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

Pré-Requisitos

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

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.

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

O Que é Ser Inteligente?

- Seria ter habilidade específicas em certos domínios do conhecimento?

  (e.g. ser um neurocirurgião)

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)

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)

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?

Inteligência e Reconhecimento de Padrões

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

Inteligência e Reconhecimento de Padrões

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

Inteligência e Reconhecimento de Padrões

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

Inteligência e Reconhecimento de Padrões

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

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
  1. Comportamentais (fulano age sempre assim!)
  2. Sonoros (Este barulho não é normal!)
  3. Táteis (Estes tecidos têm texturas distintas!)
  4. Visuais (Acho que vai chover hoje!)
  5. Olfativos (De quem é esse perfume?)

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
  1. Comportamentais (fulano age sempre assim!)
  2. Sonoros (Este barulho não é normal!)
  3. Táteis (Estes tecidos têm texturas distintas!)
  4. Visuais (Acho que vai chover hoje!)
  5. Olfativos (De quem é esse perfume?)
  6. Numéricos (Quem pegou um dos doces?)

Inteligência e Reconhecimento de Padrões

- Seres vivos são habilidosos em reconhecer diversos padrões:
  1. Comportamentais (fulano age sempre assim!)
  2. Sonoros (Este barulho não é normal!)
  3. Táteis (Estes tecidos têm texturas distintas!)
  4. Visuais (Acho que vai chover hoje!)
  5. Olfativos (De quem é esse perfume?)
  6. 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.

Inteligência e Reconhecimento de Padrões

(a) Grupo 1

(b) Grupo 2

(c) Grupo 3

Em qual dos grupos acima você colocaria o objeto abaixo:

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.

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?

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?

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???

Um Computador Pode Reconhecer Padrões?

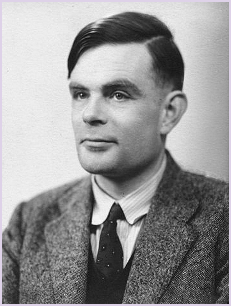


Figura : Alan Turing (23/06/1912 – 07/06/1954)[a]

[a] A. 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`

Um Computador Pode Reconhecer Padrões?

- Para comparar objetos precisamos de

Um Computador Pode Reconhecer Padrões?

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

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.

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 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.

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

  1. Qual seu formato?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

  1. Qual seu formato?
  2. É uma fruta cítrica?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

  1. Qual seu formato?
  2. É uma fruta cítrica?
  3. Qual a sua cor?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

    1. Qual seu formato?
    2. É uma fruta cítrica?
    3. Qual a sua cor?
    4. Qual a textura da casca?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?

1. Qual seu formato?
2. É uma fruta cítrica?
3. Qual a sua cor?
4. Qual a textura da casca?
5. Seu cheiro é ativo ou não?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
  1. Qual seu formato?
  2. É uma fruta cítrica?
  3. Qual a sua cor?
  4. Qual a textura da casca?
  5. Seu cheiro é ativo ou não?
- Todos esses atributos são igualmente importantes?

Um Computador Pode Reconhecer Padrões?

- Quais são os atributos que descrevem a tangerina?
  1. Qual seu formato?
  2. É uma fruta cítrica?
  3. Qual a sua cor?
  4. Qual a textura da casca?
  5. Seu cheiro é ativo ou não?
- Todos esses atributos são igualmente importantes?
- Quão difícil é a tarefa de definir os atributos de um objeto?

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

  1. Oval.

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

  1. Oval.
  2. Cítrico.

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

1. Oval.
2. Cítrico.
3. Alaranjada.

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
    1. Oval.
    2. Cítrico.
    3. Alaranjada.
    4. Rugosa.

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:

1. Oval.
2. Cítrico.
3. Alaranjada.
4. Rugosa.
5. Ativo.

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
  1. Oval.
  2. Cítrico.
  3. Alaranjada.
  4. Rugosa.
  5. Ativo.
- Provavelmente não!

Um Computador Pode Reconhecer Padrões?

- Respondendo às perguntas anteriores para a tangerina:
    1. Oval.
    2. Cítrico.
    3. Alaranjada.
    4. Rugosa.
    5. Ativo.
- Provavelmente não!
- Horrivelmente árdua! :-(

Um Computador Pode Reconhecer Padrões?

- Porém, um computador só entende números!!

Um Computador Pode Reconhecer Padrões?

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

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!

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

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!

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

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

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

Um Computador Pode Reconhecer Padrões?

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

Um Computador Pode Reconhecer Padrões?

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

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

Um Computador Pode Reconhecer Padrões?

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

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

$$x_2\text{: } \{\text{não, sim}\} = \{0, 1\}$$

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} $= \{0, 1, 2, 3, 4\}$

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

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} $= \{0, 1\}$

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.

Um Computador Pode Reconhecer Padrões?

- Usando a convenção estabelecida para os atributos:

Um Computador Pode Reconhecer Padrões?

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

Um Computador Pode Reconhecer Padrões?

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

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**: $\mathbf{z} = [0 \;\; 1 \;\; 2 \;\; 1 \;\; 1]$

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,

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 $\mathbf{x} = [0 \;\; 1 \;\; 2 \;\; 1 \;\; 0]$?

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 $\mathbf{y} = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0]$?

Componentes de um Problema de Classificação

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

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

em que

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

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 $\mathbf{x}$ ainda não-classificado, realizar uma busca em $\mathcal{X}$ pelo índice do vetor de atributos mais próximo de $\mathbf{x}$:

$$\boxed{i^* = \arg \min_{i=1,\ldots,N} \|\mathbf{x} - \mathbf{x}_i\|^2} \quad (2)$$

em que $\|\mathbf{v}\|$ denota a norma euclidiana do vetor $\mathbf{v}$.

Passo 3 - Atribuir a $\mathbf{x}$ a mesma classe que $\mathbf{x}_{i^*}$.

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 a $\mathbf{x}$ a mesma classe que $\mathbf{x}_{i^*}$, se

$$\boxed{\|\mathbf{x}-\mathbf{x}_{i^*}\|^2 < \|\mathbf{x}-\mathbf{x}_i\|^2, \quad \forall i \neq i^*} \qquad (3)$$

- Nas Expressões (2) e (3) usa-se a norma euclidiana quadrática, por dois motivos: (i) evitar o cálculo da raiz quadrada para ter a norma do vetor argumento. (ii) O valor mínimo (ou máximo) de uma função não se altera se a ela for aplicada uma operação matemática monotônica.
- Por exemplo, o mínimo da função $f(x) = |x|$ é o mesmo que o da função $g(x) = [f(x)]^2$.

Classificador Centróide Mais Próximo (MDC, sigla em Inglês)

Passo 1 - Encontrar o vetor centróide de cada classe:

$$\boxed{\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x}} \qquad (4)$$

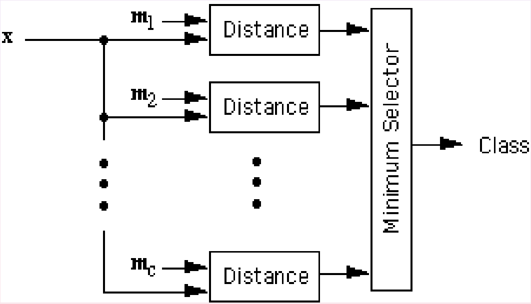
em que N_i é o número de exemplos da i-ésima classe (cujo rótulo é ω_i), $i = 1, ..., K$.

Passo 2 - Atribuir a $\mathbf{x}$ a mesma classe que $\mathbf{m}_{i^*}$, se

$$\boxed{\|\mathbf{x} - \mathbf{m}_{i^*}\|^2 < \|\mathbf{x} - \mathbf{m}_i\|^2, \quad \forall i \neq i^*} \qquad (5)$$

em que $\|\mathbf{v}\|$ denota a norma euclidiana do vetor $\mathbf{v}$.

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



- 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.

Definição: Medidas de Dissimilaridade

- Se d_{rs} é uma medida de dissimilaridade entre os objetos[a] r e s, então deve satisfar 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).

[a] Tais objetos podem ser vetores, por exemplo. Ou nuvens (distribuições) de pontos.

- **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!

Dissimilaridade vs. Similaridade

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

$$\boxed{s_{ij} = \frac{1}{1 + d_{ij}}} \tag{6}$$

ou como

$$\boxed{s_{ij} = c - d_{ij}} \tag{7}$$

onde c é uma constante positiva.

Definição: Métrica

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

$$\boxed{d_{rs} \leq d_{rt} + d_{ts}, \quad \forall r, s, t} \tag{8}$$

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.

Definição: Distância de Minkowski de Ordem m

- A *distância de Minkwoski* 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 entre $\mathbf{x}$ e $\mathbf{y}$ é definida pela seguinte expressão:

$$\boxed{d_M^{(m)}(\mathbf{x}, \mathbf{y}) = \left(\sum_{j=1}^{p} |x_j - y_j|^m \right)^{\frac{1}{m}}} \quad (9)$$

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

Se $m = 1 \Rightarrow$ Distância Quarteirão

- Se fizermos $m = 1$ na Equação (9) obtemos a expressão da distância quarteirão (*city block*, em Inglês):

$$\begin{aligned} d_{cb}(\mathbf{x}, \mathbf{y}) &= |\mathbf{x} - \mathbf{y}| && (10) \\ &= \sum_{j=1}^{p} |x_j - y_j| && (11) \\ &= |x_1 - y_1| + \cdots + |x_p - y_p| && (12) \end{aligned}$$

Se $m = 2 \Rightarrow$ Distância Euclidiana

- Se fizermos $m = 2$ na Equação (9) obtemos a conhecida expressão da distância euclidiana:

$$
\begin{aligned}
d_e(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} - \mathbf{y}\| && (13)\\
&= \left(\sum_{j=1}^{p} |x_j - y_j|^2 \right)^{\frac{1}{2}} && (14)\\
&= \sqrt{\sum_{j=1}^{p} |x_j - y_j|^2} && (15)\\
&= \sqrt{|x_1 - y_1|^2 + \cdots + |x_p - y_p|^2} && (16)
\end{aligned}
$$

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

- Se fizermos $m \to \infty$ na Equação (9) obtemos a expressão da distância de Chebyshev:

$$\begin{aligned} d_{ch}(\mathbf{x}, \mathbf{y}) &= \lim_{m \to \infty} \left(\sum_{j=1}^{p} |x_j - y_j|^m \right)^{\frac{1}{m}} && (17) \\ &= \max\{|x_1 - y_1|, \ldots, |x_p - y_p|\} && (18) \\ &= \max_i |x_i - y_i| && (19) \end{aligned}$$

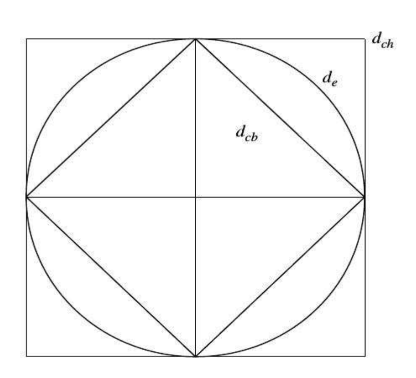
Desafio!

Demonstrar a passagem da expressão mostrada na Equação (15) para a da Equação (16), tanto computacionalmente quanto matematicamente.

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/extração 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.

- 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.



Distância Quadrática

- A distância quadrática, também chamada de distância de Mahalanobis, é outra medida de dissimilaridade muito utilizada em RP.
- Sua expressão geral é dada por

$$d_q(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{Q} (\mathbf{x}-\mathbf{y})} \quad (20)$$

$$= \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} (x_i - y_i) Q_{ij} (x_j - y_j)} \quad (21)$$

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

Distância Quadrática: Algumas considerações

- Uma matriz $\mathbf{A}$ é definida positiva se, para qualquer vetor $\mathbf{v}$, obtermos sempre

$$\mathbf{v}^T\mathbf{A}\mathbf{v} > 0 \tag{22}$$

- Assim, a matriz $\mathbf{Q}$ deve ser positiva definida para que o termo $(\mathbf{x}-\mathbf{y})^T\mathbf{Q}(\mathbf{x}-\mathbf{y})$ seja positivo, garantindo que não haja argumentos negativos na função raiz quadrada na Eq. (20).

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

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{e} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{23}$$

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

$$\mathbf{Q} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \tag{24}$$

em que a, b, c e d são números reais.

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

Assim, substituindo na Eq. (20), temos

$$
\begin{aligned}
d_q(\mathbf{x},\mathbf{y}) &= \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}}, \qquad (25) \\
&= \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 \end{bmatrix} \begin{bmatrix} a(x_1 - y_1) + b(x_2 - y_2) \\ c(x_1 - y_1) + d(x_2 - y_2) \end{bmatrix}},
\end{aligned}
$$

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.

Distância quadrática vs. distância euclidiana

- Se fizermos $\mathbf{Q} = \mathbf{I}_p$ na Eq. (20), obtemos a seguinte expressão:

$$
\begin{aligned}
d_q(\mathbf{x}, \mathbf{y}) &= \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{I}_p (\mathbf{x} - \mathbf{y})} && (26) \\
&= \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} && (27) \\
&= \sqrt{\|\mathbf{x} - \mathbf{y}\|^2} && (28) \\
&= \|\mathbf{x} - \mathbf{y}\|. && (29)
\end{aligned}
$$

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

Parte I

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 dos dados (i.e. média=0, variância=1).

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:

$$\boxed{\mathbf{x} = \frac{\mathbf{x}}{\|\mathbf{x}\|}} \quad \text{e} \quad \boxed{\mathbf{m}_i = \frac{\mathbf{m}_i}{\|\mathbf{m}_i\|}} \tag{30}$$

Método 1: Norma Constante

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

$$\mathbf{x} = \begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} \tag{31}$$

- A norma deste vetor é calculada como

$$\|\mathbf{x}\| = \sqrt{(\sqrt{3})^2 + (3)^2 + (-2)^2} = \sqrt{16} = 4. \tag{32}$$

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

$$\mathbf{x}^* = \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{1}{4}\begin{bmatrix} \sqrt{3} \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} \sqrt{3}/4 \\ 3/4 \\ -1/2 \end{bmatrix} \tag{33}$$

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.

$$\mathbf{x}^* = \frac{1}{\|\mathbf{x}\|}\mathbf{x} = \alpha\mathbf{x}, \tag{34}$$

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*.

Método 2: Mudança de escala

- A normalização pelo Método 1 é particularmente útil para o classificador de máxima correlação.
- 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 procedimento de *normalização global*.

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

$$\boxed{x^* = \frac{x - x_{min}}{x_{max} - x_{min}}} \quad (35)$$

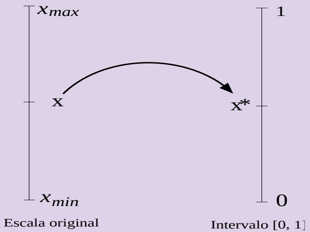


Figura : Mudança da escala original de x para o intervalo [0,1].

Método 2: Exemplo 1

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $x_{1,min} = 11,03$ e $x_{1,max} = 14,83$.
- Assim, a função de normalização é dada por

$$x_1^* = \frac{x_1 - x_{1,min}}{x_{1,max} - x_{1,min}} = \frac{x_1 - 11,03}{14,83 - 11,03} = \frac{x_1 - 11,03}{3,80} \quad (36)$$

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

$$x_1^* = \frac{13,50 - 11,03}{3,80} = 0,65. \quad (37)$$

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

$$\boxed{x^* = 2\left(\frac{x - x_{min}}{x_{max} - x_{min}}\right) - 1} \qquad (38)$$

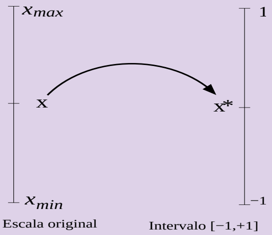


Figura : Mudança da escala original de x para o intervalo [-1,+1].

Método 2: Exemplo 2

- Considere o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $x_{1,min} = 11,03$ e $x_{1,max} = 14,83$.
- Assim, a função de normalização é dada por

$$x_1^* = 2\left(\frac{x_1 - x_{1,min}}{x_{1,max} - x_{1,min}}\right) - 1 = 2\left(\frac{x_1 - 11,03}{3,80}\right) - 1 \quad (39)$$

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

$$x_1^* = 2\left(\frac{13,50 - 11,03}{3,80}\right) - 1 = 0,30. \quad (40)$$

Método 3: Padronização da variável (média=0, 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 ($\bar{x}$) e do desvio-padrão (σ_x) da variável x.
- Por isso, a padronização também pode ser chamada de normalização estatística.
- Este procedimento também é um tipo de *normalização global*.

Método 3: Padronização da variável (média=0, variância=1)

- A normalização estatística é dada por

$$\boxed{x^* = \frac{x - \bar{x}}{\sigma_x}} \tag{41}$$

com a média e desvio-padrão amostrais de x calculados como

$$\boxed{\bar{x} = \frac{\sum_{n=1}^{N} x(n)}{N}} \quad \text{e} \quad \boxed{\sigma_x = \sqrt{\left(\frac{\sum_{n=1}^{N}(x(n) - \bar{x})^2}{N-1}\right)}} \tag{42}$$

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

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:

$$\begin{aligned} x^* &= \frac{x - x_{min}}{x_{max} - x_{min}} \\ &= \left(\frac{1}{x_{max} - x_{min}}\right) x - \left(\frac{x_{min}}{x_{max} - x_{min}}\right) \\ &= ax + b \end{aligned}$$

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

Propriedade 1 dos Métodos de Normalização

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

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 PDF é denotada por $f_X(x)$.
- Seja $y \in \mathbb{R}$ a variável aleatória resultante da seguinte operação matemática sobre x: $y = g(x)$.
- Pode-se mostrar facilmente que a PDF de y é dada por

$$f_Y(y) = \frac{f_X(x)}{\left|\frac{dy}{dx}\right|} \qquad (43)$$

- 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!)

Propriedade 2 dos Métodos de Normalização

- Por serem transformações lineares, as normalizações descritas anteriormente não alteram a natureza (ou tipo) da distribuição da variável normalizada em relação à variável original não-normalizada.
- Em outras palavras, o tipo de PDF (*probability density function*) da variável permanece o mesmo. Por exemplo, se a PDF for gaussiana, continua gaussiana após a transformação.
- 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.
- A variável normalizada pelo Método 2, tem a mesma unidade da variável original. Se for normalizada pelo

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`.

Método 3: Exemplo

- Usando o atributo X_1 (teor alcoólico) do conjunto de dados `wine.dat`.
- Para esta variável temos $\bar{x} = 13,00$ e $\sigma_1 = 0,81$.
- Assim, a função de normalização é dada por

$$x_1^* = \frac{x_1 - 13,00}{0,81} \tag{44}$$

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

$$x_1^* = \frac{13,50 - 13,00}{0,81} = 0,617. \tag{45}$$

Parte II

Coeficiente de Correlação

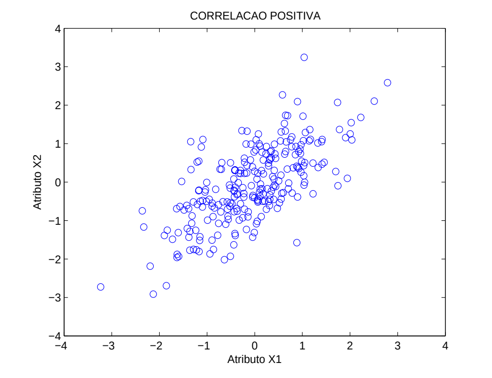
Objetivos

Objetivo: Entender como duas variáveis estão interrelacionadas do ponto de vista estatístico.

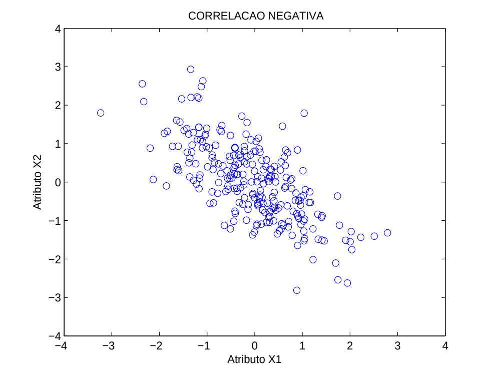
- **Método 1:** Gráfico de dispersão (qualitativo).
- **Método 2:** Coeficiente de Correlação (quantitativo).

Objetivo: Discutir relação como modelo de regressão linear simples.

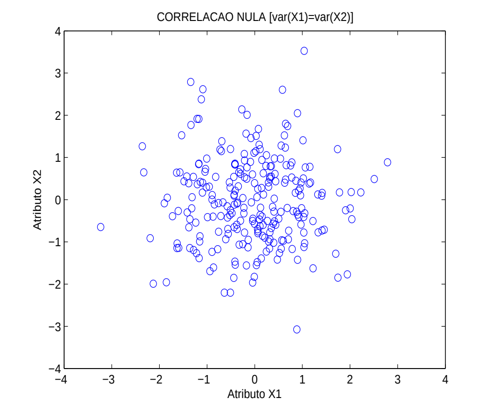
- **Correlação Positiva**: ocorre quando um atributo tende a aumentar, o outro também tende a aumentar.



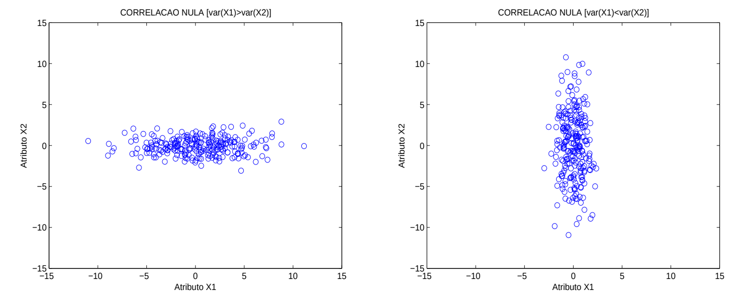
- **Correlação Negativa**: ocorre quando um atributo tende a aumentar, o outro também tende a aumentar.



- **Correlação Nula [Caso 1: var(X1)=var(X2)]**: não há um padrão definido de tendência.



- **Correlação Nula [Caso 2: var(X1) $\neq$ var(X2)]**: também ocorre quando ao aumentar um atributo não há mudança significativa nos valores do outro atributo.



Definição

- O coeficiente de correlação entre duas variáveis aleatórias quaisquer é dado pela seguinte expressão:

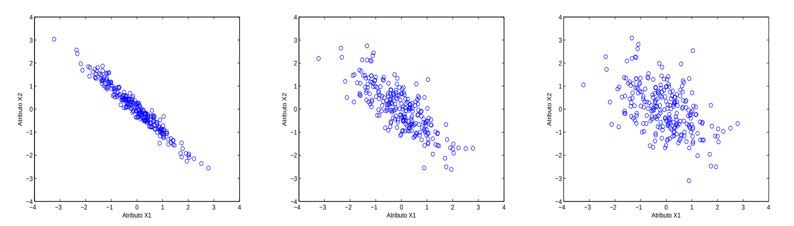
$$\boxed{\rho_{12} = \frac{\text{cov}(X_1, X_2)}{\text{dp}(X_1) \cdot \text{dp}(X_2)} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2}} \quad (46)$$

em que

$$\begin{aligned} \sigma_{12} &= \text{cov}(X1, X2) = \text{covariância entre } X_1 \text{ e } X_2. \\ \sigma_1 &= \text{dp}(X_1) = \text{desvio-padrão de } X_1 = \sqrt{\text{var}(X_1)} \\ \sigma_2 &= \text{dp}(X_2) = \text{desvio-padrão de } X_2 = \sqrt{\text{var}(X_2)} \end{aligned}$$

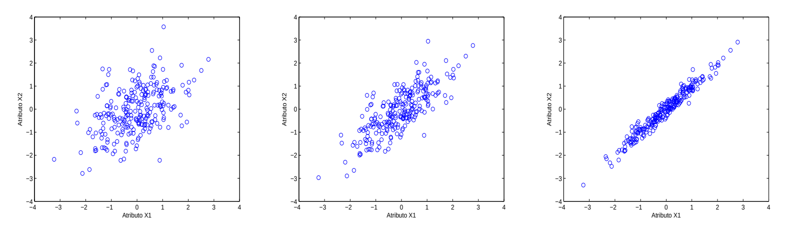
- Note que: $-1 \leq \rho(X_1, X_2) \leq 1$.

- Se $-1 < \rho_{12} < 0$, então X_1 e X_2 estão **negativamente** correlacionadas.
- Quanto mais próximo ρ_{12} estiver de -1, maior é a correlação negativa entre X_1 e X_2.



$\rho_{12} = -0.98$ $\rho_{12} = -0.8$ $\rho_{12} = -0.5$

- Se $0 < \rho_{12} \leq 1$, então X_1 e X_2 estão **positivamente** correlacionadas.
- Quanto mais próximo ρ_{12} estiver de 1, maior é a correlação positiva entre X_1 e X_2.



$\rho_{12} = 0.5$ $\rho_{12} = 0.8$ $\rho_{12} = 0.98$

- A variância de uma variável aleatória é uma medida de sua dispersão (i.e. espalhamento) em torno de seu valor médio.
- Para um conjunto de N observações de X qualquer, sua variância amostral é dada por

$$\boxed{\sigma_x^2 = \frac{\sum_{n=1}^{N}(x(n) - \bar{x})^2}{N-1}} \tag{47}$$

em que $x(n)$ é a n-ésima observação de X e $\bar{x}$ é a média amostral de X calculada como

$$\boxed{\bar{x} = \frac{\sum_{n=1}^{N} x(n)}{N}} \tag{48}$$

- O desvio padrão de X é definido como $\sigma_x = \sqrt{\sigma_x^2}$.

- A covariância entre atributos é a média dos produtos dos desvios de cada atributo em relação às suas respectivas médias.
- Para N observações conjuntas de (X_1, X_2), a covariância amostral entre X_1 e X_2 é dada por

$$\boxed{\sigma_{12} = \frac{\sum_{n=1}^{N}(x_1(n) - \bar{x}_1)(x_2(n) - \bar{x}_2)}{N-1}} \tag{49}$$

  em que $x_1(n)$ e $x_2(n)$ denotam as n-ésimas observações de X_1 e X_2, respectivamente; enquanto $\bar{x}_1$ e $\bar{x}_2$ são as médias amostrais correspondentes.
- Note que se $X_1 = X_2$, temos que $\sigma_{12} = \sigma_1^2 = \sigma_2^2$. Ou seja, covariância reduz-se à variância, se $X_1 = X_2$.

Implementação

- Dadas N observações conjuntas de (X_1, X_2), a covariância entre X_1 e X_2 e os desvios-padrão de X_1 e X_2 podem ser facilmente calculados em diferentes softwares.
  - No Matlab/Octave, usar os comandos `cov` e `std`.
  - No Excel[a], usar os comandos `COVAR` e `DESVPAD`.
  - No LibreOffice Calc, usar os comandos `COVAR` e `DESVPAD`.

[a] www.bertolo.pro.br/matematica/Disciplinas/3ano/Estatistica/Bimestre2/EstatisticaAplicada3.pdf

Implementação

- Dadas N observações conjuntas de (X_1, X_2), o coeficiente de correlação (ρ_{12}) também pode ser calculado diretamente em vários pacotes de software.
  - No Matlab/Octave, usar os comandos `corrcoef`.
  - No Excel[a], usar o comando `CORREL`.
  - No LibreOffice Calc, usar o comando `CORREL`.

[a] www.bertolo.pro.br/matematica/Disciplinas/3ano/Estatistica/Bimestre2/EstatisticaAplicada3.pdf

- Uma interessante relação teórica entre o gráfico de espalhamento e o coeficiente de correlação pode ser verificada através da inclinação da reta de regressão linear (ou linha de tendência)

$$y = ax + b \qquad (\text{ou } x_2 = ax_1 + b), \tag{50}$$

ajustada aos dados pelo método dos mínimos quadrados.

- As constante a e b são chamadas, respectivamente, de inclinação e intercepto da reta de tendência, sendo calculados como

$$\boxed{a = \left(\frac{\sigma_2}{\sigma_1}\right)\rho_{12}} \qquad \text{e} \qquad \boxed{b = \bar{x}_2 - a\bar{x}_1}. \tag{51}$$

- Note que no slide anterior escrevemos a inclinação da reta de regressão como sendo diretamente proporcional ao coeficiente de correlação das variáveis X_1 e X_2, ou seja

$$\boxed{a = k \cdot \rho_{12}}, \tag{52}$$

em que a constante de proporcionalidade é dada pela razão entre os desvios-padrão de X_2 e X_1:

$$\boxed{k = \frac{\sigma_2}{\sigma_1}}. \tag{53}$$

- Se os desvios-padrão forem o mesmo para as duas variáveis, temos que a inclinação será igual ao coeficiente de correlação: $a = \rho_{12}$, se $\sigma_1 = \sigma_2$.

Implementação

- Dadas N observações conjuntas de (X_1, X_2), a inclinação da linha de tendência ajustada a este conjunto de observações pode ser facilmente calculada em planilhas numéricas.
  - No Excel e no LibreOffice Calc, usar o comando `INCLINAÇÃO`.

Parte III

Classificador Quadrático

- Costuma-se dispor as covariâncias entre todas as variáveis, tomadas duas a duas, σ_{ij}, $i, j = 1, ..., p$, 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 $i \neq j$.

- 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

$$\boxed{\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j}} \qquad (54)$$

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

$$\boxed{\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j} \qquad (55)$$

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:

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

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

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

$$|\mathbf{C_x}| = \sigma_1^2 \cdot \sigma_2^2 \cdots \sigma_p^2 = \prod_{i=1}^{p} \sigma_i^2 \tag{56}$$

- A inversa de tal matriz também é facilmente calculada

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

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*:

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

   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.

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 **A** pode de fato ser uma matriz de covariância.

Matriz de Covariância no Matalb/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`

Classificador Quadrático

Passo 1 - Determinar o vetor centróide de cada classe:

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} \mathbf{x} \tag{57}$$

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

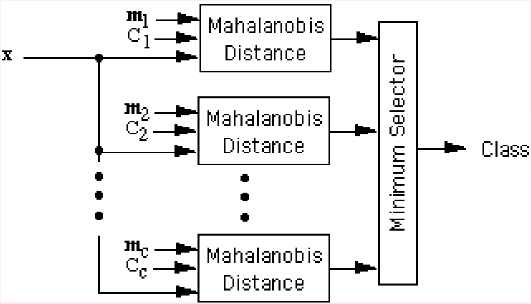
Passo 2 - Determinar a matriz de covariância de cada classe:

$$\mathbf{C}_i = \frac{1}{N_i} \sum_{\forall \mathbf{x} \in \omega_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \tag{58}$$

Passo 3 - Atribuir a $\mathbf{x}$ a mesma classe que $\mathbf{m}_{i^*}$, se

$$(\mathbf{x}-\mathbf{m}_{i^*})^T \mathbf{C}_{i^*}^{-1} (\mathbf{x}-\mathbf{m}_{i^*}) < (\mathbf{x}-\mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x}-\mathbf{m}_i), \quad \forall i \neq i^* \tag{59}$$

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



Distância Quadrática em Classificação e Normalização

- Ao usar da distância quadrática em classificação, 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.
- Para simplificar, usaremos uma única matriz de covariância para todas as classes:

$$\mathbf{C}_i = \mathbf{C}_{\mathbf{x}} = \text{diag}(\sigma_1^2, \sigma_2^2) = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}. \tag{60}$$

- Deste modo, a matriz inversa é dada por como

$$\mathbf{Q} = \mathbf{C}_{\mathbf{x}}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \tag{61}$$

Distância Quadrática em Classificação e Normalização

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

$$
\begin{aligned}
d_q(\mathbf{x},\mathbf{m}_i) &= \sqrt{\left[(x_1-m_{i1}) \quad (x_2-m_{i2})\right]\begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}\begin{bmatrix} (x_1-m_{i1}) \\ (x_2-m_{i2}) \end{bmatrix}} \\
&= \sqrt{\left[(x_1-m_{i1}) \quad (x_2-m_{i2})\right]\begin{bmatrix} \frac{(x_1-m_{i1})}{\sigma_1^2} \\ \\ \frac{(x_2-m_{i2})}{\sigma_2^2} \end{bmatrix}},
\end{aligned}
$$

Distância Quadrática em Classificação e Normalização

- Que resulta na seguinte expressão:

$$d_q(\mathbf{x}, \mathbf{m}_i) = \sqrt{\left(\frac{x_1 - m_{i1}}{\sigma_1}\right)^2 + \left(\frac{x_2 - m_{i2}}{\sigma_2}\right)^2} \quad (62)$$

- 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 (62) com a expressão da normalização estatística em (41), repetida abaixo:

$$\boxed{x^* = \frac{x - \bar{x}}{\sigma_x}} \quad (63)$$

Parte IV

Estimação da Matriz de Covariâcia

Estimação do Vetor de Médias e da Matriz de Correlação

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

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

- O vetor médio deste conjunto de observações é estimado pela seguinte expressão:

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}(n) \tag{64}$$

- A matriz de correlação $\mathbf{R_x}$ é estimada pela seguinte expressão:

$$\hat{\mathbf{R}}_{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}(n)\mathbf{x}^T(n) \tag{65}$$

Estimação da Matriz de Correlação - Forma Matricial

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

$$\mathbf{X} = [\mathbf{x}(1) \,|\, \mathbf{x}(2) \,|\, \cdots \,|\, \mathbf{x}(N)] \tag{66}$$

- Desta forma, a matriz de correlação pode ser estimada por meio da seguinte expressão:

$$\hat{\mathbf{R}}_{\mathbf{x}} = \frac{1}{N}\mathbf{X}\mathbf{X}^T \tag{67}$$

- Esta última expressão é particularmente útil na economia de notação matemática e para implementação em ambientes de computação numérica, tais como Matlab, Octave e Scilab.

Estimação da Matriz de Covariância - Forma Vetorial

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

$$\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} [\mathbf{x}(n) - \mathbf{m}][\mathbf{x}(n) - \mathbf{m}]^T \qquad (68)$$

- Ou então, através da seguinte expressão:

$$\hat{\mathbf{C}}_{\mathbf{x}} = \hat{\mathbf{R}}_{\mathbf{x}} - \mathbf{m}\mathbf{m}^T \qquad (69)$$

Estimação da Matriz de Covariância - Forma Matricial

- Usando a definição da matriz de dados $\mathbf{X}$, dada em (66), a matriz de covariância pode ainda ser estimada por meio da seguinte expressão:

$$\hat{\mathbf{C}}_{\mathbf{x}} = \frac{1}{N}[\mathbf{X} - \mathbf{M}][\mathbf{X} - \mathbf{M}]^T \quad (70)$$

em que $\mathbf{M} \in \mathbb{R}^{p \times N}$ é uma matriz cujas colunas são formadas pelo vetor médio $\mathbf{m}$, repetido N vezes, ou seja:

$$\mathbf{M} = [\mathbf{m} \,|\, \mathbf{m} \,|\, \cdots \,|\, \mathbf{m}] \quad (71)$$

Estimação do Vetor de Médias - Forma Recursiva

- A Equação (64) 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 seqüencial. Assim, deve-se estimar tal vetor recursivamente. Por exemplo:

$$\begin{aligned}\mathbf{m}(n) &= \frac{n-1}{n}\mathbf{m}(n-1) + \frac{1}{n}\mathbf{x}(n) \\ &= \alpha\mathbf{m}(n-1) + (1-\alpha)\mathbf{x}(n)\end{aligned} \tag{72}$$

  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 $p \times 1$.

Estimação da Matriz de Correlação - Forma Recursiva

- As técnicas anteriores utilizam todos os vetores aleatórios de uma única vez na estimação das matrizes $\mathbf{C_x}$ e $\mathbf{R_x}$.
- Em muitos casos, os vetores só estão disponíveis um-a-um, de forma seqüencial. Assim, deve-se estimar tais matrizes recursivamente. Por exemplo:

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{x}}(n) &= \frac{n-1}{n}\hat{\mathbf{R}}_{\mathbf{x}}(n-1) + \frac{1}{n}\mathbf{x}(n)\mathbf{x}^T(n) \\ &= \alpha\hat{\mathbf{R}}_{\mathbf{x}}(n-1) + (1-\alpha)\mathbf{x}(n)\mathbf{x}^T(n) \end{aligned} \tag{73}$$

  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 $\hat{\mathbf{R}}_{\mathbf{x}}(0) = \mathbf{I}_p$ = matriz identidade $p \times p$.

Exemplo no Matlab: Funções nativas `mean()` e `cov()`

- Os comandos a seguir ilustram como determinar o vetor de médias e a matriz de covariância para um conjunto de N observações de um vetor aleatório gaussiano de dimensão 2:

```
» p=2;     % dim.  vetor aleatorio
» N=5000;     % total de observacoes
» X=randn(N, p); % gera vetores aleatorios
» M=mean(X)
   M = 0.0097     -0.0027
» Cx=cov(X)
   Cx = 1.0224     -0.0124
       -0.0124      0.9887
```

Exemplo no Matlab: Função `mcovar()`

- A Eq. (68) pode ser implementada pela seguinte rotina:

```
function C=mcovar(X)

[N p]=size(X);
soma=zeros(p);
[N p]=size(X);
M=sum(X)/N; % vetor de medias
for i=1:N,
   soma = soma + (X(i,:)-M)'*(X(i,:)-M);
end
C=soma/N; % Matriz de covariancia
```

- Para executar no Matlab: » `Cx = mcovar(X);`

Estimando a Matriz de Covariância pela Função `mcovar()`

- A matriz estimada pela rotina `mcovar` é dada por:

```
» Cx=mcovar(X) % Funcao criada no Matlab
    Cx = 1.0222      -0.0124
        -0.0124       0.9885
```

Exercício Computacional Proposto

- Implementar a Eq. (69) como uma rotina do Matlab chamada `mcovar2` e comparar o resultado com o gerado pela função pronta `cov` e pela rotina `mcovar`.

Parte V

Distribuição Gaussiana Multivariada

FDP Gaussiana Multivariada

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

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

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

- Resumidamente, escrevemos que um vetor aleatório $\mathbf{x} \in \mathbb{R}^p$ está distribuído segundo uma FDP gaussiana multivariada da seguinte forma:

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

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*.

FDP Gaussiana Bivariada

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

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ e } \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{74}$$

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

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

FDP Gaussiana Bivariada (cont.-1)

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

$$\mathbf{C}_{\mathbf{x}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{75}$$

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

$$\boxed{\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}} \tag{76}$$

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

FDP Gaussiana Bivariada (cont.-2)

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

$$\begin{aligned} |\mathbf{C}_\mathbf{x}| &= \begin{vmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{vmatrix} \\ &= \sigma_1^2\sigma_2^2 - (\rho\sigma_1\sigma_2)(\rho\sigma_1\sigma_2) \\ &= (1-\rho^2)\sigma_1^2\sigma_2^2 \end{aligned} \tag{77}$$

- E a inversa da matriz de covariância é dada por:

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

FDP Gaussiana Bivariada (cont.-3)

- Finalmente, podemos escrever a FDP gaussiana bivariada como:

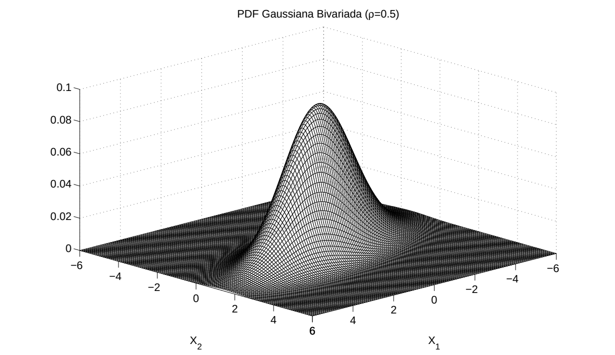
$$f_{\mathbf{x}}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q(x_1, x_2)\right\}, \qquad (78)$$

em que

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

FDP Gaussiana Bivariada (cont.-4)

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



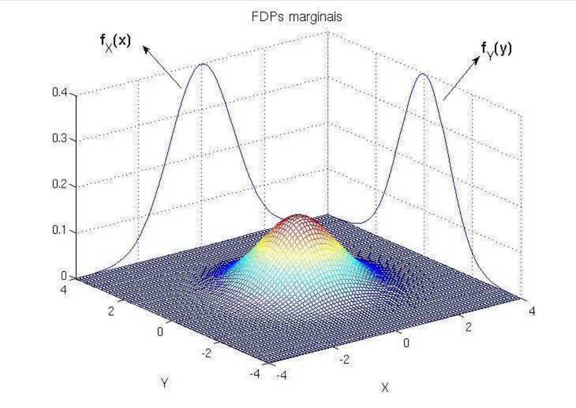
FDPs Marginais da FDP Gaussiana Bivariada

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

$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(x_1, x_2) dx_2 \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right\} \\ f_{X_2}(x_2) &= \int_{-\infty}^{\infty} f_{\mathbf{x}}(x_1, x_2) dx_1 \\ &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right\} \end{aligned}$$

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

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



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:

$$f_{\mathbf{x}}(x_1|x_2) = \frac{f_{\mathbf{x}}(x_1, x_2)}{f_{x_2}(x_2)} \quad \text{ou} \quad f_{\mathbf{x}}(x_2|x_1) = \frac{f_{\mathbf{x}}(x_1, x_2)}{f_{x_1}(x_1)} \qquad (79)$$

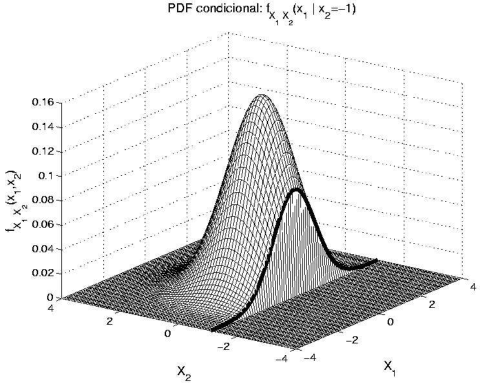
- Para um valor específico de x_1 ou x_2, digamos x^*, estas expressões podem ser escritas como:

$$f_{\mathbf{x}}(x_1|x_2 = x^*) = \frac{f_{\mathbf{x}}(x_1, x_2 = x^*)}{f_{x_2}(x_2 = x^*)} \quad \text{ou} \qquad (80)$$

$$f_{\mathbf{x}}(x_2|x_1 = x^*) = \frac{f_{\mathbf{x}}(x_1 = x^*, x_2)}{f_{x_1}(x_1 = x^*)} \qquad (81)$$

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 $\rho = 0$.



FDP Gaussiana Bivariada com $\rho = 0$.

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

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

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

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

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

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

- Ou seja, podemos escrever:

$$\begin{aligned} f_{\mathbf{x}}(x_1, x_2) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{\left\{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}} \\ &= f_{X_1}(x_1) \cdot f_{X_2}(x_2) \end{aligned} \quad (82)$$

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. (82) revela uma importante exceção à regra geral:

  Variáveis aleatórias gaussianas não-correlacionadas são **sempre** *independentes!!*

Parte VI

Contornos da 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**.

Contornos da FDP Gaussiana Bivariada

- Uma curva de contorno é a curva formada pelo **lugar geométrico** dos pontos (x_1, x_2) que satisfazem a equação $f_{\mathbf{x}}(x_1, x_2) = C$, em que $C > 0$ é uma constante.
- Matematicamente, isto é o mesmo que escrever:

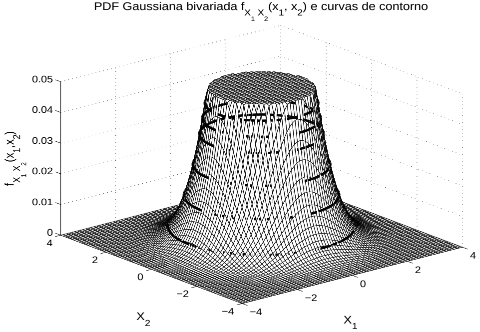
$$f_{\mathbf{x}}(x_1, x_2) \propto \exp\left\{-\frac{1}{2}Q(x_1, x_2)\right\} = C, \tag{83}$$

em que

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

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.



Contornos da FDP Gaussiana Bivariada (cont.-2)

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

$$\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} = C \quad (84)$$

- A forma da curva da Eq. (84) 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** ($\rho \neq 0$)

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

- Neste caso, a Eq. (84) reduz-se à seguinte expressão:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = C \tag{85}$$

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

$$\boxed{\frac{(x_1 - \mu_1)^2}{(\sqrt{C}\sigma_1)^2} + \frac{(x_2 - \mu_2)^2}{(\sqrt{C}\sigma_2)^2} = 1} \tag{86}$$

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

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

- Temos 3 casos a considerar:

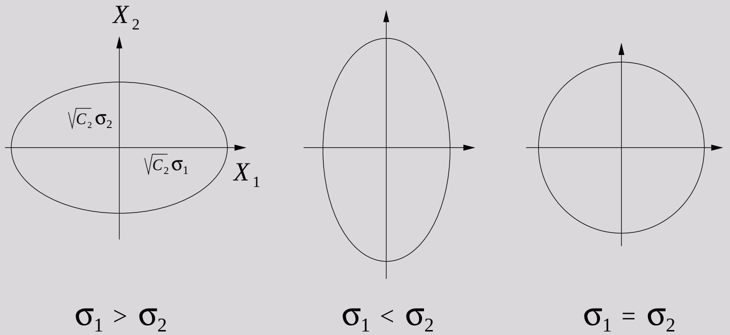
  $\sigma_1 > \sigma_2$ - Elipse com eixo maior ao longo da reta $x_2 = 0$, ou seja, mais alongada na horizontal.

  $\sigma_1 < \sigma_2$ - Elipse com eixo maior ao longo da reta $x_1 = 0$, ou seja, mais alongada na vertical.

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

- A figura no slide a seguir ilustra os três casos supracitados.

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



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

- Neste caso, a Eq. (84), continua representando uma curva de contorno em forma de elipse.
- Contudo, torna-se importante conhecer o efeito introduzido pelo termo
$$-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}, \tag{87}$$
no desenho da elipse.
- Para facilitar a análise, vamos assumir $\mu_1 = \mu_2 = 0$ e $\sigma_1 = \sigma_2 = 1$.
- Assim, a Eq. (84) passa a ser escrita da seguinte maneira:
$$x_1^2 - 2\rho x_1 x_2 + x_2^2 = C \tag{88}$$

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

- A Eq. (88) pode ser reorganizada de tal modo a representar uma equação do segundo grau em x_2:

$$a x_2^2 + b x_2 + c = 0 \tag{89}$$

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

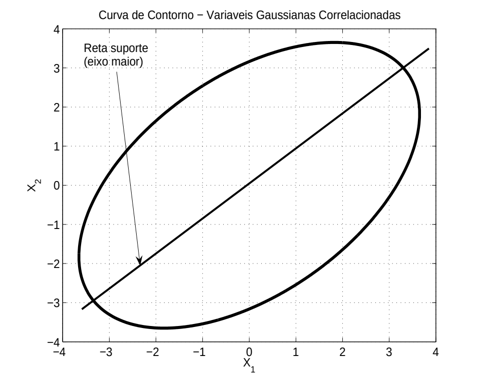
- As raízes desta equação são calculadas pela seguinte expressão:

$$x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{90}$$

- Usando a Eq. (89), 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. (90).

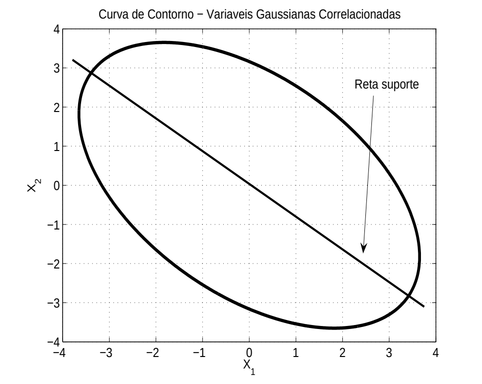
Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$) (cont.-2)

- A figura abaixo ilustra uma curva de contorno para variáveis aleatórias x_1 e x_2 correlacionadas positivamente ($\rho = 0{,}5$).



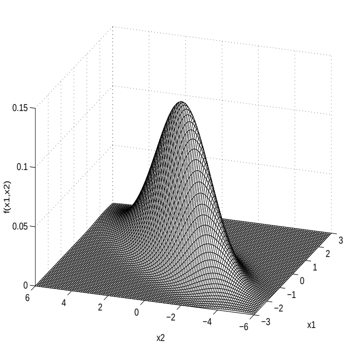
Contornos da FDP Gaussiana Bivariada ($\rho \neq 0$) (cont.-3)

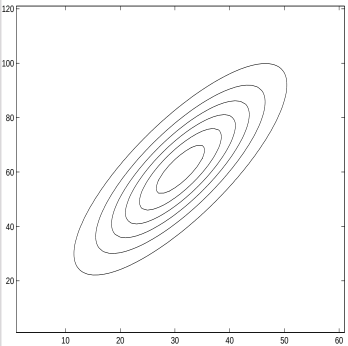
- A figura abaixo ilustra uma curva de contorno para variáveis aleatórias x_1 e x_2 correlacionadas negativamente ($\rho =$ -0,5).



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

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





Geração de Vetores Aleatórios Gaussianos Correlacionados

Passo 1 - Gerar vetores aleatórios gaussianos não-correlacionados:

```
» X=randn(p,N);
```

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

```
» Cd=[1 0.79;0.79 4];
```

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

```
» A=chol(Cd)';
```

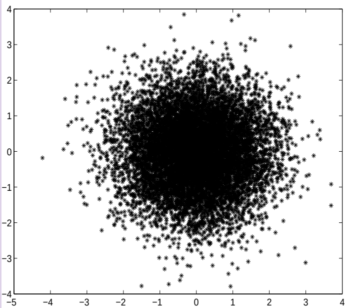
Passo 4 - Aplicar transformação linear aos dados originais:

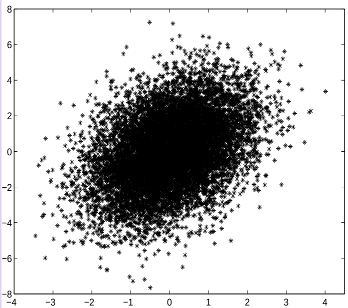
```
» Y=A*X;
```

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

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

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





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

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

```
» Cd=cov(Y)
   Cd = 0.9983     0.7836
        0.7836     3.9609
```

- 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 VII

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.

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.

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.

Conceituação

- A **especificidade** é a probabilidade que o exame dê negativo, dado que o paciente não está doente.
- Ou seja, é a razão entre pacientes cujos exames dão negativo (VN) pelo número total de pacientes que são verdadeiramente negativos (VN+FP).
- Quanto maior a especificidade, menor o número de pacientes saudáveis que são classificados com doentes.

Especificidade (*True Negative Ratio*)

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{91}$$

Conceituação

- A **sensibilidade** é a probabilidade que o exame dê positivo, dado que o paciente está doente.
- Ou seja, é a razão entre pacientes cujos exames dão positivo (VP) pelo número total de pacientes que são verdadeiramente positivos (VP+FN).
- Quanto maior a sensibilidade, menor é o número de casos de pacientes doentes que não são detectados.

Sensibilidade (*True Positive Ratio*)

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{92}$$

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.

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, Falsos Positivos, Falsos Negativos e Verdadeiros Positivos.

Matriz de Confusão para Classificação Binária

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