SCC5895 – Análise de Agrupamento de Dados

Representação de Dados e Medidas de Proximidade

Prof. Eduardo R. Hruschka

PPG-CCMC / ICMC / USP

Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - Elaborados por Eduardo R. Hruschka e Ricardo J. G. B. Campello;
  - Gentilmente cedidos por André C. P. L. F. de Carvalho
  - de (Tan et al., 2006)
  - de E. Keogh (SBBD 2003)
  - de G. Piatetsky-Shapiro (KDNuggets)

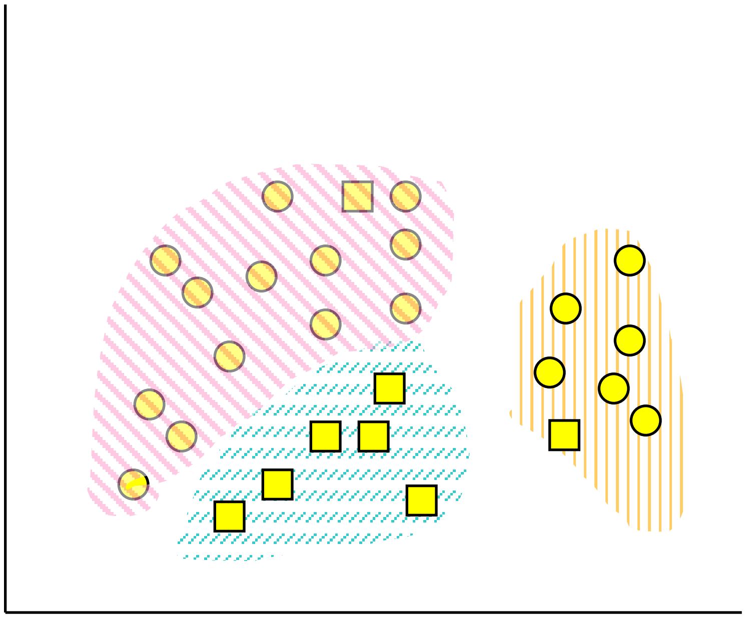
Agenda

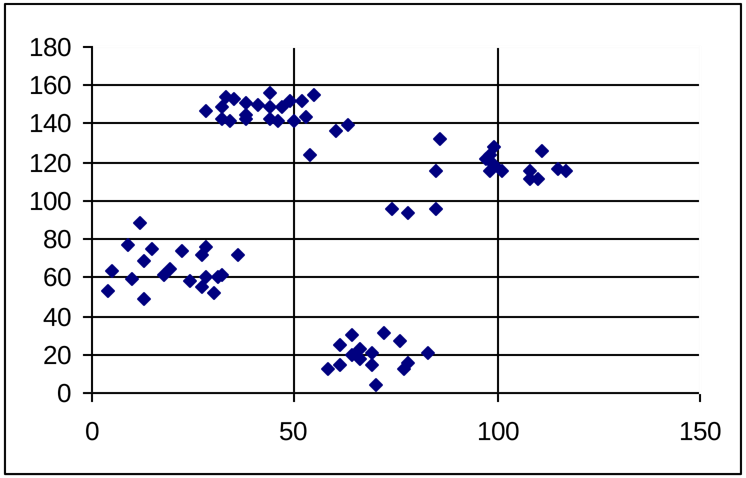
- Motivação
- Tipos e Escalas de Dados
- Normalizações
- Medidas de Proximidade
  - Similaridade
  - Dissimilaridade
- Noções de Significância Estatística

Agrupamento de Dados (*Clustering*)

- Aprendizado não supervisionado

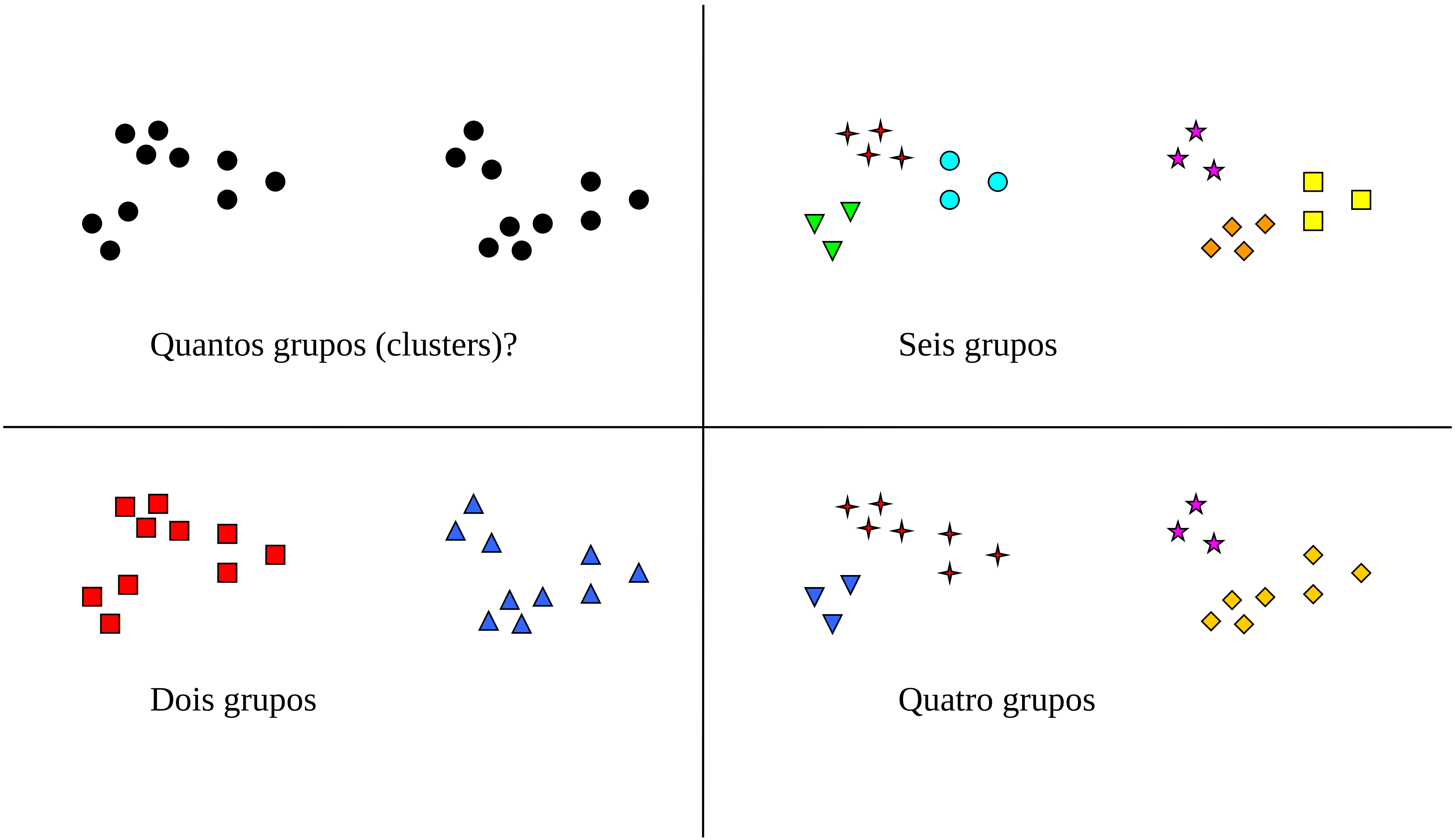
- Encontrar grupos "naturais" de objetos para um conjunto de dados não rotulados





Slide baseado no curso de Gregory Piatetsky-Shapiro, disponível em http://www.kdnuggets.com

Noção de grupo pode ser ambígua



Visualizando Clusters

- Sistema visual humano é muito poderoso para reconhecer padrões
- Entretanto...
  - "*Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent*" (Carl Sagan)
- Everitt et al., Cluster Analysis, Chapter 2 (Visualizing Clusters), Fourth Edition, Arnold, 2001

Definindo o que é um **Cluster**

- Conceitualmente, definições são subjetivas:
  - Homogeneidade (coesão interna)...
  - Heterogeneidade (separação)...
  - Densidade (concentração)...
- É preciso formalizar matematicamente
- Existem diversas medidas
  - Cada uma induz (impõe) uma estrutura aos dados...
  - Em geral, baseadas em algum tipo de **(dis)similaridade**

Medidas de (Dis)Similaridade

- Existem diversas medidas de dissimilaridade e similaridade, p/ diferentes contextos de aplicação
- Cada uma assume que os objetos são descritos por atributos de uma determinada natureza
  - qualitativos, quantitativos, ...
- Para discuti-las precisamos antes falar um pouco sobre tipos e escalas de dados...

- Reconhecer o **tipo** e a **escala** dos dados nos ajuda a escolher o algoritmo de agrupamento:

- **Tipo de dados**: no presente contexto, refere-se ao grau de quantização dos dados
  - Atributo **Binário**:
    - 2 valores
  - Atributo **Discreto**:
    - valores enumeráveis
    - **binário** é caso particular
  - Atributo **Contínuo**:
    - valores numéricos reais

- Podemos tratar qualquer atributo como assumindo valores na forma de números, em algum tipo de **escala**
- **Escala de dados**: indica a significância relativa dos números (nominal, ordinal, intervalar e taxa)
  - **Escala Qualitativa**:
    - **Nominal**: números usados como *nomes;* p. ex.
      - {M, F} = {0, 1}
      - {Solteiro, Casado, Separado, Viúvo} = {0, 1, 2, 3}
    - **Ordinal**: números possuem apenas informação sobre a ordem relativa; p. ex.
      - {ruim, médio, bom} = {1, 2, 3} = {10, 20, 30} = {1, 20, 300}
      - {frio, morno, quente} = {1, 2, 3}

Faz sentido realizar cálculos diretamente com escalas qualitativas como aquelas acima?

- **Escala Quantitativa**:
  - **Intervalar**:
    - Interpretação dos números depende de uma unidade de medida, cujo zero é arbitrário
    - Exemplos:
      - Temperatura 26ºC = 78F não é 2 vezes mais quente do que 13ºC (55F) e do que 39F (4ºC)
      - 400D.C. não é 2 vezes mais tempo histórico de uma sociedade que 200D.C.
  - **Razão**:
    - Interpretação não depende de qualquer unidade
    - Exemplos:
      - 2x Temperatura em Kelvin = 2 vezes mais quente
      - 2x Salário = dobro do poder de compra, não interessa moeda

Medidas de (Dis)similaridade

"*A escolha da medida de dis(similaridade) é importante para aplicações, e a melhor escolha é freqüentemente obtida via uma combinação de experiência, habilidade, conhecimento e sorte*..."

Gan, G., Ma, C., Wu, J., **Data Clustering: Theory, Algorithms, and Applications**, SIAM Series on Statistics and Applied Probability, 2007

Notação

- **Matriz de Dados X:**
  - N linhas (objetos) e n colunas (atributos):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

- Cada **objeto** (linha da matriz) é denotado por um vetor $\mathbf{x}_i$
  - Exemplo:

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} & \cdots & x_{1n} \end{bmatrix}$$

Notação

- **Matriz de Dados X:**
  - N linhas (objetos) e n colunas (atributos):

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

- Cada **atributo** (coluna) da matriz será denotada por um vetor $\mathbf{a}_i$
  - Exemplo:

$$\mathbf{a}_1 = \begin{bmatrix} x_{11} & \cdots & x_{N1} \end{bmatrix}^T$$

Notação

- **Matriz de Proximidade (**Dissimilaridade ou Similaridade**):**
  - N linhas e N colunas:

$$\mathbf{D} = \begin{bmatrix} d(\mathbf{x}_1, \mathbf{x}_1) & d(\mathbf{x}_1, \mathbf{x}_2) & \cdots & d(\mathbf{x}_1, \mathbf{x}_N) \\ d(\mathbf{x}_2, \mathbf{x}_1) & d(\mathbf{x}_2, \mathbf{x}_2) & \cdots & d(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & & \ddots & \vdots \\ d(\mathbf{x}_N, \mathbf{x}_1) & d(\mathbf{x}_N, \mathbf{x}_2) & \cdots & d(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

- Simétrica se proximidade d apresentar propriedade de simetria

Similaridade e Dissimilaridade

- **Similaridade**
  - Mede o quanto dois objetos são parecidos
    - quanto mais parecidas, maior o valor
  - Geralmente valor $\in [0, 1]$
- **Dissimilaridade**
  - Mede o quanto dois objetos são diferentes
    - quanto mais diferentes, maior o valor
  - Geralmente valor $\in [0, d_{max}]$ ou $[0, \infty]$

Similaridade x Dissimilaridade

- Saber converter dissimilaridades (**d**) em similaridades (**s**) e vice-versa é muitas vezes útil e nos permite tratar com apenas uma das formas
  - Se ambas forem definidas em [0,1], a conversão é direta:
    - $\mathbf{s} = 1 - \mathbf{d}$ ou $\mathbf{d} = 1 - \mathbf{s}$ (linear, não distorce os valores)
  - Caso contrário, algumas alternativas são:
    - se limitantes para **s** (s_{min} e s_{max}) ou **d** (d_{min} e d_{max}) forem conhecidos, podemos re-escalar em [0,1] e usar $\mathbf{s} = 1 - \mathbf{d}$
    - se $\mathbf{d} \in [0,\infty]$, não há como evitar uma transformação não linear...
      - por exemplo, $\mathbf{s} = 1/(1 + \alpha\mathbf{d})$ ou $\mathbf{s} = e^{-\alpha\mathbf{d}}$ ($\alpha \rightarrow$ constante positiva)
      - melhor forma depende do problema...

Dissimilaridade e Distância

- Em agrupamento de dados, dissimilaridades são em geral calculadas utilizando medidas de **distância**
- Uma medida de distância é uma medida de dissimilaridade que apresenta um conjunto de propriedades

Propriedades de Distâncias

- Seja **d**(**p**, **q**) a distância entre dois objetos **p** e **q**
- Então valem a seguintes propriedades:
  - **Positividade e Reflexividade**:
    - $d(p, q) \geq 0 \quad \forall$ p e q
    - $d(p, q) = 0$ se e somente se $p = q$
  - **Simetria**:
    - $d(p, q) = d(q, p)\ \forall$ p e q
- Além disso, **d** é dita uma **métrica** se também vale:
  - $d(p, q) \leq d(p, r) + d(r, q)\ \forall$ p, q e r (**Desigualdade Triangular**)

Desigualdade Triangular:

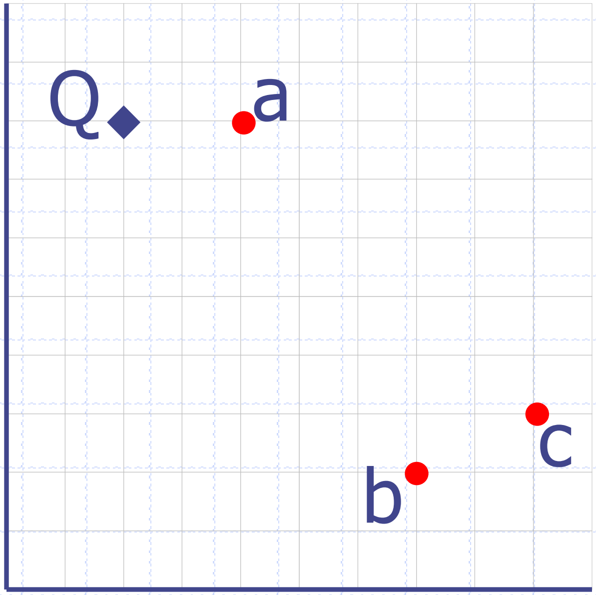
- Encontrar o objeto mais próximo de **Q** em uma base de dados formada por três objetos (**a**,**b**,**c**)

- Assumamos que já se disponha de algumas distâncias entre pares de objetos: $d(\mathbf{a},\mathbf{b})$, $d(\mathbf{a},\mathbf{c})$, $d(\mathbf{b},\mathbf{c})$

- Calculamos $d(\mathbf{Q},\mathbf{a}) = 2$ e $d(\mathbf{Q},\mathbf{b}) = 7.81$

- Não é necessário calcular explicitamente $d(\mathbf{Q},\mathbf{c})$:

$$d(\mathbf{Q},\mathbf{b}) \leq d(\mathbf{Q},\mathbf{c}) + d(\mathbf{c},\mathbf{b})$$
$$d(\mathbf{Q},\mathbf{b}) - d(\mathbf{c},\mathbf{b}) \leq d(\mathbf{Q},\mathbf{c})$$
$$7.81 - 2.30 \leq d(\mathbf{Q},\mathbf{c})$$
$$5.51 \leq d(\mathbf{Q},\mathbf{c})$$



	a	b	c
a		6.70	7.07
b			2.30
c			

➢ Já se pode afirmar que **a** está mais próximo de **Q** do que qualquer outro objeto da base de dados

➢ Veremos mais adiante no curso um possível uso desta propriedade em agrupamento de dados

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Propriedades de Similaridade

- As seguintes propriedades são desejáveis e em geral são válidas para similaridades:
  - Seja **s**(**p**, **q**) a similaridade entre **p** e **q**
    - $s(p, q) = 1$ apenas se $p = q$ (similaridade máxima)
    - $s(p, q) = s(q, p) \ \forall$ p e q (simetria)

Medidas de (Dis)similaridade:

a) Atributos **contínuos**

b) Atributos **discretos**

c) Atributos **mistos**

- Nos concentraremos em estudar medidas amplamente utilizadas na prática

- Há uma vasta literatura sobre este assunto

    - ver bibliografia da disciplina

a) Atributos Contínuos

a.1) **Distância Euclidiana**:

$$d^{E}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left\| \mathbf{x}_i - \mathbf{x}_j \right\| = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

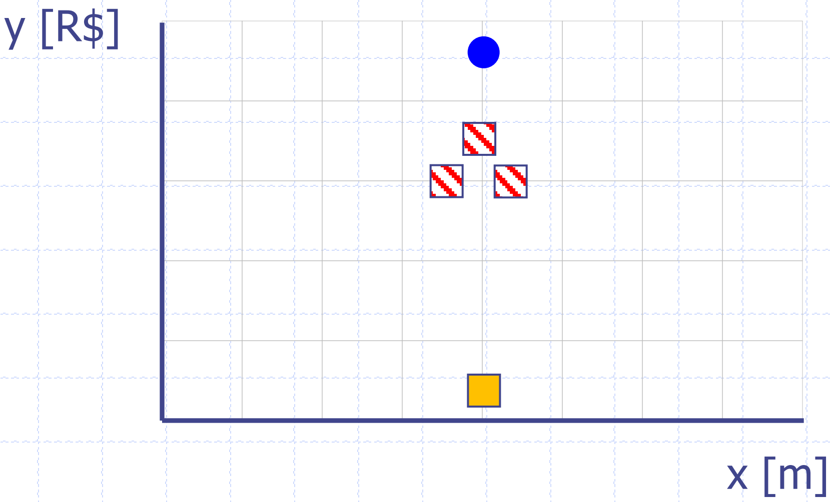
- Métrica
- Tende a induzir *clusters* hiper-esféricos
- *Clusters* invariantes com rel. a translação e rotação no espaço dos atributos (Duda et al., Pattern Classification, 2001)
- Implementações computacionais eficientes usam $(d^E)^2$
- Atributos com maiores valores e variâncias tendem a *dominar* os demais...

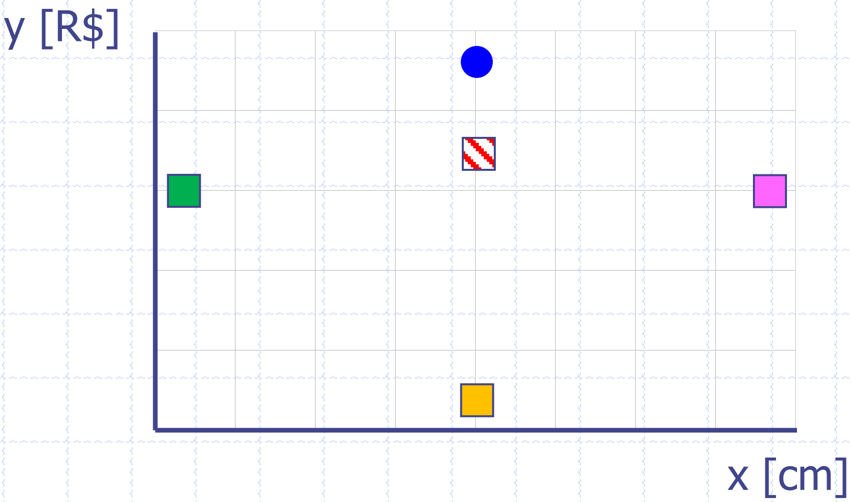
Exemplo 1:

	a_1	a_2	a_3	a_4
x_1	1	2	5	803
x_2	1	1	5	712
x_3	1	1	5	792
x_4	0	2	6	608
x_5	0	1	5	677
x_6	1	1	5	927
x_7	1	1	5	412
x_8	1	1	6	368
x_9	1	1	6	167
x_{10}	0	2	5	847
Média	**0,70**	**1,30**	**5,30**	**631,30**
Variância	**0,23**	**0,23**	**0,23**	**59045,34**

$$d^E(\mathbf{x}_1,\mathbf{x}_2) = ?$$

Exemplo 2:





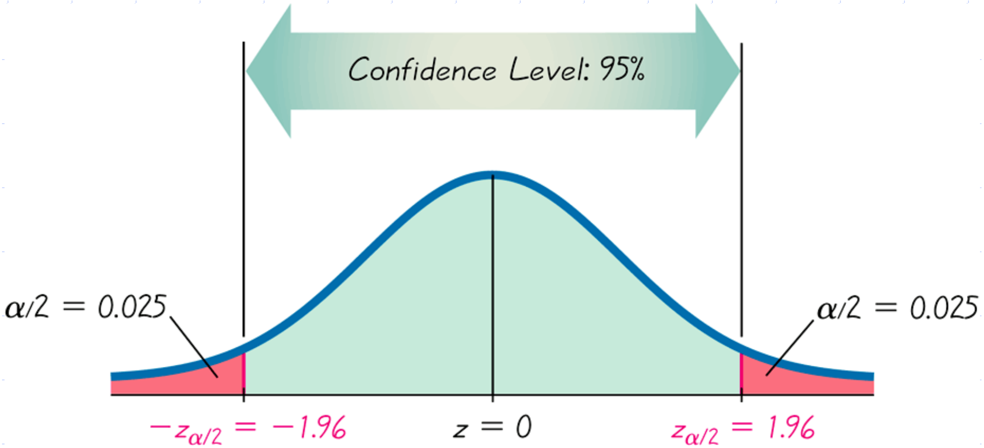
- Pode-se lidar com tais problemas por meio do que usualmente se denomina **normalização**

- Estudaremos as formas de normalização mais comuns...

Normalização

- Re-escala Linear [0,1]: $$l_{ij} = \frac{x_{ij} - \min(\mathbf{a}_j)}{\max(\mathbf{a}_j) - \min(\mathbf{a}_j)}$$

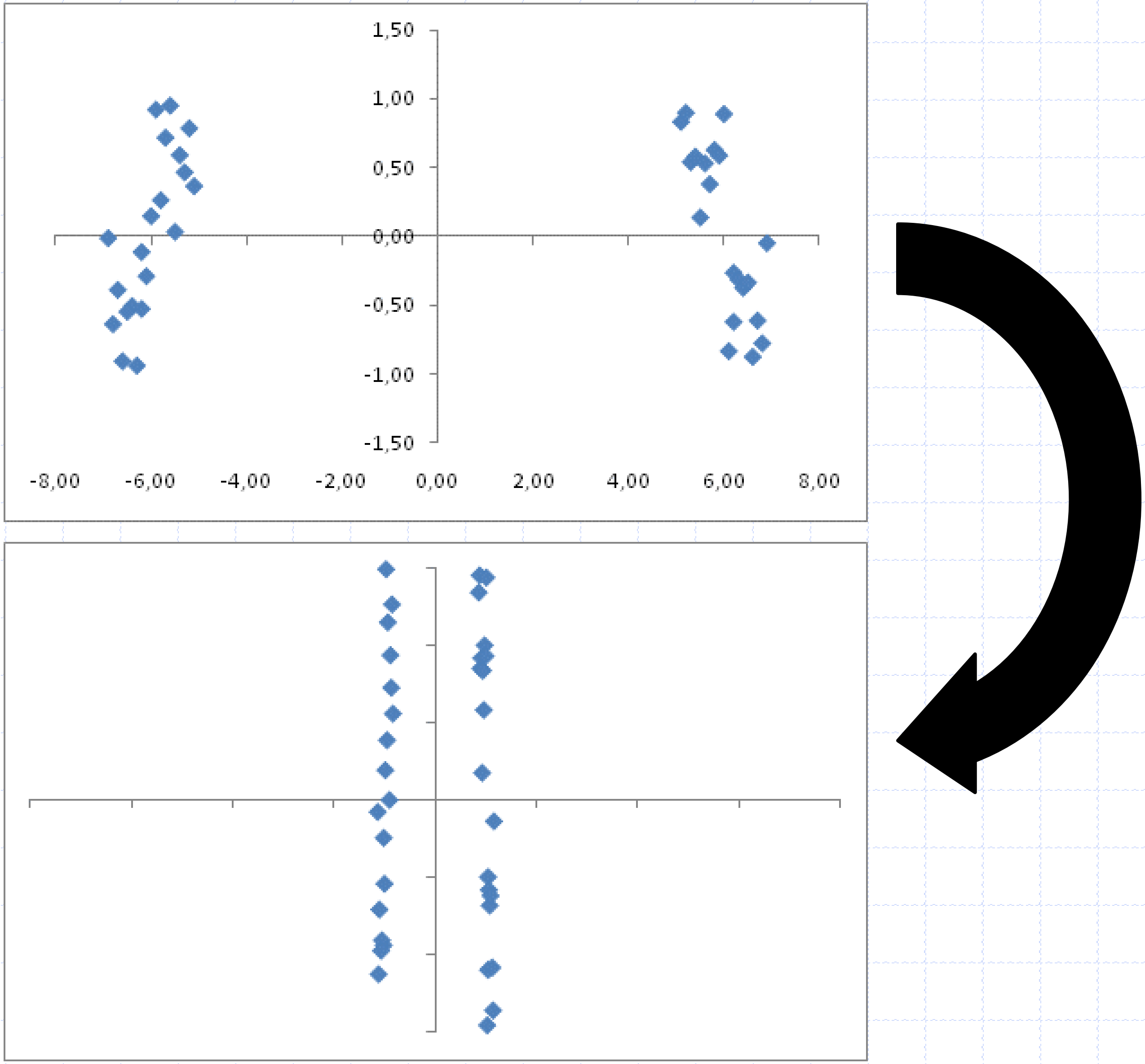
- Padronização Escore z: $$z_{ij} = \frac{x_{ij} - \mu_{\mathbf{a}_j}}{\sigma_{\mathbf{a}_j}}$$



N(0,1) se atributo possui dist. Normal

Triola, Notas de Aula, Copyright © 2004 Pearson Education

Normalização não é necessariamente sempre apropriada ...



escore z
(efeito semelhante para linear [0,1])

❑ **Em Resumo**:

- ➢ Atributos com escala mais ampla / maior variabilidade tendem a ter maior peso nos cálculos de distâncias

  - ➢ Isso representa uma espécie de pré-**ponderação** implícita dos dados

- ➢ Normalização busca eliminar esse efeito, assumindo ser artificial

  - ➢ p. ex., simples consequência do uso de unidades de medida específicas

  - ➢ porém, também impõe uma (contra) ponderação aos dados originais...

  - ➢ pode introduzir distorções se (ao menos parte das) diferentes variabilidades originais refletiam corretamente a natureza do problema

❑ Por essas e (tantas) outras, agrupamento de dados é considerada uma das área de DM mais desafiadoras.

Recomendações ?

- Difícil fornecer sugestões independentes de domínio

- Everitt et al. (2001) sugerem que *escores z* e normalizações lineares [0,1] não são eficazes em geral

- Lembremos que ADs envolve, em essência, **análise exploratória de dados**

  - Quais são os pesos mais apropriados ?

    - para pesos 0 e 1 ⇒ quais são os melhores atributos ?

    - questão remete a agrupamento em sub-espaços...

a.2) Distância de **Minkowski**:

$$d^{p}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_p = \left(\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^p \right)^{1/p}$$

- Para p = 2: Distância Euclidiana

- Para p = 1: Distância de **Manhattan** (*city block*, *taxicab*)

  - recai na distância de **Hamming** para atributos binários

- Para $p \to \infty$: Dist. **Suprema** $d^{\infty}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_\infty = \max_{1 \le k \le n} \left| x_{ik} - x_{jk} \right|$

- Em 2-dimensões, quais seriam as superfícies formadas pelos pontos equidistantes de um ponto de origem ?

❑ a.2.1) Distância de **Minkowski Normalizada**:

$$d^{p}_{(\mathbf{x}_i,\mathbf{x}_j)} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_p = \left(\frac{\sum_{k=1}^{n} \delta_{ijk} \left| x_{ik} - x_{jk} \right|^{p}}{\sum_{k=1}^{n} \delta_{ijk}} \right)^{1/p}$$

δ_{ijk} = 0 se x_{ik} **ou** x_{jk} forem ausentes

δ_{ijk} = 1 se x_{ik} **e** x_{jk} forem conhecidos

- Permite cálculos na presença de valores faltantes
- Alternativa à imputação
- Qual a melhor abordagem?
  - **análise exploratória de dados**...

Distância com Valores Ausentes

Exemplo (Distância Euclidiana Normalizada entre $\mathbf{x}_1$ e $\mathbf{x}_3$):

Obj. / Atrib.	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	$\mathbf{a_4}$
$\mathbf{x_1}$	2	-1	**?**	0
$\mathbf{x_2}$	7	0	-4	8
$\mathbf{x_3}$	**?**	3	5	2
$\mathbf{x_4}$	**?**	10	**?**	5

- **Exercício:** calcule todas as demais distâncias !

a.3) Distância de **Mahalanobis**:

$$\left(d^{m}_{(\mathbf{x}_i,\mathbf{v}_j)}\right)^2 = \left(\mathbf{x}_i - \mathbf{v}_j\right)^T \mathbf{\Sigma}_j^{-1}\left(\mathbf{x}_i - \mathbf{v}_j\right)$$

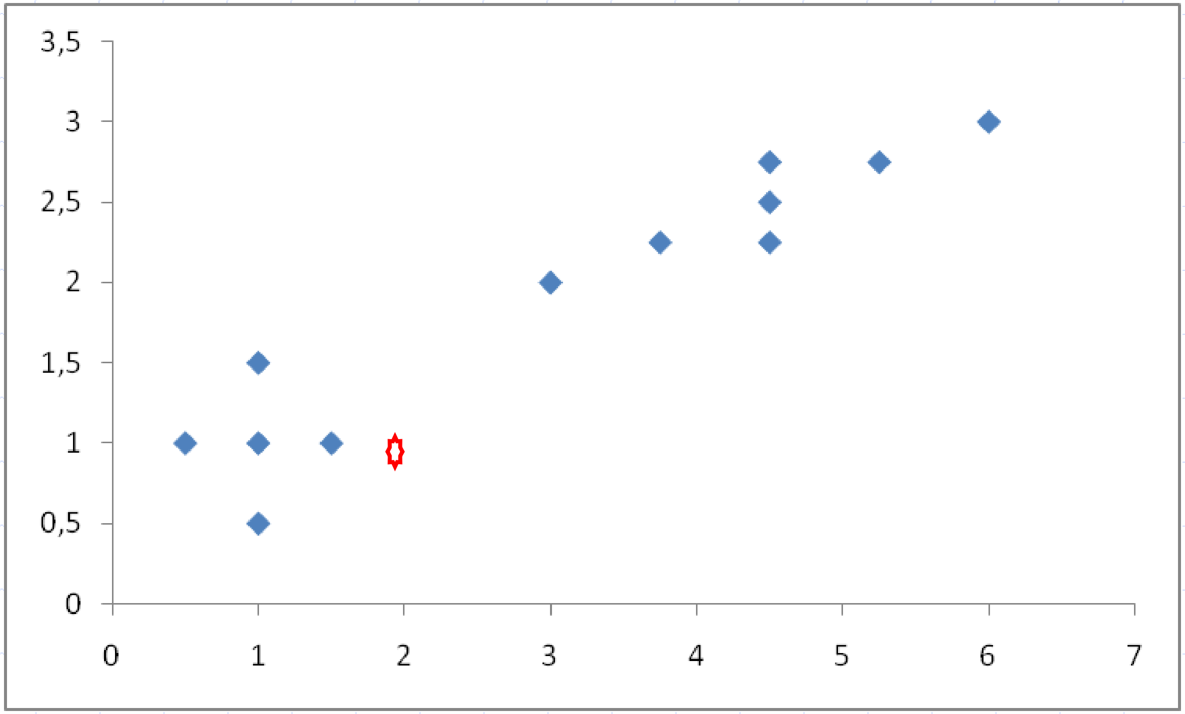
Σ_j = matriz de covariâncias do j-ésimo grupo de dados, com objetos $\mathbf{x}_l$ (l = 1, ..., N_j) e centro $\mathbf{v}_j$:

$$\mathbf{\Sigma}_j = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} = \frac{1}{N_j}\sum_{l=1}^{N_j}\left(\mathbf{x}_l - \mathbf{v}_j\right)\left(\mathbf{x}_l - \mathbf{v}_j\right)^T$$

simétrica

$$\mathbf{v}_j = \frac{1}{N_j}\sum_{l=1}^{N_j}\mathbf{x}_l$$

Exemplo pedagógico:



Considere o pto. (2,1) e suas distâncias aos centros dos grupos:

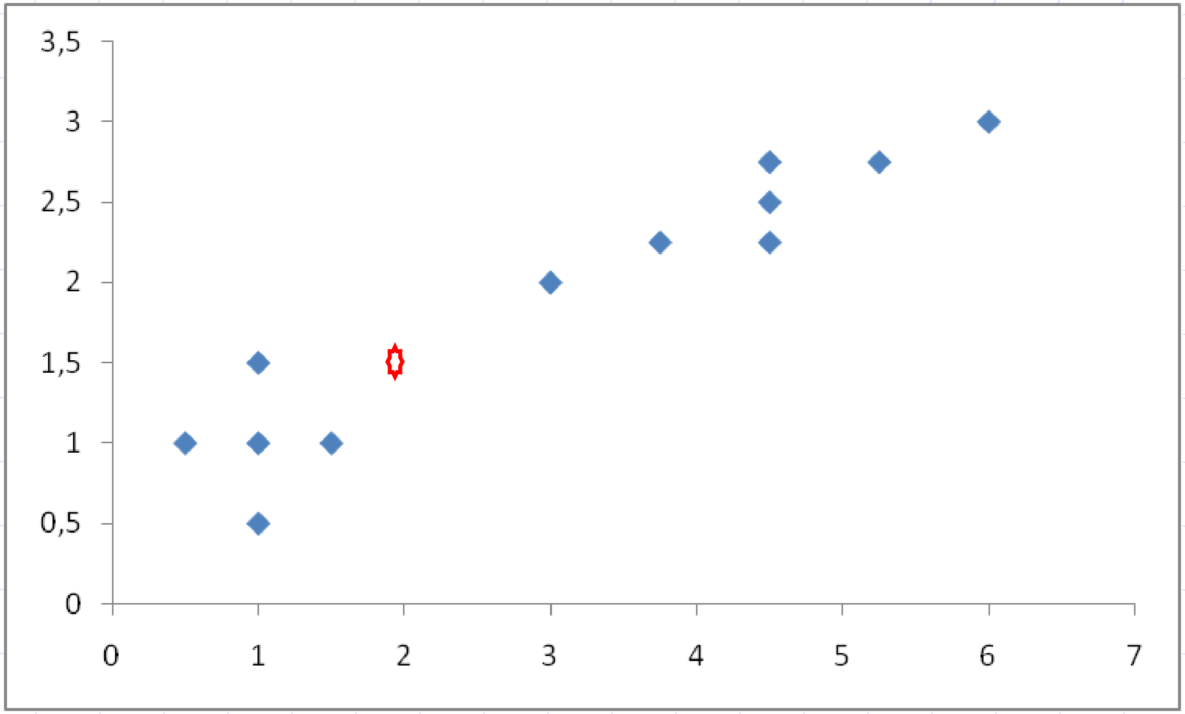
$d^m(2,1)_c=10$
$d^m(2,1)_e=29$

Consideremos agora que esse ponto se mova para cima...

$$\Sigma_c = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \qquad \Sigma_e = \begin{bmatrix} 0.80 & 0.27 \\ 0.27 & 0.11 \end{bmatrix}$$

$$\Sigma_c^{-1} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \qquad \Sigma_e^{-1} = \begin{bmatrix} 7 & -17 \\ -17 & 50 \end{bmatrix}$$

Exemplo pedagógico:

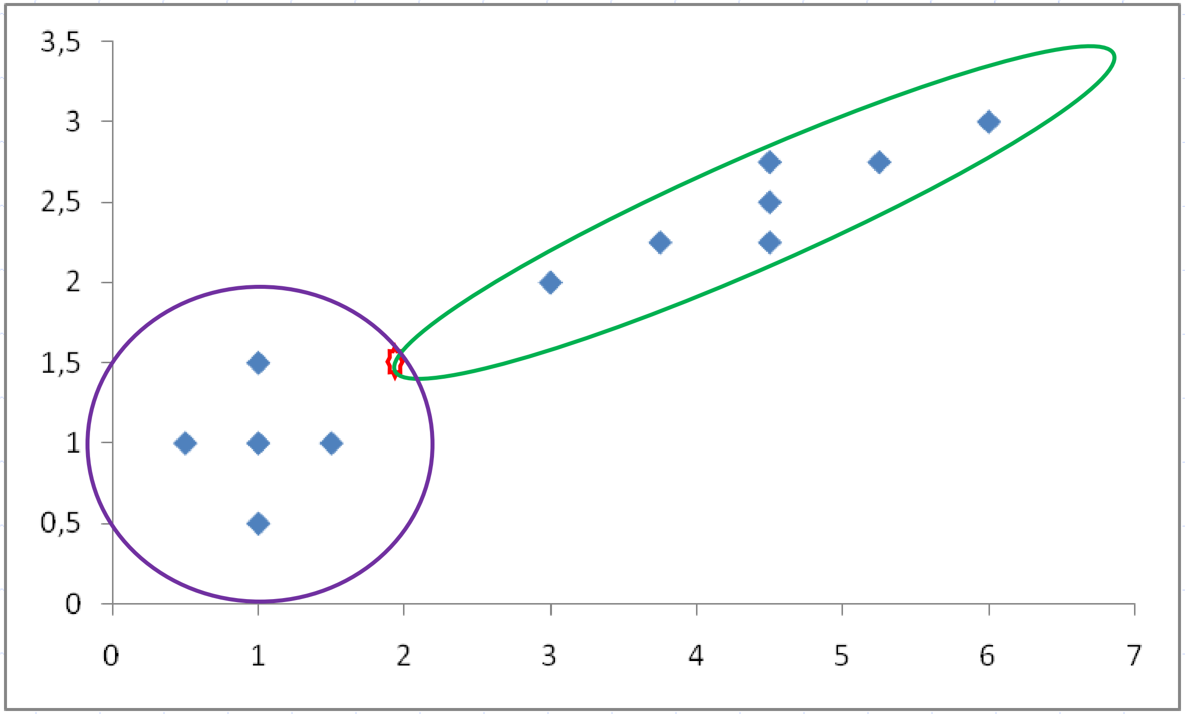


Qual é o *cluster* mais próximo?

$$\Sigma_c = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \qquad \Sigma_e = \begin{bmatrix} 0.80 & 0.27 \\ 0.27 & 0.11 \end{bmatrix}$$

$$\Sigma_c^{-1} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \qquad \Sigma_e^{-1} = \begin{bmatrix} 7 & -17 \\ -17 & 50 \end{bmatrix}$$

Exemplo pedagógico...



$d^m(2.0,1.5)_c=12.5$

$d^m(2.0,1.5)_e= 8.80$

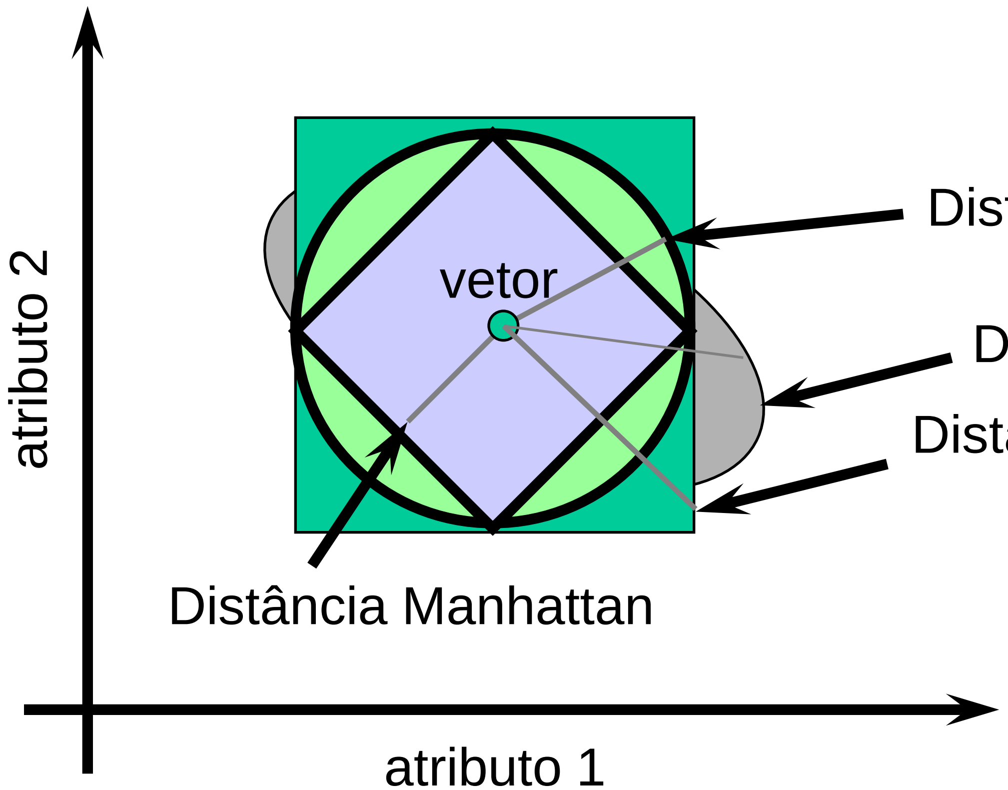
- Voltaremos a esse assunto quando estudarmos GK e EM
- Problemas apresentados pela distância de Mahalanobis?
  - Cálculo da inversa da matriz de covariâncias...

Nota:

- A distância de Mahalanobis é uma distância de um objeto a um grupo de pontos (em particular, ao seu centro)
- Se calculada entre dois objetos, assume implicitamente que um deles é o centro de um grupo com covariância Σ_j
- Generalizações, por exemplo para distância entre 2 grupos, são discutidas em (Everitt et al., 2001)

Visão Geométrica

- Onde se situam os pontos eqüidistantes de um vetor



a.4) Correlação Linear de **Pearson**

$$r(\mathbf{x}_i, \mathbf{x}_j) = \frac{\frac{1}{n}\sum_{k=1}^{n}\left(x_{ik} - \mu_{\mathbf{x}_i}\right)\left(x_{jk} - \mu_{\mathbf{x}_j}\right)}{\frac{1}{n}\sqrt{\sum_{k=1}^{n}\left(x_{ik} - \mu_{\mathbf{x}_i}\right)^2 \sum_{i=1}^{n}\left(x_{jk} - \mu_{\mathbf{x}_j}\right)^2}} = \frac{\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i} \cdot \sigma_{\mathbf{x}_j}}$$

- medida de similaridade
- interpretação intuitiva ?

Pearson, K., Mathematical contributions to the theory of evolution, III Regression, Heredity and Panmixia, *Philos. Trans. Royal Soc. London Ser. A*, v. 187, pp. 253-318, 1896.

Correlação

- Mede interdependência entre vetores numéricos
  - Por exemplo, interdependência linear
- Pode ser portanto usada para medir similaridade
  - entre 2 objetos descritos por atributos numéricos
  - entre 2 atributos numéricos (seleção de atributos)
- Correlação de **Pearson** mede a compatibilidade linear entre as tendências dos vetores
  - despreza média e variabilidade
  - muito útil em bioinformática

Correlação de Pearson

- Cálculo do coeficiente de Pearson:
  - Padronizar vetores via score-z
  - Calcular produto interno

$$p'_k = (p_k - \mu_{\mathbf{p}}) / \sigma_{\mathbf{p}}$$

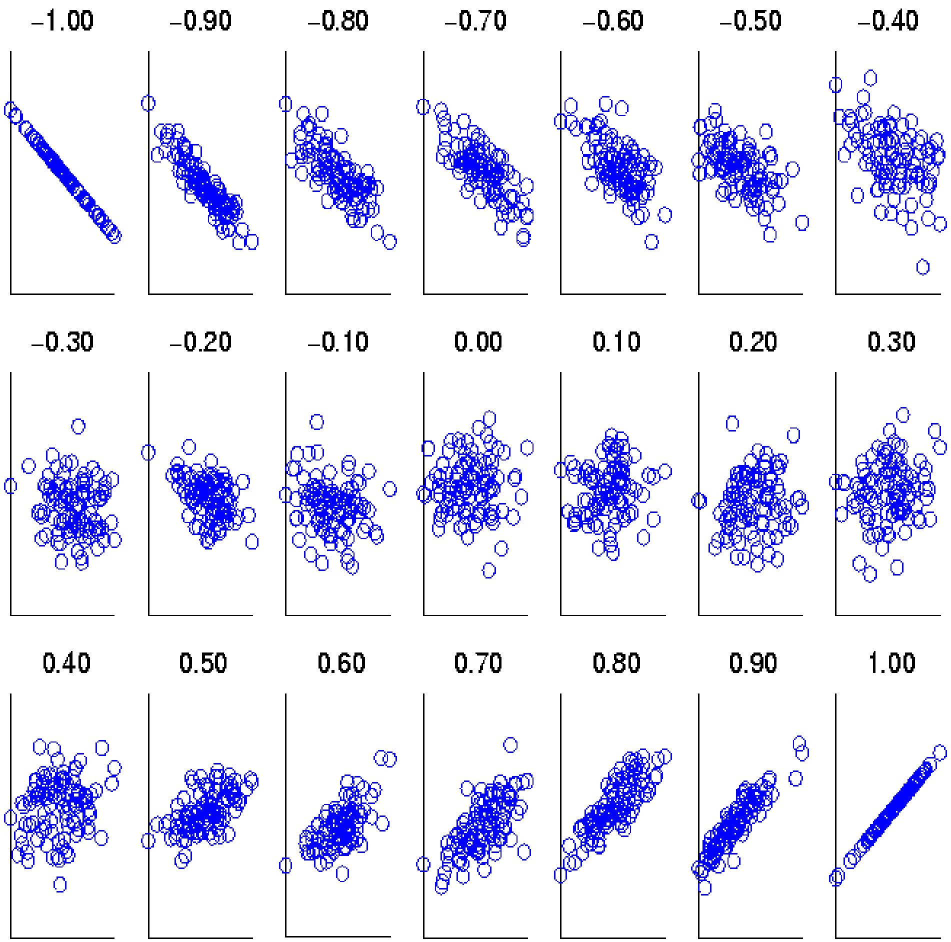
$$q'_k = (q_k - \mu_{\mathbf{q}}) / \sigma_{\mathbf{q}}$$

$$\text{correlação}(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}' \bullet \mathbf{q}'}{n}$$

Correlação

- Valor no intervalo [-1, +1]
  - Correlação ($\mathbf{p}$, $\mathbf{q}$) = +1
    - Objetos p e q têm um relacionamento linear positivo perfeito
  - Correlação ($\mathbf{p}$, $\mathbf{q}$) = −1
    - Objetos p e q têm um relacionamento linear negativo perfeito
  - Correlação ($\mathbf{p}$, $\mathbf{q}$) = 0
    - Não existe relacionamento linear entre os objetos p e q
  - Relacionamento linear: $\mathbf{p}_k = a\mathbf{q}_k + b$

Avaliação Visual de Correlação

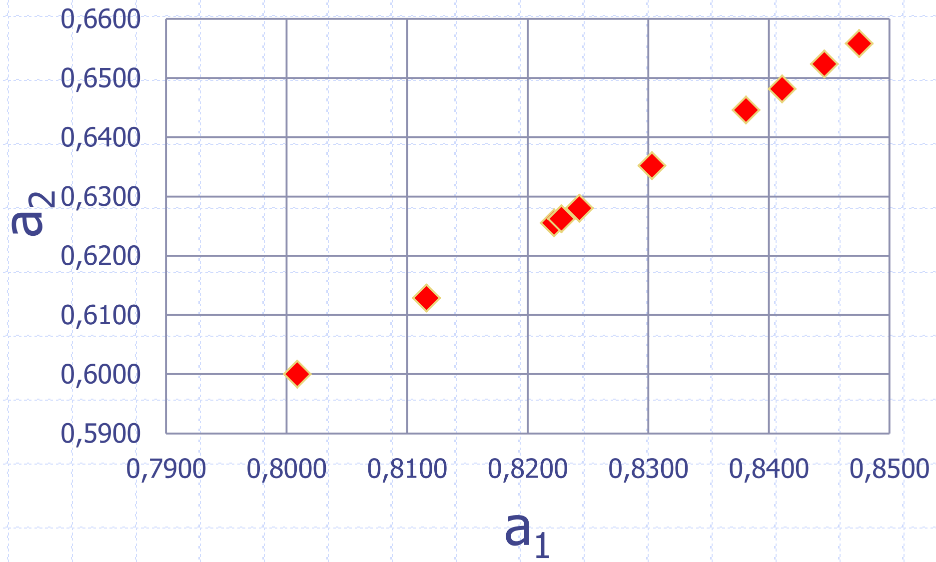


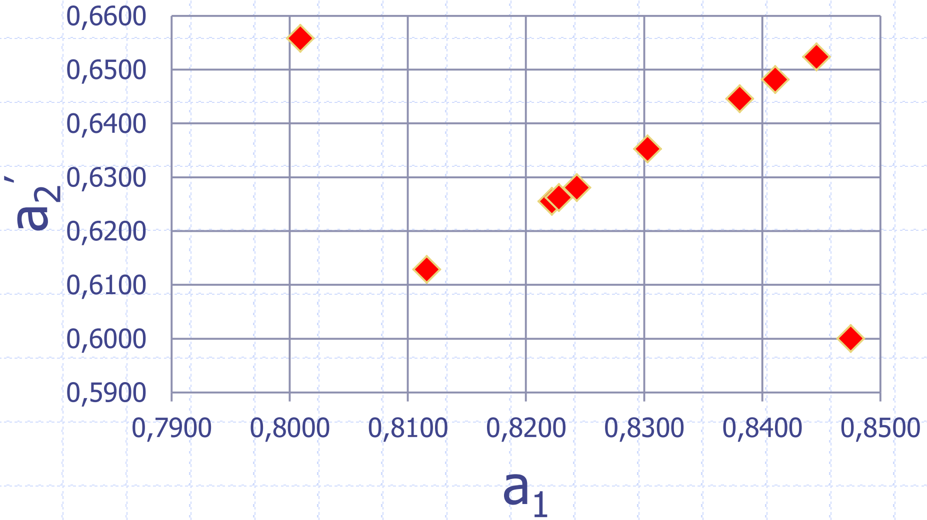
Scatter plots para dois objetos considerando diferentes pares de atributos.

Pearson possui seus problemas; por exemplo...

a_1	a_2	a_2'
0.8009	0.6000	0.6558
0.8116	0.6128	0.6128
0.8222	0.6255	0.6255
0.8228	0.6262	0.6262
0.8243	0.6280	0.6280
0.8303	0.6352	0.6352
0.8381	0.6446	0.6446
0.8411	0.6482	0.6482
0.8446	0.6523	0.6523
0.8475	0.6558	0.6000

	(a_1, a_2)	(a_1, a_2')
Pearson	1.000	-0.080





... mas existem outras medidas de correlação...

Campello, R. J. G. B., Hruschka, E. R., *On Comparing Two Sequences of Numbers and Its Applications to Clustering Analysis*, **Information Sciences**, 2009

Exercício

- Calcular correlação de Pearson entre os seguintes objetos **p** e **q**

$$\mathbf{p} = [1\ \ -3\ \ 0\ \ 4\ \ 1\ \ 0\ \ 3]$$
$$\mathbf{q} = [0\ \ 1\ \ 4\ \ -2\ \ 3\ \ -1\ \ 4]$$

a.5) **Cosseno**

- Correlação de Pearson tende a enxergar os vetores como sequências de valores e capturar as semelhanças de forma / tendência dessas sequências

    - Não trata os valores como assimétricos

    - Valores nulos interferem no resultado

- Similaridade baseada no **Cosseno**, embora seja matematicamente similar, possui características diferentes:

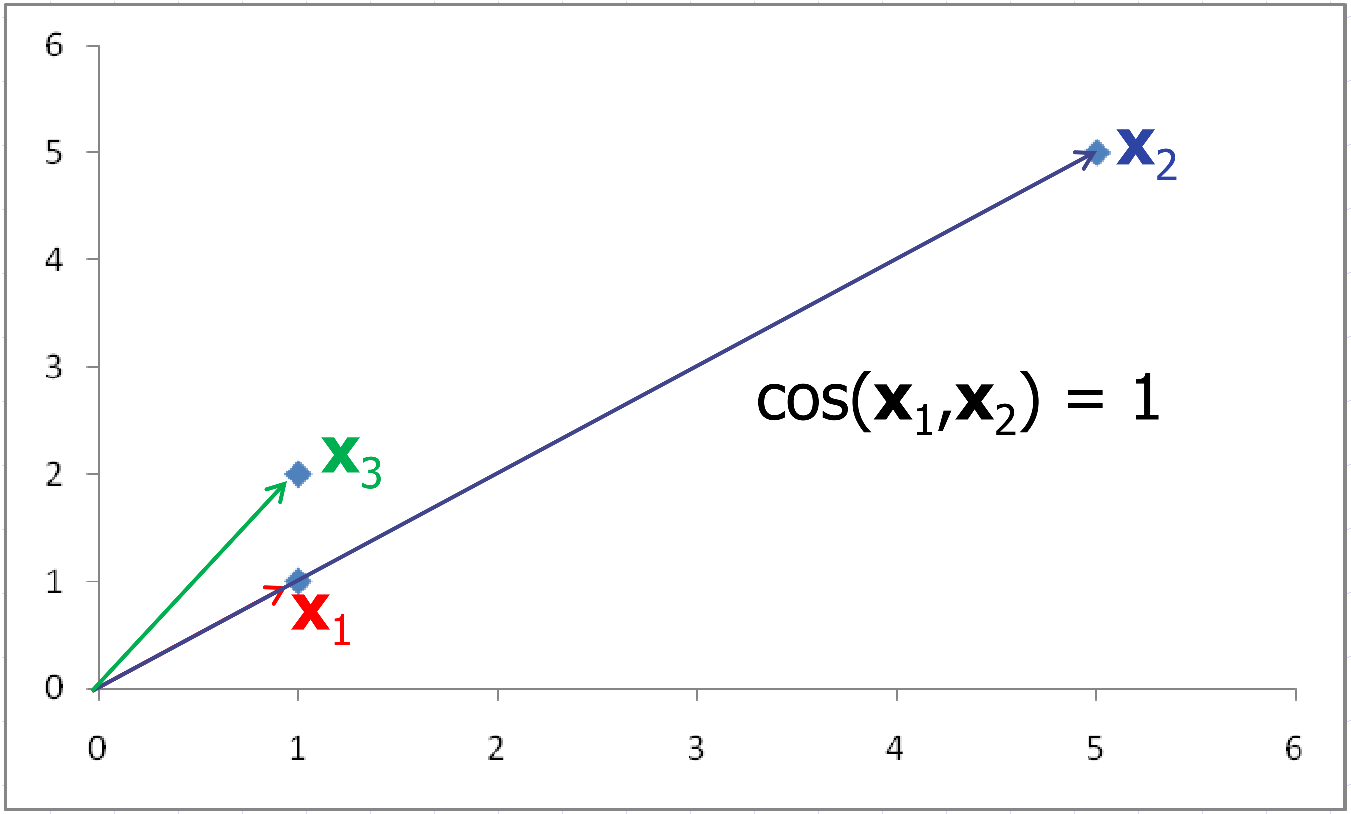
$$\cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}$$

Similaridade Cosseno

- Apropriada para **atributos assimétricos***
  - Muito utilizada em mineração de textos
    - grande número de atributos, poucos não nulos (dados esparsos)
- Sejam $\mathbf{d}_1$ e $\mathbf{d}_2$ vetores de valores assimétricos
  - $\boxed{\cos(\mathbf{d}_1, \mathbf{d}_2) = (\mathbf{d}_1 \bullet \mathbf{d}_2) / ||\mathbf{d}_1|| \; ||\mathbf{d}_2||}$
    - $\bullet$: produto interno entre vetores
    - $||\ \mathbf{d}\ ||$: é o tamanho (norma) do vetor **d**
  - Mede o cosseno do ângulo entre os respectivos versores

* Definidos em breve.

Exemplo (Gráfico):



$\cos(\mathbf{x}_1,\mathbf{x}_3) = \cos(\mathbf{x}_2,\mathbf{x}_3) = 0.95$

(ângulo de aproximadamente 18º)

- Para calcular distâncias (entre documentos):

$d(\mathbf{x}_i,\mathbf{x}_j) = 1 - \cos(\mathbf{x}_i,\mathbf{x}_j)$

Exemplo (Numérico)

- Sejam os vetores (instâncias) $\mathbf{d}_1$ e $\mathbf{d}_2$ abaixo
  - $\mathbf{d}_1$ = [3 2 0 5 0 0 0 2 0 0]
  - $\mathbf{d}_2$ = [1 0 0 0 0 0 0 1 0 2]

$\cos(\mathbf{d}_1,\mathbf{d}_2) = (\mathbf{d}_1 \bullet \mathbf{d}_2) / ||\mathbf{d}_1||\ ||\mathbf{d}_2||$

$\mathbf{d}_1 \bullet \mathbf{d}_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||\mathbf{d}_1|| = (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$

$||\mathbf{d}_2|| = (1^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)^{\mathbf{0.5}} = (6)^{\mathbf{0.5}} = 2.245$

$\cos(\mathbf{d}_1,\mathbf{d}_2) = .3150$

Exercício

- Calcular dissimilaridade entre **p** e **q** usando medida de similaridade cosseno:

$$\mathbf{p} = [1\ \ 0\ \ 0\ \ 4\ \ 1\ \ 0\ \ 0\ \ 3]$$
$$\mathbf{q} = [0\ \ 5\ \ 0\ \ 2\ \ 3\ \ 1\ \ 0\ \ 4]$$

b) Atributos Discretos

Motivação:

	Sexo	País	Estado Civil	Comprar
$\mathbf{x}_1$	M	França	solteiro	Sim
$\mathbf{x}_2$	M	China	separado	Sim
$\mathbf{x}_3$	F	França	solteiro	Sim
$\mathbf{x}_4$	F	Inglaterra	casado	Sim
$\mathbf{x}_5$	F	França	solteiro	Não
$\mathbf{x}_6$	M	Alemanha	viúvo	Não
$\mathbf{x}_7$	M	Brasil	casado	Não
$\mathbf{x}_8$	F	Alemanha	casado	Não
$\mathbf{x}_9$	M	Inglaterra	solteiro	Não
$\mathbf{x}_{10}$	M	Argentina	casado	Não

$d(\mathbf{x}_1,\mathbf{x}_6)=?$

$d(\mathbf{x}_1,\mathbf{x}_7)=?$

b.1) **Atributos Binários**:

- Calcular a distância entre $\mathbf{x}_1$ = [1 0 0 1 1 0 0 1 0 0] e
  $\mathbf{x}_2$ = [0 0 0 1 0 1 1 0 0 0]
- Usando uma tabela de contingências temos:

	Objeto $\mathbf{x}_j$			
Objeto $\mathbf{x}_i$		1	0	Total
	1	n_{11}	n_{10}	$n_{11}+n_{10}$
	0	n_{01}	n_{00}	$n_{01}+n_{00}$
	Total	$n_{11}+n_{01}$	$n_{10}+n_{00}$	n

$$S^{SM}_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{n_{11}+n_{00}}{n_{11}+n_{00}+n_{10}+n_{01}} = \frac{n_{11}+n_{00}}{n}$$

Coeficiente de Casamento Simples (Zubin, 1938)

$$1-S^{SM}_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{n_{10}+n_{01}}{n} = \frac{d^{\text{Hamming}}_{(\mathbf{x}_i,\mathbf{x}_j)}}{n}$$

Entretanto, podemos ter:

- **Atributos simétricos**: valores igualmente importantes
  - Exemplo típico $\rightarrow$ Sexo (M ou F)
- **Atributos assimétricos**: valores com importâncias distintas – presença de um efeito é mais importante do que sua ausência
  - Depende do contexto...
  - Exemplo: sejam 3 objetos que apresentam (1) ou não (0) dez sintomas para uma determinada doença

$\mathbf{x}_1 = [1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1]$

$\mathbf{x}_2 = [1\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 0\ 0]$

$\mathbf{x}_3 = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$

$S^{SM}(\mathbf{x}_1,\mathbf{x}_2) = 0.5;$

$S^{SM}(\mathbf{x}_1,\mathbf{x}_3) = 0.5;$

- Conclusão?

- Para atributos assimétricos, pode-se usar, por exemplo, o *Coeficiente de Jaccard* (1908):

$$S_{(\mathbf{x}_i,\mathbf{x}_j)}^{Jaccard} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

  - Focada nos *casamentos* do tipo 1-1
  - Despreza *casamentos* do tipo 0-0
- Existem outras medidas similares na literatura, mas CCS e Jaccard são as mais utilizadas
  - vide (Kaufman & Rousseeuw, 2005)

Em Resumo...

- **Coeficiente de Casamento Simples**

$$\boxed{CCS = (n_{11} + n_{00}) / (n_{01} + n_{10} + n_{11} + n_{00})}$$

= no. de coincidências / no. de atributos

- Conta igualmente 1s e 0s, portanto é adequado quando ambos os valores são de fato equivalentes
  - Atributos binários **simétricos**

Em Resumo...

- **Coeficiente Jaccard**

$$J = n_{11} / (n_{01} + n_{10} + n_{11})$$

  - Despreza as coincidências de 0s, para lidar adequadamente com atributos **assimétricos**
    - 0s indicam apenas ausência de uma característica
    - similaridade se dá pelas características presentes

Outro Exemplo

p = [1 0 0 0 0 0 0 0 0 0]
q = [0 0 0 0 0 0 1 0 0 1]

$n_{01} = 2$ (número de atributos em que **p** = 0 e **q** = 1)
$n_{10} = 1$ (número de atributos em que **p** = 1 e **q** = 0)
$n_{00} = 7$ (número de atributos em que **p** = 0 e **q** = 0)
$n_{11} = 0$ (número de atributos em que **p** = 1 e **q** = 1)

$$CCS = (n_{11} + n_{00})/(n_{01} + n_{10} + n_{11} + n_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = n_{11} / (n_{01} + n_{10} + n_{11}) = 0 / (2 + 1 + 0) = 0$$

Exercício

- Calcular disssimilaridade entre **p** e **q** usando coeficientes:
  - Casamento Simples
  - Jaccard

$$\mathbf{p} = [1\ 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1\ 1\ 1\ 0]$$
$$\mathbf{q} = [0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1]$$

b.2) **Atributos Nominais** (não binários)

b.2.1) **Codificação 1-de-n**

- Exemplo:
  - Estado civil $\in$ {solteiro, casado, divorciado, viúvo}:
  - Criar 4 atributos binários: solteiro $\in \{0,1\}$, ... , viúvo $\in \{0,1\}$

- Atributos assimétricos
- Pode introduzir um número elevado de atributos.

b.2.2) **CCS e Jaccard (Adaptados)**

- Conversão nominal $\rightarrow$ binário não é estritamente necessária;
- Eventualmente ponderar contribuições individuais de cada atributo em função da cardinalidade do seu conjunto de valores

b.3) **Atributos Ordinais**

Ex.: Gravidade de um efeito: {nula, baixa, média, alta}

- Ordem dos valores é importante

- Normalizar e então utilizar medidas de (dis)similaridade para valores contínuos (p. ex. Euclidiana, cosseno, etc):

    - {1, 2, 3, 4} $\rightarrow$ (rank – 1) / (número de valores – 1)

    - {0, 1/3, 2/3, 1}

➢ Abordagem comum

c) Atributos Mistos (Contínuos e Discretos)

Método de Gower (1971):

$$s_{(\mathbf{x}_i,\mathbf{x}_j)} = \frac{1}{n}\sum_{k=1}^{n} s_{ijk} \longrightarrow d_{(\mathbf{x}_i,\mathbf{x}_j)} = 1 - S_{(\mathbf{x}_i,\mathbf{x}_j)}$$

Para atributos nominais / binários:

$$\begin{cases} (x_{ik} = x_{jk}) \Rightarrow s_{ijk} = 1; \\ (x_{ik} \neq x_{jk}) \Rightarrow s_{ijk} = 0; \end{cases}$$

Para atributos ordinais ou contínuos:

$$s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k \qquad R_k = \max_m \mathbf{x}_{mk} - \min_m \mathbf{x}_{mk}$$

R_k= faixa de observações do *k-ésimo* atributo *(termo de normalização)*

Sumário:

- Medidas de dis(similaridade) mais populares foram descritas, mas há várias outras na bibliografia

- Diferentes medidas de dis(similaridade) afetam a formação (indução) dos *clusters*

    - Como selecionar a medida de (dis)similaridade?

    - Devemos padronizar? Caso afirmativo, como?

- Infelizmente, não há respostas definitivas e globais...

- Análise de agrupamento de dados é, em essência, um processo subjetivo, dependente do problema

- Lembrem: **análise exploratória de dados**!

Algumas Questões Complementares...

Suponha que já se conheça um conjunto de pontos que pertençam a um grupo G_1 e que se considere esses pontos como mais ou menos próximos ao grupo como um todo segundo alguma medida de distância a partir do seu centro

Questão: Dado que a distância de um novo ponto (até então desconhecido) para o centro de G_1 é, digamos, d=5, o **quão próximo** de G_1 é de fato este ponto ?

- A quantificação (d=5) é absoluta, mas a interpretação é relativa

- **Teoria de Probabilidades** pode ajudar

A discussão anterior remete a uma questão fundamental quando se lida com diferentes medidas, índices, critérios para quantificar um determinado evento

Questão: Como interpretar um dado valor medido ?

Note que 0.9, por exemplo, não é necessariamente um valor significativamente alto de uma medida c/ escala 0 a 1

Depende de distribuições de probabilidade !

- Precisamos de uma distribuição de referência para avaliar a magnitude do valor da medida

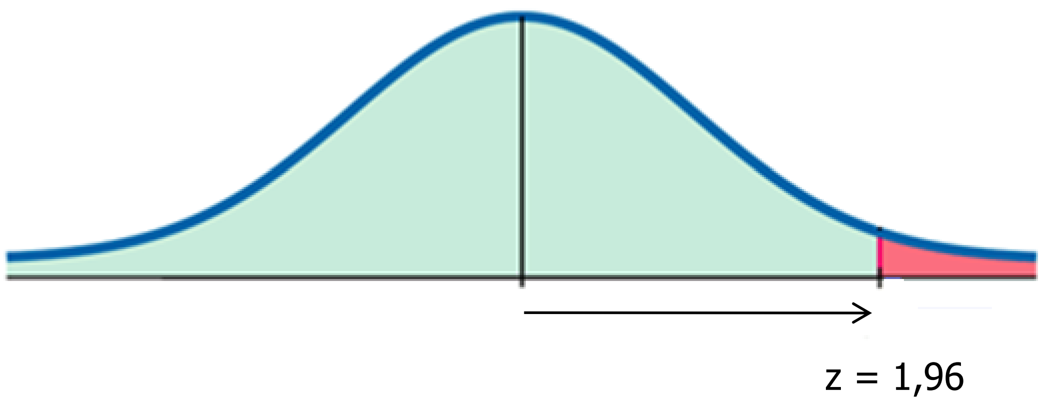
- Por hora, para fins do nosso exemplo simples, a distribuição de ref. pode ser a da distância “d” de interesse
  - de pontos gerados pelo fenômeno descrito por G_1 ao seu centro
- Suponha hipoteticamente que se conheça essa distribuição:
  - p. ex. normal com média μ e desvio padrão σ, ou seja, $N(\mu,\sigma)$
- Fazendo a padronização escore-z tem-se N(0,1)
  - $z = (d - \mu) / \sigma$
- Suponha mais uma vez hipoteticamente que a média e desvio sejam tais que nossa medida d = 5 implica z = 1,96
- **O que poderíamos concluir...?**

➢ Poderíamos concluir que a probabilidade de se observar um valor de distância d < 5 para um ponto de G_1 é 97,5%

➢ Isso pode sugerir que:

➢ um novo ponto observado com d = 5 não foi gerado pelo mesmo fenômeno descrito por G_1, ou

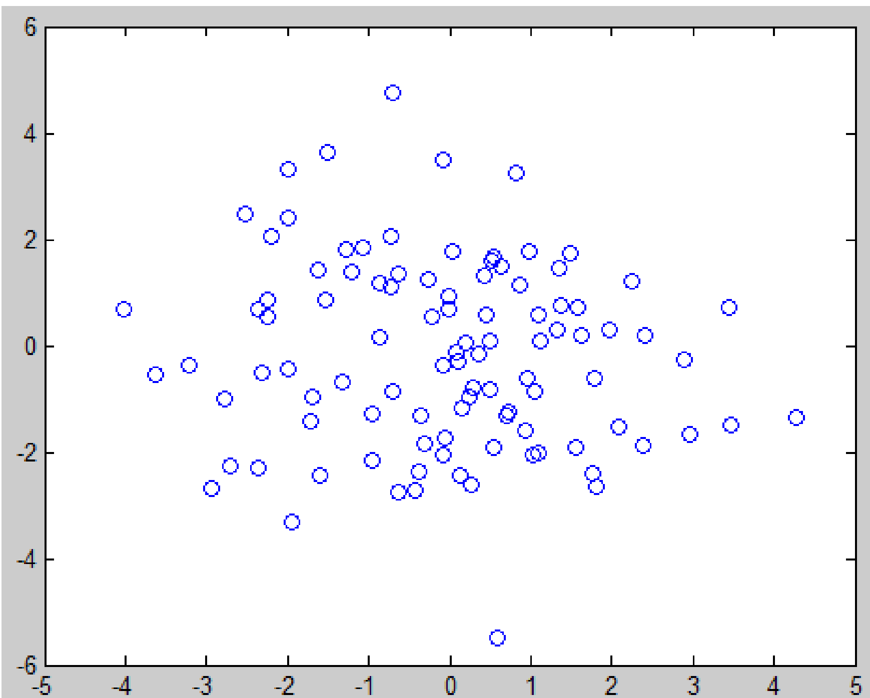
➢ esse ponto é um evento relativamente raro de G_1



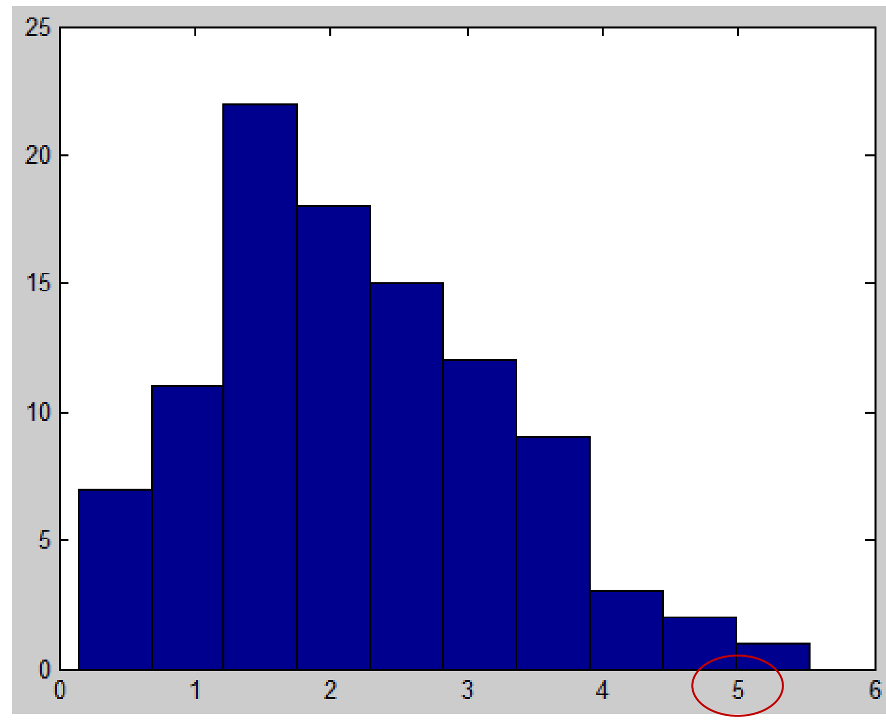
➢ Mas... e se não conhecemos a distribuição ?

- Se não conhecemos, podemos tentar estimar...

- No caso do nosso exemplo simples, podemos montar um histograma das distâncias dos pontos conhecidos de G_1



Dados (100 pontos 2D)



Histograma (10 bins) – Dist. Euclidiana

- Aprofundaremos essas questões mais adiante no curso...

Principais referências usadas para preparar essa aula:

- Xu, R., Wunsch, D., **Clustering**, IEEE Press, 2009
    - Capítulos 1 e 2, pp. 1-30
- Jain, A. K., Dubes, R. C., **Algorithms for Clustering Data**, Prentice Hall, 1988
    - Capítulos 1 e 2, pp. 1-25
- Gan, G., Ma, C., Wu, J., **Data Clustering: Theory, Algorithms, and Applications**, SIAM Series on Statistics and Applied Probability, 2007
    - Capítulos 1 e 2, pp. 1-24
- Kaufman, L., Rousseeuw, P. J., **Finding Groups in Data: An Introduction to Cluster Analysis**, 2a Edição, Wiley, 2005
    - Capítulo 1, seção 2

Outras Referências

- Everitt, B. S., Landau, S., Leese, M., *Cluster Analysis*, Hodder Arnold Publication, 2001
- P.-N. Tan, Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006
- Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, 2nd Edition, Wiley, 2001
- Triola, M. F., *Elementary Statistics*, 8ª Ed., Prentice-Hall, 2000