*T}], \quad (128)$$

$$= E[(\mathbf{x} + \mathbf{r})(\mathbf{x} + \mathbf{r})^T], \quad (129)$$

$$= E[\mathbf{x}\mathbf{x}^T + \mathbf{x}\mathbf{r}^T + \mathbf{r}\mathbf{x}^T + \mathbf{r}\mathbf{r}^T], \quad (130)$$

$$= E[\mathbf{x}\mathbf{x}^T] + E[\mathbf{x}\mathbf{r}^T] + E[\mathbf{r}\mathbf{x}^T] + E[\mathbf{r}\mathbf{r}^T], \quad (131)$$

$$= E[\mathbf{x}\mathbf{x}^T] + E[\mathbf{x}]E[\mathbf{r}^T] + E[\mathbf{r}]E[\mathbf{x}^T] + E[\mathbf{r}\mathbf{r}^T], \quad (132)$$

$$= \mathbf{R}_i + \mathbf{m}_i \mathbf{0}_p^T + \mathbf{0}_p \mathbf{m}_i^T + \sigma^2 \mathbf{I}_p, \quad (133)$$

$$= \mathbf{R}_i + \sigma^2 \mathbf{I}_p, \quad (134)$$

onde assumimos que o ruído gaussiano é descorrelacionado e, consequentemente, independente dos dados da classe ω_i .

- A suposição de independência entre \mathbf{x} e \mathbf{r} permite fatorar os termos $E[\mathbf{x}\mathbf{r}^T]$ e $E[\mathbf{r}\mathbf{x}^T]$.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Combinando os dois desenvolvimentos anteriores, a matriz de covariância de \mathbf{x}^* é dada por

$$\mathbf{C}_i^* = \mathbf{R}_i^* - \mathbf{m}_i^* \mathbf{m}_i^{*T}, \quad (135)$$

$$= (\mathbf{R}_i + \sigma^2 \mathbf{I}_p) - \mathbf{m}_i \mathbf{m}_i^T, \quad (136)$$

$$= (\mathbf{R}_i - \mathbf{m}_i \mathbf{m}_i^T) + \sigma^2 \mathbf{I}_p, \quad (137)$$

$$= \mathbf{C}_i + \lambda \mathbf{I}_p, \quad (138)$$

em que fizemos $\lambda = \sigma^2$.

- Concluímos que a matriz de covariância dos dados com ruído branco é igual à matriz de covariância dos dados originais somada da matriz diagonal $\lambda \mathbf{I}_p$.
- O parâmetro de regularização pode ser interpretado como a variância do ruído branco.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Assim, para que a matriz de covariância da classe ω_i seja invertível, devemos regularizá-la pela Eq. (121) ou adicionar ruído branco aos dados da classe ω_i .
- Em suma, **por que adicionar ruído aos dados regulariza a matriz de covariância?**
- Porque do ponto de vista geométrico, adicionar ruído aos dados significa “desalinear” aqueles vetores que são co-lineares, ou seja, que são linearmente dependentes.
- Ao desalinear tais vetores, os vetores que formam as p linhas da matriz de dados ficam LI. De modo que o posto da matriz de covariância fica completo.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- **Variante 2:** Usa a matriz de covariância agregada (*pooled*), calculada como

$$\mathbf{C}_{pool} = \frac{1}{N} (N_1 \mathbf{C}_1 + \cdots + N_K \mathbf{C}_K) = \sum_{k=1}^K \frac{N_i}{N} \mathbf{C}_i, \quad (139)$$

em que N_i denota o número de exemplos de treinamento da i -ésima classe.

- A distância quadrática no **Passo 3** passa ser escrita como

$$d(\mathbf{x} | \mathbf{m}_i, \mathbf{C}_{pool}) = (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_{pool}^{-1} (\mathbf{x} - \mathbf{m}_i), \quad i = 1, \dots, K. \quad (140)$$

- Esta variante é útil para problemas em que as classes são desbalanceadas, ao evitar problemas numéricos resultantes de uma inversa mal-condicionada.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- **Variante 3:** Usa o método de regularização de Friedman¹, cuja matriz é calculada como

$$\mathbf{C}_i(\lambda) = \frac{(1 - \lambda)\mathbf{S}_i + \lambda\mathbf{S}_{pool}}{(1 - \lambda)N_i + \lambda N}, \quad (141)$$

em que $0 \leq \lambda \leq 1$, $\mathbf{S}_i = N_i \mathbf{C}_i$ e $\mathbf{S}_{pool} = N \mathbf{C}_{pool}$.

- A distância quadrática no **Passo 3** passa ser escrita como

$$d(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_i(\lambda)) = (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1}(\lambda) (\mathbf{x} - \mathbf{m}_i), \quad i = 1, \dots, K. \quad (142)$$

¹J. H. Friedman (1989). "Regularized Discriminant Analysis", *Journal of the American Statistical Association*, 84:165–175.

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- Sobre a Variante 3 do classificador quadrático, percebe-se que é uma abordagem que cobre o discriminante quadrático original ($\lambda = 0$) e a Variante 2 ($\lambda = 1$):

$$\mathbf{C}_i(\lambda) = \begin{cases} \mathbf{C}_i, & \lambda = 0 \\ \mathbf{C}_{pool}, & \lambda = 1 \end{cases} \quad (143)$$

- O parâmetro λ deve ser entendido como um hiperparâmetro, pois o método só funciona APÓS a sua determinação.
- Normalmente, a especificação de λ é inicialmente feita por tentativa-e-erro, para em seguida proceder com um método mais adequado (e.g. validação cruzada).

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- **Variante 4:** Usar a matriz diagonal extraída das matrizes de covariância da classe:

$$\mathbf{C}_{i,diag} = \text{diag}(\mathbf{C}_i), \quad (144)$$

$$= \begin{bmatrix} \sigma_{i,1}^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_{i,2}^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_{i,p}^2 \end{bmatrix} \quad (145)$$

- A distância quadrática no **Passo 3** passa ser escrita como

$$d(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_{i,diag}) = (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_{i,diag}^{-1} (\mathbf{x} - \mathbf{m}_i), \quad i = 1, \dots, K. \quad (146)$$

Introdução à Classificação de Padrões

Variantes do Classificador Quadrático

- É fácil perceber que a matriz diagonal $\mathbf{C}_{i,diag}$ será sempre invertível, com inversa dada por

$$\mathbf{C}_{i,diag}^{-1} = \begin{bmatrix} \frac{1}{\sigma_{i,1}^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_{i,2}^2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sigma_{i,p}^2} \end{bmatrix} \quad (147)$$

- Usar a matriz $\mathbf{C}_{i,diag}$ significa assumir que os atributos são descorrelacionados para aquela classe, ou seja, desconsidera a informação provida pela correlação entre atributos.
- Esta é a mesma suposição feita pelo famoso classificador *Naive Bayes*, que com a suposição adicional de gaussianidade dos atributos, implica em considerar os atributos independentes.

Introdução à Classificação de Padrões

Classificador Quadrático

Matriz de Covariância Diagonal e Normalização

- Ao usar a matriz diagonal na distância quadrática, também não é necessário normalizar os dados previamente.
- Isto é feito automaticamente pela matriz \mathbf{C}_i^{-1} .
- A título de exemplo, vamos usar apenas 2 atributos.

$$\mathbf{C}_{i,diag} = \text{diag}(\sigma_{i,1}^2, \sigma_{i,2}^2) = \begin{bmatrix} \sigma_{i,1}^2 & 0 \\ 0 & \sigma_{i,2}^2 \end{bmatrix}. \quad (148)$$

- Deste modo, a matriz inversa é dada por como

$$\mathbf{C}_{i,diag}^{-1} = \begin{bmatrix} \frac{1}{\sigma_{i,1}^2} & 0 \\ 0 & \frac{1}{\sigma_{i,2}^2} \end{bmatrix} \quad (149)$$

Introdução à Classificação de Padrões

Classificador Quadrático

Distância Quadrática em Classificação e Normalização

- Podemos então determinar a distância quadrática do vetor \mathbf{x} ao centróide da i -ésima classe:

$$\begin{aligned} d(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_{i,diag}) &= [(x_1 - m_{i1}) \quad (x_2 - m_{i2})] \begin{bmatrix} \frac{1}{\sigma_{i,1}^2} & 0 \\ 0 & \frac{1}{\sigma_{i,2}^2} \end{bmatrix} \begin{bmatrix} (x_1 - m_{i1}) \\ (x_2 - m_{i2}) \end{bmatrix} \\ &= [(x_1 - m_{i1}) \quad (x_2 - m_{i2})] \begin{bmatrix} \frac{(x_1 - m_{i1})}{\sigma_{i,1}^2} \\ \frac{(x_2 - m_{i2})}{\sigma_{i,2}^2} \end{bmatrix}, \end{aligned}$$

Introdução à Classificação de Padrões

Classificador Quadrático

Distância Quadrática em Classificação e Normalização

- Que resulta na seguinte expressão:

$$d(\mathbf{x}|\mathbf{m}_i, \mathbf{C}_{i,diag}) = \left(\frac{x_1 - m_{i1}}{\sigma_{i,1}} \right)^2 + \left(\frac{x_2 - m_{i2}}{\sigma_{i,2}} \right)^2 \quad (150)$$

- De onde concluímos que, ao usar a distância quadrática, os atributos x_1 e x_2 acabam sendo automaticamente normalizados pelos seus respectivos desvios-padrão!
- Compare os termos da expressão (150) com a expressão da normalização estatística em (72), repetida abaixo:

$$x^* = \frac{x - \bar{x}}{\sigma_x} \quad (151)$$

Parte VII

Classificação Binária

Classificação Binária

Classificação Binária

Definição

Classificação binária é a tarefa de categorizar os elementos de um dado conjunto de objetos em *dois* grupos apenas.

Aplicações Típicas

- Diagnóstico médico: determinar se um dado indivíduo possui uma doença ou não.
- Controle de qualidade na indústria: decidir se um dado produto é bom o suficiente para ser vendido ou deve ser descartado.
- Detecção de falhas: decidir se um dado equipamento está operando normalmente ou não.

Classificação Binária

Teste de Hipóteses

Conceituação

Em um teste de hipóteses estatístico, o usuário define inicialmente uma *hipótese nula* (H_0) e uma *hipótese alternativa* (H_1). Em seguida, realiza um experimento e então decide se rejeita H_0 em favor de H_1 .

Resultados Possíveis

- *Falso Positivo* (erro Tipo I): rejeitar a hipótese nula, quando ela é verdadeira.
- *Verdadeiro Positivo*: rejeitar a hipótese nula, quando ela é falsa.
- *Falso Negativo* (erro Tipo II): aceitar a hipótese nula, quando ela é falsa.
- *Verdadeiro Negativo*: aceitar a hipótese nula, quando ela é verdadeira.



Classificação Binária

Avaliação de Classificadores Binários

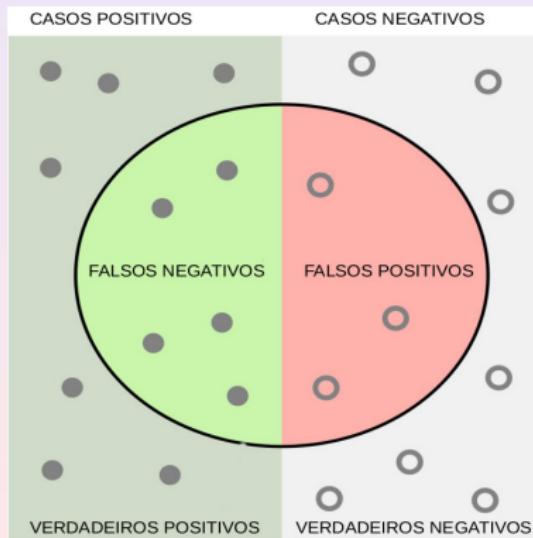
Conceituação

- Digamos que se queira avaliar pacientes sobre a presença de uma determinada doença ou patologia.
- **Verdadeiros Positivos (VP)**: Pacientes que têm a doença e seus exames acusam corretamente a presença da doença, ou seja, dão positivo.
- **Falsos Negativos (FN)**: Pacientes que têm a doença, porém seus exames não acusam a presença da doença, ou seja, dão negativo.
- **Verdadeiros Negativos (VN)**: Pacientes que não têm a doença e seus exames também não acusam a presença da doença.
- **Falsos Positivos (FP)**: Pacientes que não têm a doença, porém seus exames acusam a presença da mesma.

Classificação Binária

Avaliação de Classificadores Binários

- VP, VN, FP e FN como conjuntos em um diagrama de Venn.



Classificação Binária

Matriz de Confusão

Conceituação

- Em problemas de classificação binária, a *matriz de confusão* é uma tabela com duas linhas e duas colunas que reporta o número de Verdadeiros Negativos (VN), Falsos Positivos (FN), Falsos Negativos (FN) e Verdadeiros Positivos (VP).

Matriz de Confusão para Classificação Binária

		Resultado ou Condição Real	
		Positivo	Negativo
Saída Preditiva	Positivo	VP	FP (erro tipo I, valor p)
	Negativo	FN (erro tipo II)	VN

Classificação Binária

Sensibilidade

Figuras de Mérito Baseadas na Matriz de Confusão

- A partir da matriz de confusão de um problema de classificação binária, pode-se obter várias figuras de mérito que servem para avaliar o desempenho de um classificador.
- Entre as que serão definidas a seguir, destacam-se as seguintes:
 - ➊ Acurácia
 - ➋ Sensibilidade (ou *recall*)
 - ➌ Especificidade (ou seletividade)
 - ➍ Precisão
 - ➎ F1-Score

Classificação Binária

Acurácia

Conceituação

- A **acurácia** é a probabilidade de detecção global do exame (ou, por extensão, do classificador).
- Usando-se a matriz de confusão, a acurácia é calculada como razão entre o número total acertos ($VP+VN$) e o número total de pessoas ou objetos testados ($VP+VN+FP+FN$).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (152)$$

- Quanto maior a acurácia, melhor é o desempenho geral do teste (e do classificador).

Classificação Binária

Sensibilidade

Conceituação

- A **sensibilidade** (ou *recall*) é a probabilidade de o classificador corretamente indicar positivo para a presença de certa doença, quando o paciente está realmente doente.
- Em outras palavras, é a habilidade do classificador de identificar os verdadeiros positivos (VP), sendo por isso também chamada de *taxa de VP* (*TP rate*).
- Matematicamente, tem-se

$$\text{Sensibilidade} = \frac{VP}{VP + FN} = 1 - FNR \quad (153)$$

em que $FNR = FN/P$ é a taxa de falsos negativos, sendo $P = VP + FN$ os casos totais de positivos.

Classificação Binária

Especificidade

Conceituação

- A **especificidade** ou seletividade é a probabilidade de o classificador corretamente indicar negativo, quando o paciente não está realmente doente.
- Em outras palavras, é a habilidade do classificador de identificar os verdadeiros negativos (VN), sendo por isso também chamada de *taxa de VN* (*TN rate*).
- Matematicamente, tem-se

$$\text{Especificidade} = \frac{VN}{VN + FP} = 1 - FPR \quad (154)$$

em que $FPR = FP/N$ é a taxa de falsos positivos, sendo $N = VN + FP$ os casos totais de negativos.

Classificação Binária

Especificidade e Sensibilidade

Exemplo

- Testes químicos de gravidez usam uma medida indireta (marcador) para detectar se uma mulher está grávida.
- Tais testes usam a gonadotrofina coriônica (hCG) presente na urina de mulheres grávidas.
- Como a hCG pode ser produzida também por um tumor, a especificidade de um teste moderno de gravidez não pode ser 100%, ou seja, podem ocorrer falsos positivos.
- Além disso, como a hCG está presente em pequenas concentrações após a fertilização e no início da embriogênese, a sensibilidade do teste de gravidez também não pode ser 100%, ou seja, podem ocorrer falsos negativos.

Classificação Binária

Precisão

Conceituação

- A **precisão** avalia a capacidade de um modelo de classificação em identificar apenas os dados relevantes. Neste caso, daqueles identificados como positivos, quantos são verdadeiramente positivos.
- Em exames/testes, corresponde à probabilidade de identificar entre os pacientes positivados, aqueles que são relevantes (ou seja, os verdadeiros positivos).

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (155)$$

Classificação Binária

F1-Score

Conceituação

- A medida **F1-Score** é a média harmônica entre precisão e sensibilidade.

$$\text{F1-Score} = 2 \cdot \frac{\text{precisão} \times \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}} \quad (156)$$

$$= \frac{2VP}{2VP + FP + FN} \quad (157)$$

- A medida F1-score atinge seu valor máximo em 1, correspondendo à uma precisão e sensibilidade perfeitas. Caso contrário, assume valores menores que 1.

Classificação Binária

Índices de Desempenho

Questão de Concurso



proframonsouza
Patrocinado

RESOLVENDO QUESTÃO

APRENDIZADO DE MÁQUINA



(FGV - 2022 – CGU – AUDITOR FEDERAL DE FINANÇAS E CONTROLE – TECNOLOGIA DA INFORMAÇÃO)

Considere uma matriz de confusão de um modelo de classificação binária de relatórios financeiros. O modelo classifica os relatórios em fraudulentos ou não fraudulentos.

Se essa matriz apresenta 200 verdadeiros positivos, 100 verdadeiros negativos, 40 erros do "tipo 1" e 20 erros do "tipo 2", podem-se calcular as métricas de desempenho aproximadas como:

- a) Precision = 0.71. Recall = 0.83;
- b) Precision = 0.83. Recall = 0.71;
- c) Precision = 0.83. Recall = 0.90;
- d) Precision = 0.90. Recall = 0.71;
- e) Precision = 0.90. Recall = 0.83.

←

Professor Ramon Souza

@proframonsouza
t.me/proframonsouza

Parte VIII

Detecção de Anomalias

Classificação Binária

Detecção de Anomalias

Definição

Tipo de classificação binária, também chamada de classificação de classe única (*one-class classification*).

- Útil quando o número de exemplos de uma das classes (N_1) é muito maior que os da outra classe (N_2): $N_1 \gg N_2$.
- A ideia é construir um modelo matemático apenas da classe com o maior número de exemplos.
- E em seguida definir um limiar de decisão (K), que será usado para definir se um dado vetor de atributos é portador de informação anômala; ou seja, estranha ao modelo construído.

Classificação Binária

Detecção de Anomalias

Aplicações Típicas

- Deteção de crises epilépticas convulsivas: costuma-se ter mais dados, na forma de sinais de EEG, representativos do estado *não-convulsivo* do que dados correspondentes ao estado convulsivo. Constroi-se então um modelo para representar a classe *não-convulsivo*.
- Deteção de falhas em equipamentos elétricos (e.g., motor): dados obtidos do funcionamento normal (i.e., não-faltoso) são mais comuns do que dados de falhas/faltas. Constroi-se então um modelo para representar o motor em seu funcionamento normal classe *não-falha*. Pode-se dizer que o modelo representa o motor em seu estado norma de funcionamento.

Classificação Binária

Detecção de Anomalias

Caso Univariado: $x \in \mathbb{R}$

- Suponha que estejamos monitorando uma grandeza física X que representa o estado de um processo industrial.
- Se há razões para assumir que $X \sim N(\mu, \sigma^2)$, então podemos usar a seguinte regra de decisão para detectar anomalias:

$$\text{SE } \left| \frac{X-\mu}{\sigma} \right| > K, \\ \text{ENTÃO } X \text{ é um estado anômalo.}$$

- Costuma-se adotar $K = 2$ ou 3 , desde que a variável seja gaussiana. Caso não seja, pode-se usar a transformação de Box-Cox ou outra do gênero.

Classificação Binária

Detecção de Anomalias

Caso p -variado: $\mathbf{x} \in \mathbb{R}^p$

- Suponha que estejamos medindo várias grandezas físicas, agrupadas em um vetor de atributos $\mathbf{x} \in \mathbb{R}^p$, que juntas representam o estado de um processo industrial, de um equipamento, ou até de um paciente.
- Se há razões para assumir que $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{C}_x)$, então podemos usar a seguinte regra de decisão para detectar anomalias no processo/equipamento/paciente:

$$\begin{array}{ll} \text{SE} & (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}_x^{-1} (\mathbf{x} - \boldsymbol{\mu}) > K, \\ \text{ENTÃO} & \mathbf{x} \text{ é um estado anômalo.} \end{array}$$

- Neste caso, o limiar de decisão pode ser calculado de duas maneiras, conforme será descrito a seguir.

Parte IX

Critério MAP e Classificadores Bayesianos Gaussianos