Aula 7 – Medidas de Distância

Profa. Elaine Faria
UFU - 2017

Agradecimentos

Este material é baseado
- No livro Tan et al, 2006
- Nos slides do prof Andre C. P. L. F. Carvalho

- Agradecimentos
  - Ao professor André C. P. L. F. Carvalho que gentilmente cedeu seus slides

Transformação de dados

- Tarefa
  - Converter dados de
    - Numérico para categórico
    - Categórico para numérico
  - Normalizar dados

- Por que transformar dados?
  - Algumas técnicas trabalham apenas com dados numérico ou apenas com categóricos

Discretização e Binarização

- Discretizar
  - Transformar atributos contínuos em categórico
- Binarizar
  - Transformar atributos contínuos ou discretos em binário

O melhor método de discretização e binarização é aquele que produz o melhor resultado para o algoritmo de MD que será usado. No free lunch!

Binarização

- Codificação inteira-binária
  - Se há *m* valores categóricos
    - Associar cada valor original a um inteiro no intervalo [0,m-1]
      - Se o valor é ordinal → manter a ordem
    - Converter cada um dos *m* inteiros para um número binário
      - São necessários $n = \log_2 m$ dígitos binários
    - Ex: Variável categórica com 5 valores: péssimo, ruim, ok, bom, ótimo → 3 variáveis binárias

Binarização

- Codificação inteira-binária

Valor Categórico	Valor Inteiro	x_1	x_2	x_3
Péssimo	0	0	0	0
Ruim	1	0	0	1
Ok	2	0	1	0
Bom	3	0	1	1
Ótimo	4	1	0	0

Binarização

- Codificação 1-de-n
  - 1 atributo binário para cada valor categórico
  - Ex: Variável categórica com 5 valores: péssimo, ruim, Ok, bom, ótimo → 5 variáveis binárias

Quais os problemas com a codificação inteira?

Quais os problemas com a codificação 1-de-n?

Binarização

- Codificação 1-de-n

Valor Categórico	**Valor Inteiro**	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$
Péssimo	0	1	0	0	0	0
Ruim	1	0	1	0	0	0
Ok	2	0	0	1	0	0
Bom	3	0	0	0	1	0
Ótimo	4	0	0	0	0	1

Binarização

- Codificar usando codificação 1-de-n os valores:

    amarelo,

    vermelho,

    verde,

    azul,

    laranja,

    branco

Binarização

- Imagine que um atributo seja nome de país
  - Existem 193 países (192 representados na ONU + Vaticano)
  - Transformar valores nominais em valores numéricos utilizando a codificação 1-de-n

Qual o problema em usar a codificação 1-de-n?

Discretização de Atributos Contínuos

- Tarefas
  - Decidir quantos categorias
    - Dividir os valores dos atributos em *n* intervalos, especificando *n-1* pontos de divisão
  - Decidir como mapear os valores contínuos em categorias
    - Todos os valores em um intervalos são mapeados para o mesmo valor categórico
- Representação
  - $x_0 < x <= x_1$, $x_1 < x <= x_2$, ... → intervalos
  - $\{(x_0, x_1], (x_1, x_2], \ldots$ → desigualdade

Discretização de Atributos Contínuos

- Discretização Não-supervisionada
  - Prop. 1: Larguras Iguais
    - Dividir o atributo em um número de intervalos especificado pelo usuário (todos do mesmo tamanho)
  - Prop. 2: Frequências Iguais
    - Dividir o atributo em intervalos, de modo que cada um tenha a mesma quantidade de exemplos

Discretização de Atributos Contínuos

- Discretização Não-supervisionada
  - Prop 3. Inspeção Visual
    - Determinar visualmente qual é a melhor forma de discretizar os dados
  - Prop 4: Algoritmos de agrupamento
    - Usar algoritmos de agrupamento para encontrar a melhor forma de discretizar os dados

Discretização de Atributos

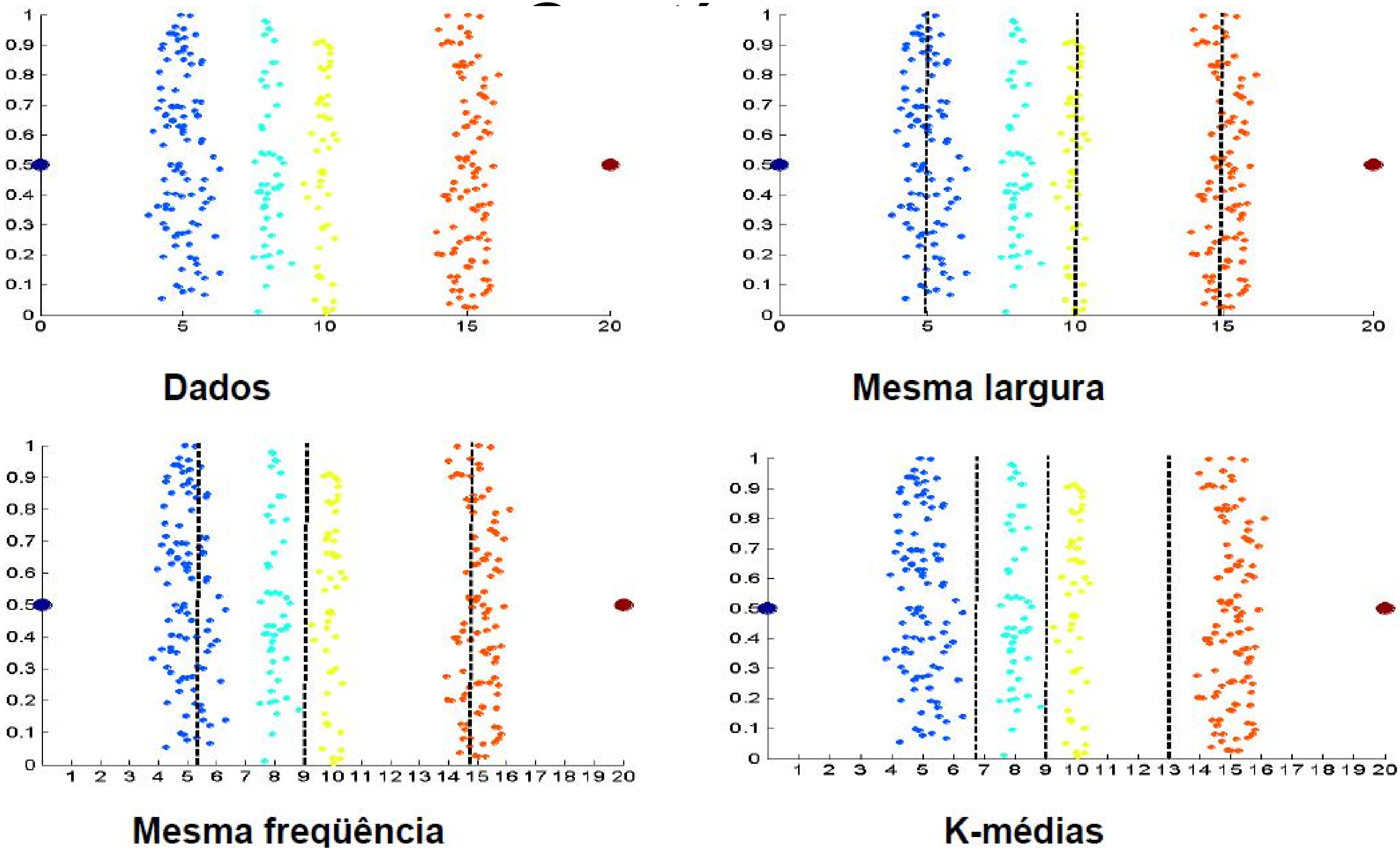


Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Discretização de Atributos Contínuos

- Discretizar o atributo que possui os valores abaixo em 3 intervalos

  0, 1, 3, 6, 6, 9, 10, 10, 10, 13, 18, 20, 21,21, 25

- Usar:
  - Larguras iguais
  - Frequencias iguais

Transformação de Atributos

- Transformação aplicada a todos os valores da variável
- Motivação
  - Grande variação de valores
  - Limites dos valores são muitos diferentes
  - Evitar que um atributo predomine sobre o outro
  - Propriedades estatísticas desejadas
- Tipo de transformação
  - Função simples
  - Normalização

Transformação de Atributos

- Por que é importante aplicar transformação de atributos?
  – Ex: comparar duas pessoas usando duas variáveis: idade e salário
    - A diferença entre salário será muito maior do que entre idade
    - A diferença entre duas pessoas será dominada pelo atributo salário

Transformação de Atributos

- Função simples
  - Uma função matemática simples é aplicada a cada valor individualmente
  - Ex: Seja x a variável
    - Exemplo de funções: x^k, $\log x$, $\sin x$, $1/x$ ou $|x|$

Qual função escolher?

R: Depende do problema

Transformação de Atributos

- Cuidado no uso de funções simples
  - Podem mudar a ordem dos valores
  - Ex.: Uso da função 1/x para x = 0,2;0,5;1;2; 4
    - Novos valores: 5; 2; 1; 0,5; 0,25
    - Reverte a ordem dos valores
      - Valores menores se tornam maiores (e vice-versa)
    - Se um dos valores fosse 0?

Transformação de Atributos

- Normalização
  - Objetivo: fazer um conjunto de valores ter uma propriedade particular
  - Tipos de normalização
    - Re-escalar
    - Padronizar

Transformação de Atributos

- Re-escalar
  - Mudar a unidade de medida dos dados
  - Propriedade: colocar os valores mínimos e máximos iguais
  - Como fazer
    - Adicionar ou subtrair uma constante
    - Multiplicar ou dividir por uma constante
  - Ex: converter os valores para o intervalo [0,1]

$$d' = \frac{(d - \min_d)}{(\max_d - \min_d)}$$

Transformação de Atributos

- Padronização
  - Como fazer:
    - Adicionar ou subtrair uma medida de localização
    - Multiplicar ou dividir por uma medida de escala
  - Ex: $\bar{x}$ é o valor médio de um atributo e s_x é o seu desvio padrão, então

    $x' = (\bar{x} – x)/ s_x$

    Cria uma variável que tem média zero e desvio padrão 1

Transformação de Atributos

- Converter os seguintes valores numéricos utilizando re-escala e padronização
  - [0, 1] e normal (0,1)

Valores	Re-escala	Padronização
3		
9		
5		
11		
5		
7		

Medidas de Similaridade e Dissimilaridade

- Importância
  - São usadas em uma série de técnicas de MD e AM. Ex: agrupamento, KNN e detecção de novidade

- Pode ser visto com uma transformação dos dados para um espaço de similaridade (dissimilaridade)
  - Em muitos casos o conjunto de dados inicial não é necessário para executar a técnica de MD → apenas as medidas de similaridade ou dissimilaridade são suficientes

- Proximidade entre objetos refere-se à proximidade entre seus atributos

Medidas de Similaridade e Dissimilaridade

- Similaridade entre dois objetos
  - É uma medida numérica do quão parecido dois objetos são
    - Objetos parecidos → similaridade alta
  - É um número não negativo entre 0 (não similar) e 1 (completamente similares)

- Dissimilaridade entre dois objetos
  - É uma medida numérica do quão diferente dois objetos são
    - Objetos similares → dissimilaridade baixa
  - Está no intervalo [0,1] ou [0, ∞]
  - Distância é um sinônimo (tipo especial de dissimilaridade)

Medidas de Similaridade e Dissimilaridade

- Transformação
  - Converter similaridade para dissimilaridade ou vice-versa
  - Transformar uma medida de proximidade para um intervalo particular, ex: [0,1]

  Ex: medida de similaridade no intervalo [1,10], mas o algoritmo só trabalha com similaridade entre [0,1] → aplicar transformação

$$s' = (s - min_s)/ (max_s - min_s)$$

$$s' = (s - 1)/9$$

Medidas de Similaridade e Dissimilaridade

- Transformação

Ex: Medida no intervalo [0,∞], converter para [0,1] → transformação não-linear

$$d' = d/(1+d)$$

  - Os valores não terão o mesmo relacionamento entre si na nova escala
  - Ex: 0; 0,5; 2; 10; 100 e 1000 serão convertidos para 0; 0,33; 0,67; 0,90; 0,99; 0,999

Medidas de Similaridade e Dissimilaridade

- Transformação: similaridade para dissimilaridade
  - Se está no intervalo [0,1]

    $d = 1 – s$ (ou $s = d – 1$)

  - Se não está no intervalo [0,1]

    $s = 1/(d+1)$, $s = e^{(-d)}$, $s = 1 – ((d – min)/(max – min))$

Similaridade e Dissimilaridade entre Atributos Simples

- Proximidade com 1 atributo

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Tabela retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Dissimilaridade entre Objetos

- Existem várias medidas de dissimilaridade
  - Diferentes medidas podem ser aplicadas a diferentes problemas
- Objetos (ou Instâncias) são descritos por *n* atributos
  - Calcular a medida de dissimilaridade usando os *n* atributos
  - Em geral, usa-se medidas de distância
- Distância
  - Medida de dissimilaridade que possui certas propriedades (ver slide 35)

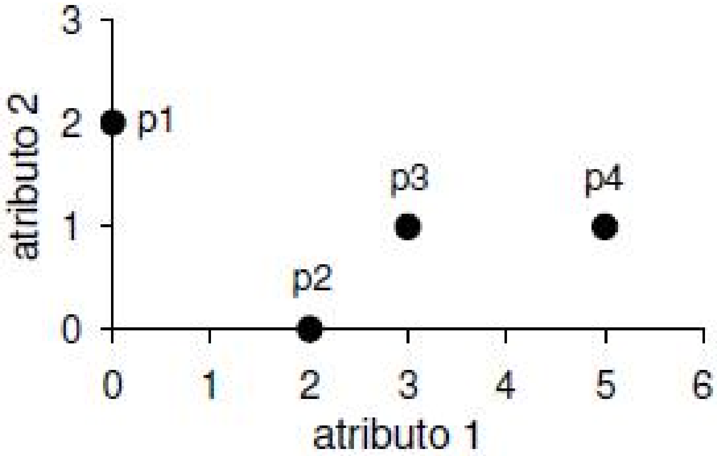
Medidas de Distância

- Distância Euclidiana
  - Distância *d* entre dois objetos x e y em um espaço *n* dimensional

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

X_k e y_k são o k-ésimo atributo dos objetos x e y

Medidas de Distância



	atributo 1	atributo 2
p1	0	2
p2	2	0
p3	3	1
p4	5	1

matriz de distância

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Imagem retirada dos slides do prof. Ricardo J. G. B. Campello – disciplina Aprendizado de Máquina – ICMC-USP

Medidas de Distância

- Distância de Minkowski
  – Generalização da distância Euclidiana

$$d(x, y) = \left(\sum_{k=1}^{n} | x_k - y_k |^r \right)^{\frac{1}{r}}$$

  - r = 1 : distância *city block* (*Manhattan* ou L_1 norm)
  - r = 2 : distância Euclidiana (L_2 norm)
  - r = ∞ : distância Suprema (L_{max} ou L_∞ norm)

$$d(x, y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} | x_k - y_k |^r \right)^{\frac{1}{r}}$$

Medidas de Distância

- Construir a matriz de distâncias para o exemplo do slide 32 usando
  - L1

    Ver solução no Livro “Introduction to Data Mining”

Propriedades das distâncias

- Positividade
  - $d(x,x) >= 0$ para todo x e y
  - $d(x,y)=0$ somente se $x = y$
- Simetria
  - $d(x,y) = d(y,x)$ para todo x e y
- Desigualdade triangular
  - $d(x,z) <= d(x,y) + d(y,z)$ para todos os objetos x, y e z.

Propriedades das distâncias

- Medidas que satisfazem as 3 propriedades → métricas
- Ex. de medida de dissimilaridade que não é métrica

  Conjuntos A e B

  A – B: elementos que estão em A e não estão em B

  dist(A,B) = tamanho (A – B)

  Não atende a 2ª parte da propriedade da positividade, nem a simetria, nem a desigualdade triangular.

Similaridade entre Objetos

- Propriedades
  - s(x,y) = 1 somente se x = y (0 <= s <= 1)
  - s(x,y)=s(y,x) para todo x e y

  - Não há uma propriedade análoga à desigualdade triangular para medidas de similaridade

Medidas de Proximidade

- Medidas de similaridade para vetores binários
  - Chamadas de coeficiente de similaridade
  - Possuem valores entre 0 e 1 → 1: objetos completamente similares, 0: objetos não similares
  - Comparando objetos x e y que consistem de *n* atributos binários (vetores binários)
    - f_{00} = nro de atributos em que x=0 e y=0
    - f_{01} = nro de atributos em que x=0 e y=1
    - f_{10} = nro de atributos em que x=1 e y=0
    - f_{11} = nro de atributos em que x=1 e y=1

Medidas de Proximidade

- Medidas de similaridade para vetores binários: **Coeficiente de casamento simples**

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{00} + f_{11}}$$

  - Conta as presenças e ausências igualmente
  - Ex: encontrar os estudantes de que responderam de forma similar a um teste que consiste de questões true/false.

Medidas de Proximidade

- Medidas de similaridade para dados binários: **Coeficiente de Jaccard**

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

  - Usado para atributos binários assimétricos
  - Não considera as coincidências de 0s

Medidas de Proximidade

- Ex:

x = (1,0,0,0,0,0,0,0,0,0)

y = (0,0,0,0,0,0,1,0,0,1)

$f_{01} = 2$ número de atributos em que x= 0 e y=1

$f_{00} = 7$ número de atributos em que x= 0 e y=0

$f_{10} = 1$ número de atributos em que x= 1 e y=0

$f_{11} = 0$ número de atributos em que x= 1 e y=1

SMC = 0 + 7/ (2+1+0+7) = 0.7

J = 0/ (2+1+0) = 0

Exercício

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

p = 1 0 0 1 1 0 1 0 1 1 1 0

q = 0 1 0 0 1 1 0 0 1 0 1 1

Medidas de Proximidade

- Similaridade Cosseno
  - É uma medida do ângulo entre x e y. Se a similaridade é 1, o ângulo entre x e y é 0^0; se a similaridade é 0, o ângulo é 90^0

$$\cos(x, y) = \frac{x.y}{\|x\|\|y\|}$$

$$x.y = \sum_{k=1}^{n} x_k y_k$$

$$\|x\| = \sqrt{\sum_{k=1}^{n} {x_k}^2}$$

. → produto interno de dois vetores,

||x|| → é o tamanho (norma) do vetor x

Medidas de Proximidade

- Similaridade Cosseno

Ex. Sejam os vetores

x = (3,2,0,5,0,0,0,2,0,0)

y = (1,0,0,0,0,0,0,1,0,2)

x.y = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5

||x|| = sqrt(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0) = 6.48

||y|| = sqrt(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) = 2.24

cos(x,y) = **0.31**

Medidas de Proximidade

- Similaridade Cosseno
  - Muito usado em mineração de texto
    - Documentos são vetores, cada atributo representa a frequência de ocorrência de um termo (palavra) no documento
    - Cada documento é esparso (poucos atributos não zero)

Medidas de Proximidade

- Similaridade Cosseno
  - Calcular disssimilaridade entre p e q usando medida de similaridade cosseno:

  p = 1 0 0 4 1 0 0 3

  q = 0 5 0 2 3 1 0 4

Medidas de Proximidade

- Correlação
  - Medida de relacionamento linear entre os atributos dos objetos
  - Pode também ser usada para medir o relacionamento entre dois atributos
  - Correlação muito usada na literatura → Correlação de Pearson

Medidas de Proximidade

- Correlação de Pearson

$$corr(x, y) = \frac{\operatorname{cov} ariancia(x, y)}{desvio\_padrao(x) * desvio\_padrao(y)}$$

$$\operatorname{cov} ariancia(x, y) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

$$desvio\_padrao(x) = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Medidas de Proximidade

- Correlação de Pearson

$x'_k = (x_k – \text{media}(x))/\text{desvio\_padrao}(x)$

$y'_k = (y_k – \text{media}(y))/\text{desvio\_padrao}(y)$

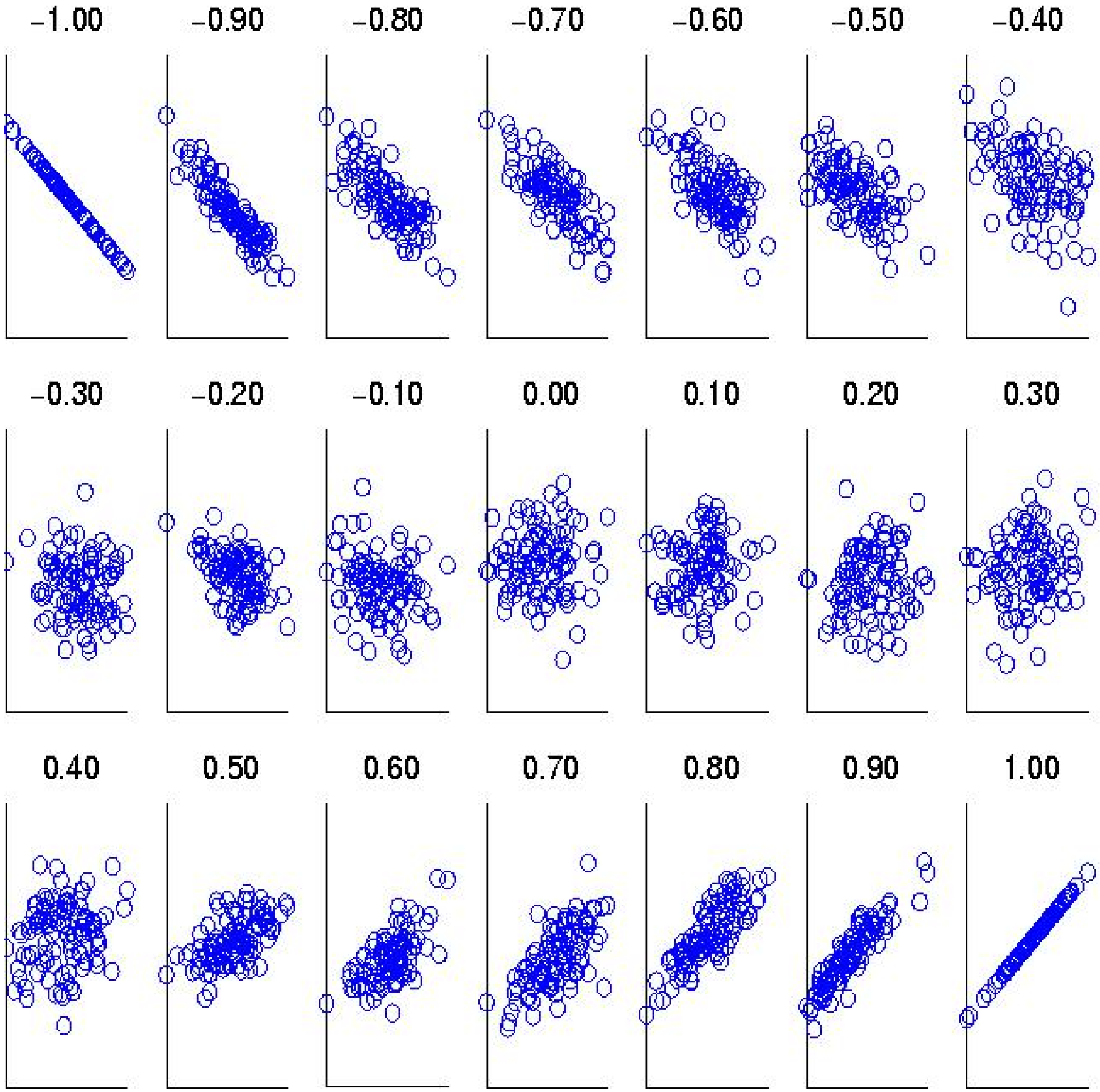
$$\text{correlacao}(x,y) = x' \,.\, y'$$

Medidas de Proximidade

- Correlação de Pearson
  - Valor no intervalo [-1,1]
    - +1: objetos tem um relacionamento linear positivo

      $x_k = ay_k + b$, sendo *a* e *b* constantes
    - -1: objetos tem um relacionamento linear negativo
    - 0: não há correlação

Medidas de Proximidade



- Similaridade entre objetos x e y, cada um com 30 atributos
- Similaridade variando de -1 a 1

Tabela retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Problemas no Cálculo de Medidas de Proximidade

- Como tratar a situação quando os atributos não tem o mesmo intervalo de valores?

- Como tratar a situação na qual os atributos tem pesos diferentes?

Distância de Mahalanobis

- Generalização da distância Euclidiana
  - Não é esférica, mas elipsoidal

- Usada quando
  - Há correlação entre alguns atributos
  - Os atributos possuem diferentes escalas
  - Distribuição dos dados é aproximadamente Gaussiana (normal)

- Desvantagem
  - Cara computacionalmente

Calculando a similaridade entre objetos com diferentes tipos de atributos

Algorithm 2.1 Similarities of heterogeneous objects.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range $[0, 1]$.
2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is an asymmetric attribute and} \\ & \text{both objects have a value of 0, or if one of the objects} \\ & \text{has a missing value for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3: Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k} \tag{2.15}$$

Imagem reitrada de Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Usando Pesos

- Modificação da medida de proximidade para ponderar a contribuição de cada atributo
  – Peso (*w*) sumariza 1

$$similaridade(x,y) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

$$dist(x,y) = \left(\sum_{k=1}^{n} w_k \mid x_k - y_k \mid^{r} \right)^{\frac{1}{r}}$$

inserindo peso na distância de Minkowski

Tarefa

- Leitura do Capítulo 2 (Seção 2.4) do livro Tan et al, 2006

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.