Sistemas Inteligentes Aplicados: Pré-processamento dos dados Parte 2

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Similaridade e dissimilaridade

- Similaridade: mede o quanto dois atributos são semelhantes
  - Quanto mais parecidos, maior o valor
  - Geralmente valor entre [0, 1]
- Dissimilaridade: mede o quanto dois atributos são diferentes
  - Quanto mais diferentes, maior o valor
  - Caso especial: medidas de distância

Similaridade e dissimilaridade (4)

- Exemplos de similaridade *s* e dissimilaridade *d* medidas para dois valores v_1 e v_1.
- Nominal:
  - $s = 1$ se $v_1 = v_2$
    0 se $v_1 \neq v_2$
  - $d = 0$ se $v_1 = v_2$
    1 se $v_1 \neq v_2$

Similaridade e dissimilaridade (5)

- Ordinal:
  - $s = 1 - (|v_1 - v_2| / (n - 1))$
  - $d = |v_1 - v_2| / (n - 1)$
  - n = número de valores ordinais

Similaridade e dissimilaridade (6)

- Intervalar ou razão:
  - $s_1 = 1 / (1 + |v_1 - v_2|)$ ou
  - $s_2 = 1 - d_2$
  - $d_1 = |v_1 - v_2|$
  - $d_2 = |v_1 - v_2| / \max(|v_i - v_j|)$
- Distâncias (a seguir) são boas medidas de dissimilaridade

Distância Euclidiana

- Medida clássica de distância
- Para duas instâncias û e v̂ no R^n, a distância é dada por:

$$d(\hat{u}, \hat{v}) = \sqrt{\sum_{i=0}^{n} (u_i - v_i)^2}$$

Distância Euclidiana (2)

- Exemplo:
  - $\hat{u} = (2, 1)$
  - $\hat{v} = (3, 4)$
  - $d = \sqrt{(2 - 3)^2 + (1 - 4)^2}$
    $= \sqrt{(-1)^2 + (3)^2}$
    $= \sqrt{10}$

Distâncias de Minkowski

- Generalização da distância euclidiana
- *r* é uma parâmetro escolhido pelo usuário

$$m_r(\hat{u}, \hat{v}) = \sqrt[r]{\sum_{i=0}^{n} \left(|u_i - v_i|\right)^r}$$

Distância de Minkowski

- Representa diferentes distâncias
  - m_1: distância bloco cidade (Manhattan), distância de hamming (valores binários e strings)
  - m_2: distância euclidiana

Distância de Minkowski

- m_{∞}: distância suprema (distância do eixo mais distante; distância dos atributos mais distantes)
  - Também chamada de distância quadrática ou distância de Chebyshev
  - $m_{\infty} = \max(|u_i - v_i|)$

Distância de Minkowski (2)

- Exemplos para û = (1, 2) e v̂ = (3, 5)
  - $m_1 = 5{,}000$
  - $m_2 = 3{,}606$
  - $m_3 = 3{,}271$
  - $m_4 = 3{,}138$
  - $m_5 = 3{,}075$
  - …
  - $m_\infty = 3{,}000 = \max(|u_i - v_i|)$

Avançado: Distância de Mahalanobis

- Generalização da distância euclidiana para atributos correlacionados que tenham distribuição normal
  - Leva em conta escala dos atributos
  - Leva em conta distribuição estatística das instâncias ao calcular distâncias
  - S: matriz de covariância (medida estatística obtida através da análise das instâncias)

Avançado: Distância de Mahalanobis (2)

- Cálculo da esperança **E(X)** para cada valor $\mathbf{x_i}$ com probabilidade $\mathbf{p_i}$ que o atributo **X** assume

$$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k \ .$$

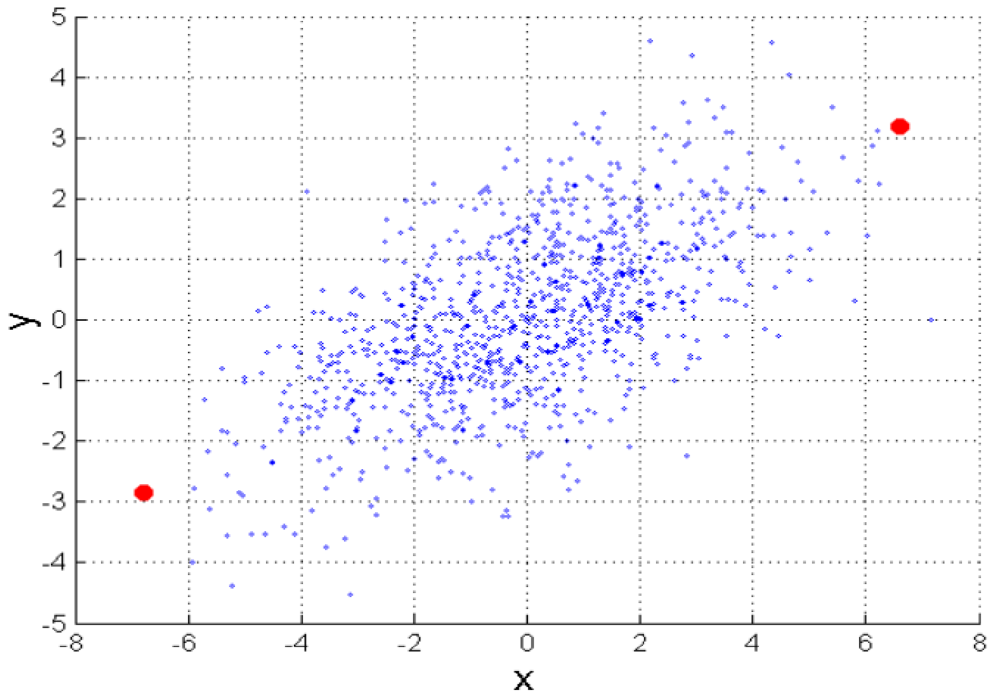
- Cálculo da Matrix de covariância S
para dois atributos $\mathbf{X_i}$ e $\mathbf{X_j}$
$S_{i,j} = E[(x_i - E[x_i])(x_j - E[x_j])]$

- Cálculo da distância de Mahalanobis

$$d(\hat{u}, \hat{v}) = \sqrt{(\hat{u} - \hat{v})^T S^{-1} (\hat{u} - \hat{v})}$$

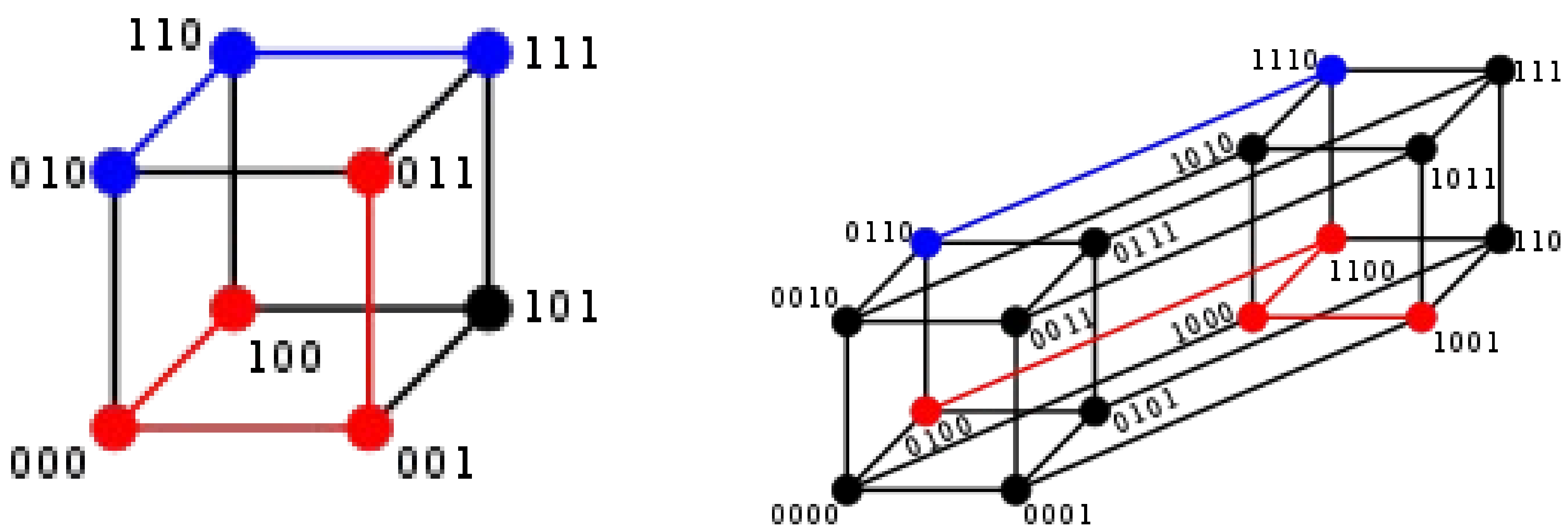
Avançado: Distância de Mahalanobis (3)

- Para pontos vermelhos: distância euclidiana = 14.7; Malahanobis = 6



Distância de Hamming

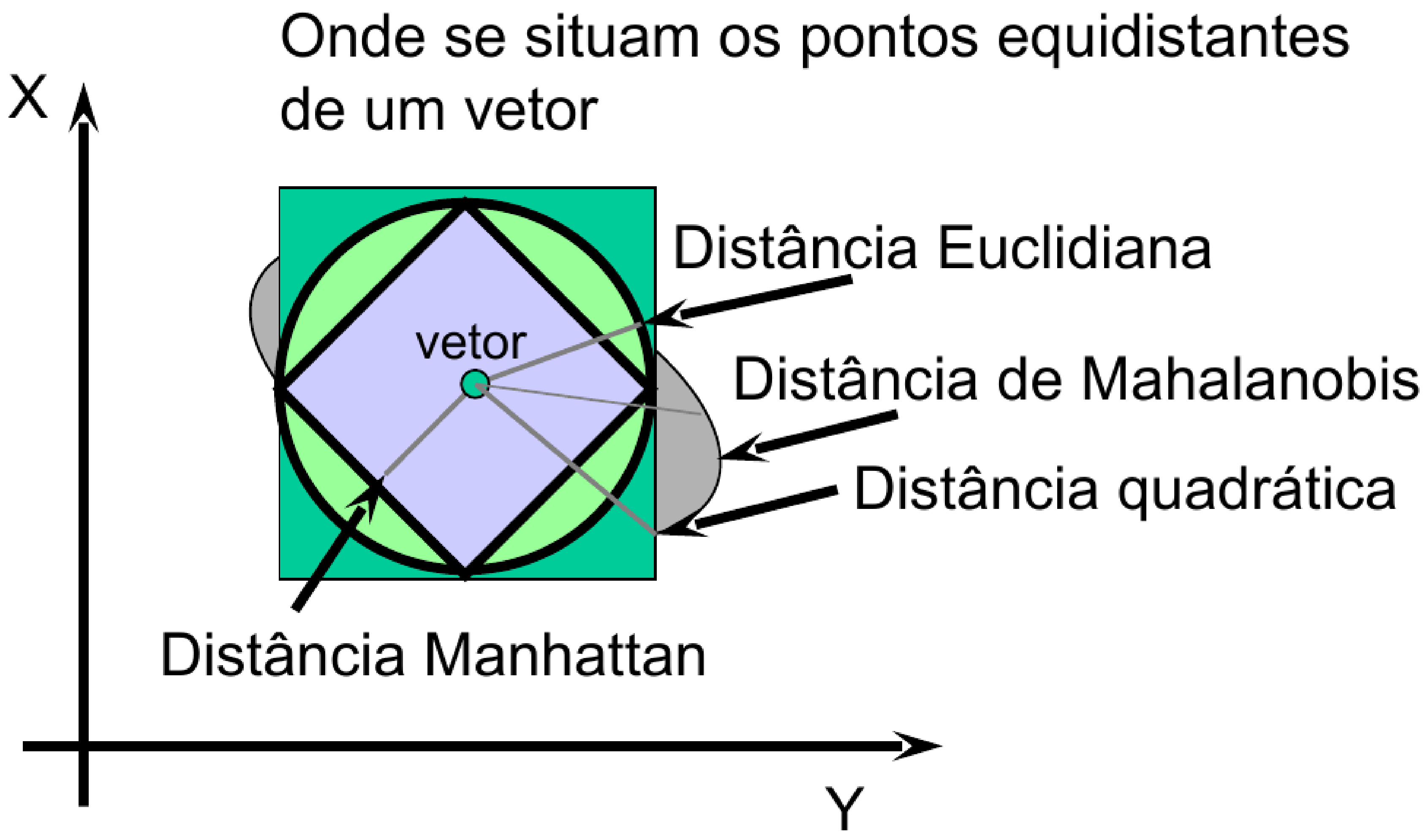
- Caso especial da distância Manhattan para valores binários (e também para Strings)
  - O cálculo é o mesmo



Distâncias

- Exercício
  - Para $\hat{u} = (1, 2, -3, 2)$ e $\hat{v} = (0, 6, 2, -1)$
  - Calcule as distâncias:
    - Manhattan
    - Euclidiana
    - Suprema

Propriedades das distâncias



Propriedades das distâncias (2)

- Medidas que satisfazem essas propriedades são chamadas **métricas**
  - $d(\hat{u}, \hat{v}) \geq 0 \ \forall \hat{u}, \hat{v}$
  - $d(\hat{u}, \hat{v}) = 0 \leftrightarrow \hat{u} = \hat{v}$
  - $d(\hat{u}, \hat{v}) = d(\hat{v}, \hat{u})$ (simetria)
  - $d(\hat{u}, \hat{w}) \leq d(\hat{u}, \hat{v}) + d(\hat{v}, \hat{w})$ (desigualdade triangular)

Propriedades das distâncias (3)

- Medidas de similaridade também possuem propriedades bem definidas
  - Seja $s(\hat{u}, \hat{v})$ a similaridade entre duas instâncias
  - $s(\hat{u}, \hat{v}) = 1 \leftrightarrow \hat{u} = \hat{v}$
  - $s(\hat{u}, \hat{v}) = s(\hat{v}, \hat{u})$

Conjuntos e vetores binários: (dis)similaridades

- Instâncias com apenas valores binários não são uma ocorrência incomum
- Conjuntos (exemplo, bag of words) podem ser mapeados para vetores binários
  - $X = \{A, B, C, D\}$
  - $Y = \{B, E, F\}$
  - $X \cup Y = \{A, B, C, D, E, F\}$
  - $X' = (1, 1, 1, 1, 0, 0)$
  - $Y' = (0, 1, 0, 0, 1, 1)$

Similaridade entre vetores binários

- Considerar duas instâncias originais û e v̂
  - $m_{0,0}$ número de atributos que possuem valor 0 tanto em û quanto v̂
  - $m_{1,0}$ número de atributos que possuem valor 1 em û e valor 0 em v̂
  - $m_{0,1}$ número de atributos que possuem valor 0 em û e valor 1 em v̂
  - $m_{1,1}$ número de atributos que possuem valor 1 tanto em û quanto v̂

Similaridade entre vetores binários (2)

- **Coeficiente de casamento simples**
  - $cs = (m_{0,0} + m_{1,1}) / (m_{0,0} + m_{0,1} + m_{1,0} + m_{1,1})$
- **Coeficiente de Jaccard** (recomendado para vetores muito esparsos)
  - $j = m_{1,1} / (m_{0,1} + m_{1,0} + m_{1,1})$

Similaridade entre vetores binários (3)

- Exemplo
  - $A = (1, 1, 1, 1, 0, 0)$
  - $B = (0, 1, 0, 0, 1, 0)$
  - $m_{0,0} = 1$
  - $m_{0,1} = 1$
  - $m_{1,0} = 3$
  - $m_{1,1} = 1$
  - $cs = (1 + 1)/ (1 + 1 + 3 + 1) \cong 0.333$
  - $j = 1 / (1 + 3 + 1) = 0.20$

Similaridade entre vetores binários (4)

- Exercício
  - Calcular similaridade usando casamento simples e coeficiente de Jaccard
  - `A = 1 0 0 1 1 0 1`
  - `B = 0 1 0 0 1 1 0`

Similaridade do cosseno

- Bastante usada em mineração de textos
  - Atributos assimétricos e esparsos
  - Trata instâncias como vetores no espaço
  - Extrair cosseno
  - Ideia geral: instâncias cujos vetores apontam na mesma direção devem ser similares
  - $\cos(\hat{u}, \hat{v}) = (\hat{u} \cdot \hat{v})/(||\hat{u}|| * ||\hat{v}||)$

Similaridade do cosseno (2)

- Exemplo 1
  - $\hat{u} = (2,3)$
  - $\hat{v} = (4,3)$
  - $|\hat{u} \cdot \hat{v}| = 2*4 + 3*3 = 17$
  - $||\hat{u}|| = \sqrt{(2^2 + 3^2)} = \sqrt{(13)} \cong 3.6$
  - $||\hat{v}|| = \sqrt{(4^2 + 3^2)} = \sqrt{(25)} = 5.0$
  - $\cos(\hat{u},\hat{v}) \cong 17 / (3.6 * 5.0) \cong 17/18$
    $\cong 0.94$

Similaridade do cosseno (3)

- Exemplo 2
  - $\hat{u} = (1,1)$
  - $\hat{v} = (3,3)$
  - $|\hat{u} \cdot \hat{v}| = 3 + 3 = 6$
  - $||\hat{u}|| = \sqrt{(1^2 + 1^2)} = \sqrt{2}$
  - $||\hat{v}|| = \sqrt{(3^2 + 3^2)} = \sqrt{18}$
  - $\cos(\hat{u},\hat{v}) = 6 / (\sqrt{2} * \sqrt{18}) = 6/\sqrt{36}$
    $= 6/6 = 1$

Similaridade do cosseno (4)

- Exercício: calcular o cosseno para
  - $\hat{u} = (1, 0, 0, 4)$
  - $\hat{v} = (0, 5, 0, 2)$

Correlação

- Medida se dois atributos x e y estão linearmente relacionadas
  - Valores no intervalo [-1, 1]

$$\frac{\sum\left(X_i-\overline{X}\right)\cdot\left(Y_i-\overline{Y}\right)}{\sqrt{\sum\left(X_i-\overline{X}\right)^2\cdot\sum\left(Y_i-\overline{Y}\right)^2}}$$

Correlação (2)

- Caso particular: podemos comparar a similaridade entre duas instâncias R^n:
- Isto é, ver se são proporcionais, semelhante a cosseno)
- Fazemos isso quebrando as instâncias originais em vetores do R^2

Correlação (2)

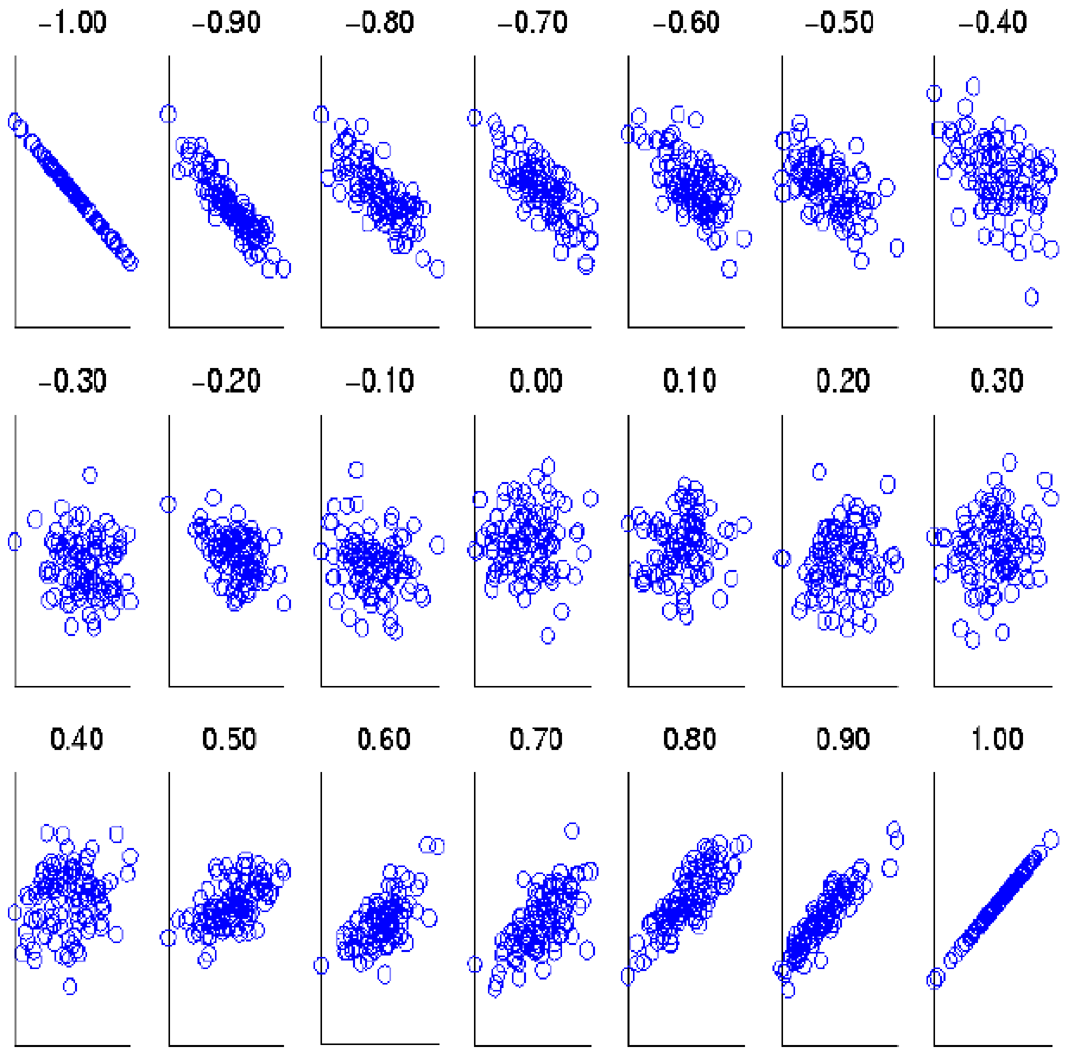
- Exemplo: **$\hat{u}$ = (0, 3, 4, -3)**; **$\hat{v}$ = (1, 1, 0, 0)**
  - $\hat{w}_1 = (0, 1)$; $\hat{w}_2 = (3, 1)$; $\hat{w}_3 = (4, 0)$; $\hat{w}_4 = (-3, 0)$;
  - Calcule a correlação:

$$\frac{\sum\left(X_i - \overline{X}\right).\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2 \cdot \sum\left(Y_i - \overline{Y}\right)^2}}$$

Correlação (3)

- $c(\hat{u}, \hat{v}) = 1$
  - Relacionamento linear positivo perfeito
- $c(\hat{u}, \hat{v}) = -1$
  - Relacionamento linear negativo perfeito
- $c(\hat{u}, \hat{v}) = 0$:
  - Não existe relacionamento
- Relacionamento linear: $u_i = av_i + b$

Correlação (4)



- Comparando instâncias û e v̂ (eixos)
- Pontos: valor de um atributo em um û e em v̂

Exercícios vistos em aula

- 1. Para û = (1, 2, -3, 2) e v̂ = (0, 6, 2, -1), calcule as distâncias: euclidiana; Manhattan; suprema
- 2. Calcular similaridade usando casamento simples e coeficiente de Jaccard

  ```
  A = 1 0 0 1 1 0 1
  B = 0 1 0 0 1 1 0
  ```
- 3. Calcular o cosseno para
  u = (1, 0, 0, 4)
  v = (0, 5, 0, 2)

Exercício

- 4. Correlação (visto em sala):
  $\hat{u} = (0, 3, 4, -3)$; $\hat{v} = (1, 1, 0, 0)$
- 5. No weka, rodar KNN para base Iris

Pontos chaves

- Escalas
- Similaridade: cosseno, casamento simples, Jaccard
- Dissimilaridade (todas as distâncias)

Agradecimentos/referências

- Notas de aula do Prof. André de Carvalho (USP)