SCC0173 – Mineração de Dados Biológicos

Preparação de Dados: Parte B

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

Créditos

- O material a seguir consiste de adaptações e extensões:
  - dos originais gentilmente cedidos pelo professor André C. P. L. F. de Carvalho
  - dos originais de Tan et al., *Introduction to Data Mining*, Addison-Wesley, 2006

Tópicos

- Transformação de dados
  - Conversões e Discretização
- Amostragem
- Cálculo de Proximidade
  - Medidas de (dis)similaridade

Transformação de Dados

- Normalização de valores numéricos
  - visto na aula anterior...
- Conversão de valores simbólicos para numéricos
- Conversão de valores numéricos para simbólicos

Conversão de Valores Categóricos

- Algumas técnicas trabalham apenas com variáveis numéricas
  - Por exemplo, redes neurais
  - Variáveis categóricas precisam ser convertidas
- Conversão depende da existência ou não de ordem entre os valores
  - Variáveis nominais ou ordinais

Conversão de Valores Ordinais

- Para variáveis **ordinais**, a ordem dos valores deve ser de alguma maneira mantida
  - Normalmente associa-se valores inteiros crescentes a cada valor simbólico
    - Por exemplo, {frio, morno, quente} = {1, 2, 3}

Conversão de Valores Nominais

- Atributos **nominais**
  - Conversão é feita por **binarização**
  - Codificação mais usual
    - Codificação 1-de-n (canônica)

Conversão de Valores Nominais

- Codificação 1-de-n
  - Um atributo binário associado a cada valor nominal
  - Exemplo:
    - {amarelo, vermelho, verde, azul, branco}
      - no quadro...
  - Pode gerar um número grande de atributos binários, mas possui propriedades interessantes

Conversão de Valores Nominais

- **Nota:** se o atributo nominal já for binário, pode não ser necessária a conversão em dois atributos
  - Depende do contexto
  - Exemplo: "Matriculado na Disciplina A" ∈ {F, V}
    - Convertido em "Matriculado na Disciplina A" ∈ {0, 1}

Exercício

- Converter os dados abaixo para valores numéricos e normalizá-los em [0, 1]

Febre	Enjôo	Mancha	Dor	Diagnóstico
baixa	sim	pequena	A	doente
média	não	média	C	saudável
alta	sim	grande	B	saudável
alta	não	pequena	A	doente
baixa	não	grande	D	saudável
média	não	ausente	C	doente

Discretização

- Alguns algoritmos de DM aceitam apenas valores categóricos
  - Demandam discretizar valores contínuos em intervalos
- Melhor discretização depende de:
  - Algoritmo que utilizará os valores discretizados
  - Demais atributos
  - ...

Discretização

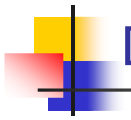
- Transformar valores contínuos em intervalos
  - podem ser vistos como valores categóricos ordinais
- Sub-tarefas
  - Definição do número de categorias
    - Geralmente feito pelo usuário
  - Definição dos limites e tamanho dos intervalos
    - Geralmente feito pelo algoritmo

Discretização

- Passo 1: definir no. e limites dos intervalos
  - Ordenar atributos pelos seus valores
  - Dividir em n intervalos
    - Definindo n-1 pontos de corte ou divisão
- Passo 2: mapear para categorias
  - Todos os valores dentro de um intervalo são mapeados para o mesmo valor categórico
- Problema se resume ao Passo 1
  - Quantas divisões e onde colocá-las

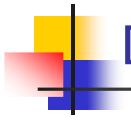
Discretização

- Existem vários algoritmos na literatura
- Algoritmos podem ser divididos como:
  - Não supervisionados
    - utilizam somente os valores do atributo a ser discretizado
  - Supervisionados
    - direcionados para classificação
    - usam informação das classes das respectivas instâncias



Discretização Não Supervisionada

- Algoritmos Simples
  - **Larguras Iguais**
    - Divide intervalo original de valores em n sub-intervalos com mesma largura
  - **Freqüências Iguais**
    - Atribui o mesmo no. de objetos a cada sub-intervalo



Discretização Não Supervisionada

- **Inspeção Visual**
  - Observa gráfico com valores dos atributos e determina visualmente os intervalos de acordo com a distribuição natural dos dados
- **Clustering**
  - Utiliza algum algoritmo de agrupamento de dados para descobrir automaticamente a distribuição dos dados

Exercício

- Discretizar 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
  - Freqüências iguais
  - Inspeção visual

Discretização Supervisionada

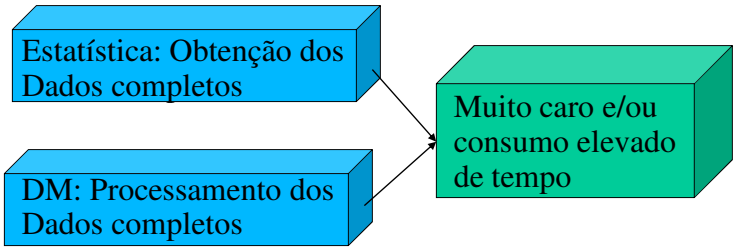
- Discutiremos posteriormente no curso...

Amostragem de Dados

- Com os dados pré-processados e transformados, pode ser necessário ou interessante selecionar sub-amostras...

Amostragem dos Dados

- Seleção de um subconjunto de instâncias (**amostra**)
- Técnica fundamental em Estatística e também em Mineração de Dados
  - tanto para investigações preliminares como definitivas



Amostragem dos Dados

- **Amostragem Aleatória Simples**
  - Tipo mais comum em DM, com 2 variações
    - Sem reposição
    - Com reposição
      - Mais simples de analisar, pois probabilidade de escolher qualquer objeto se mantém constante
      - Porém permite inserção de duplicatas

Amostragem dos Dados

- Espera-se levar à mesma acurácia (ou similar) com um esforço computacional muito menor
  - Algoritmo de DM só processa parte das instâncias
- Amostra deve ser representativa
  - Se não for suficientemente representativa, o tamanho da amostra passa a representar um compromisso eficiência computacional × eficácia

Amostragem dos Dados

- Influência do tamanho:

8000 pontos | 2000 Pontos | 500 Pontos

Amostragem dos Dados

- Amostra **representativa**
  - Aproximadamente as mesmas propriedades de interesse do conjunto de dados original
    - Ex.: $média_{pop\text{-}original} = média_{amostra}$
  - Deve fornecer uma estimativa da informação desejada contida na população original
    - Assim, uso da amostra tem efeito semelhante ao uso de toda a população

Amostragem dos Dados

- Amostra **representativa**
  - Deve permitir tirar conclusões de um todo a partir de uma parte
  - Não é possível garantir que isso ocorra
    - É particularmente difícil em tarefas não supervisionadas (p. ex. agrupamento de dados)

Amostragem dos Dados

- Amostra **representativa**
  - Para aumentar as chances que a amostra seja representativa, existem diferentes técnicas de amostragem já bem investigadas
  - Por exemplo, **amostragem estratificada**
    - Usada em problemas de **classificação** para garantir a representatividade de todas as classes nos dados

Amostragem dos Dados

- Qual o melhor tamanho?
  - Difícil responder
  - Grande:
    - Aumenta chance da amostra ser representativa
    - Reduz vantagens da amostragem
  - Pequeno:
    - Reduz custo computacional
    - Aumenta chance de perda de informação

Amostragem dos Dados

- **Amostragem progressiva**
  - Começa com pequenas amostras
    - Progressivamente aumenta tamanho da amostra enquanto houver variabilidade significativa nos modelos obtidos
      - por exemplo, na acurácia de um classificador

Conheça seus Dados!

- Conhecer bem a natureza dos dados é algo fundamental antes de querer aprender qualquer coisa a partir deles
  - Por exemplo, saber de antemão que dois atributos como salário e imposto retido na fonte podem ser redundantes é muito útil !
- Domínios específicos podem requer ferramentas específicas, completamente distintas de outros domínios !
  - Conhecer bem os dados passa por conhecer bem o domínio de aplicação que produziu esses dados

Similaridade e Dissimilaridade

- Muitos algoritmos de mineração de dados operam totalmente ou parcialmente com base em cálculos e comparações de algum tipo de similaridade ou dissimilaridade entre instâncias (ou atributos) dos dados
  - classificador K-NN, agrupamento k-médias, ...
- Veremos algumas das mais comuns dentre as tantas maneiras de se calcular (dis)similaridades

Similaridade e Dissimilaridade

- **Similaridade**
  - Mede o quanto duas instâncias são parecidas
    - quanto mais parecidos, 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, ∞]

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...

Atributos Numéricos e Distância

- Muitos problemas de DM envolvem apenas atributos numéricos
- Além disso, já vimos que é possível converter atributos categóricos em numéricos para a aplicação de ferramentas de DM que só lidam com esse tipo de atributo
- Para duas instâncias descritas por um conjunto de **n** atributos numéricos, a forma mais usual de se medir dissimilaridade entre elas é o uso de uma **medida de distância**
  - Medida de distância mais popular: Euclidiana

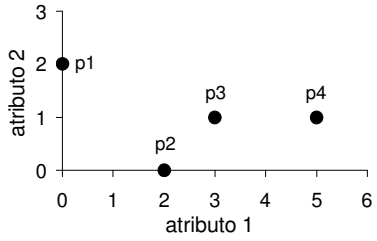
Distância Euclidiana

- Distância entre duas instâncias $\mathbf{p_i}$ e $\mathbf{p_j}$ definida como:

$$d = \sqrt{\sum_{k=1}^{n} (p_{ik} - p_{jk})^2}$$

  - onde $\mathbf{p_{ik}}$ e $\mathbf{p_{jk}}$ para k = 1, ..., **n** são os **n** atributos que descrevem as instâncias $\mathbf{p_i}$ e $\mathbf{p_j}$, respectivamente
- Dá o mesmo peso para todos os atributos...
  - pode ser necessário padronização ou re-escala

Distância Euclidiana

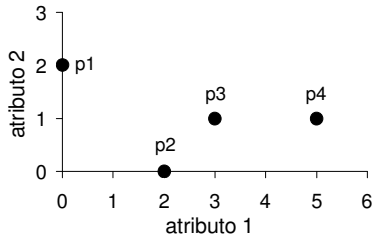


	atributo 1	atributo 2
p1	0	2
p2	2	0
p3	3	1
p4	5	1

matriz de distância

	p1	p2	p3	p4
p1				
p2				
p3				
p4				

Distância Euclidiana



	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

Distância de Minkowski

- Generalização da distância Euclidiana:

$$d = (\sum_{k=1}^{n} | p_{ik} - p_{jk} |^r)^{\frac{1}{r}}$$

- Valor de **r** leva a diferentes distâncias, por exemplo:
  - 1 (L_1): Distância de Manhattan
  - 2 (L_2): Distância Euclidiana

Distância de Manhattan

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

Matriz de Distância

L1	p1	p2	p3	p4
p1				
p2				
p3				
p4				

Distância de Manhattan

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

Matriz de Distância

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

Exercício: Transforme o conjunto de distâncias acima em similaridades em [0,1], das diferentes formas vistas que forem aplicáveis

Exercício

- Calcular dissimilaridade entre **p** e **q** usando as distâncias:
  - Manhattan
  - Euclidiana

p = [1 2 –3 2 0 8]
q = [0 6 2 –1 2 5]

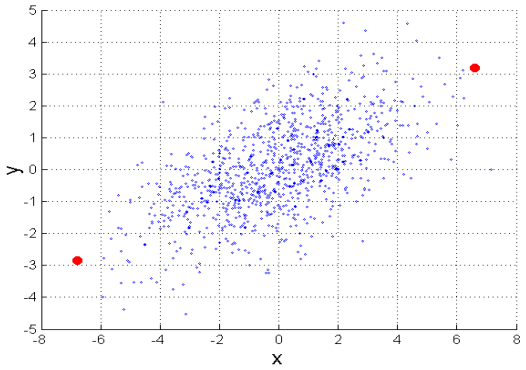
Distância de Mahalanobis

- Outra generalização da distância Euclidiana
  - distância "elíptica", não mais "esférica"...
- Particularmente útil quando:
  - Atributos são correlacionados
  - Mas é computacionalmente pesada...

Distância de Mahalanobis

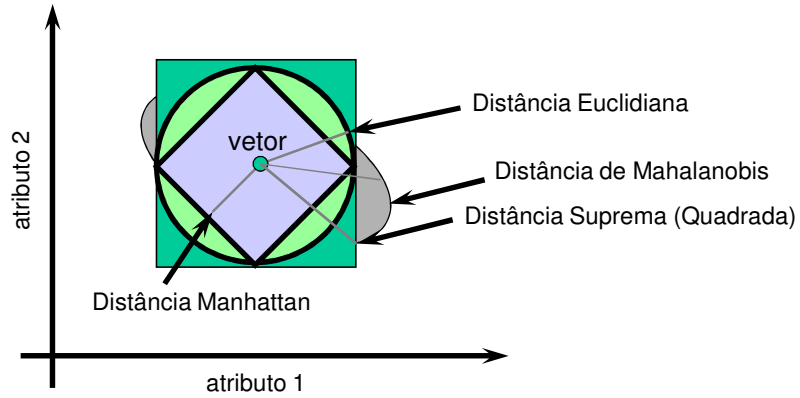
Para pontos vermelhos:

Distância Euclidiana = 14.7
Distância de Mahalanobis = 6



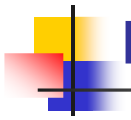
Medidas de Distância

- Onde se situam os pontos eqüidistantes de um vetor



Propriedades de Distâncias

- Seja **d**(**p**, **q**) a distância entre duas instâncias **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 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**)



Propriedades de Similaridade

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

Similaridade com Atributos Binários

- Freqüentemente, instâncias **p** e **q** são descritas apenas por atributos binários
- Similaridades podem ser computadas usando:
  - M_{01} = número de atributos em que **p** = 0 e **q** = 1
  - M_{10} = número de atributos em que **p** = 1 e **q** = 0
  - M_{00} = número de atributos em que **p** = 0 e **q** = 0
  - M_{11} = número de atributos em que **p** = 1 e **q** = 1

Similaridade com Atributos Binários

- **Coeficiente de Casamento Simples**

$$CCS = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

= 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**

Similaridade com Atributos Binários

- **Coeficiente Jaccard**

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

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

Exemplo

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

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

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

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

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]

Observação

- Pode-se generalizar as similaridades CCS e Jaccard para atributos nominais não binários
  - CCS(**p**, **q**) = M_{AA} / n
    - M_{AA} = no. atributos com o mesmo valor em **p** e **q**
    - n = no. total de atributos
  - Jaccard(**p**, **q**) = (M_{AA} – M_{00}) / (n – M_{00})
    - M_{00} = no. atributos com valor "nulo" em **p** e **q**
      - atributo nominal assimétrico, p. ex.
      - mancha = {ausente, circular, amorfa} se apenas presença importa

Observação

- Exemplo:
  - CCS para comparar 2 pares de seqs. de bases (A, G, C, T)
    - 1º par possui 100 bases cada, 98 iguais ⇒ CCS = 2/100 = 0,02
    - 2º par possui 10 bases cada, 8 iguais ⇒ CCS = 2/10 = 0,2
  - **Nota**: valores são comensuráveis
    - Embora se refiram a sequencias de tamanhos distintos

Similaridade Cosseno

- Para **atributos assimétricos não binários numéricos**
  - 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
  - $\cos(\mathbf{d}_1, \mathbf{d}_2) = (\mathbf{d}_1 \bullet \mathbf{d}_2) / ||\mathbf{d}_1|| \, ||\mathbf{d}_2||$
    - $\bullet$: produto interno entre vetores
    - $||\mathbf{d}||$: é o tamanho (norma) do vetor $\mathbf{d}$
  - Mede o cosseno do ângulo entre os respectivos versores

Exemplo

- 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]

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

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

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

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

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

Exercício

- 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]

Correlação

- Mede interdependência entre vetores numéricos
  - Por exemplo, interdependência linear
- Pode ser portanto usada para medir similaridade
  - entre 2 instâncias descritas por atributos numéricos
  - entre 2 atributos numéricos
- A correlação mais difundida é a de **Pearson**
  - Mede a similaridade entre as **tendências** dos vetores
    - Muito útil em bioinformática
      - magnitudes de seqüências de expressão gênica podem não importar

Correlação de Pearson

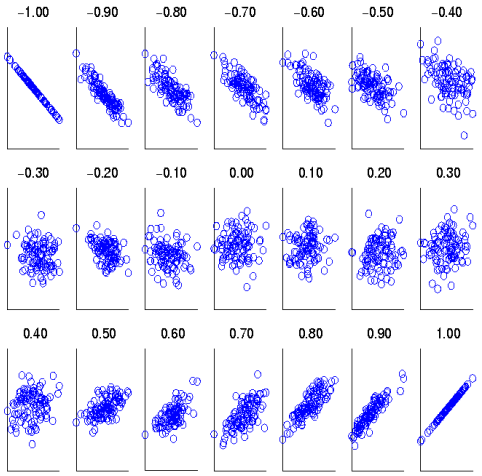
- Cálculo do coeficiente de Pearson:
  - Padronizar vetores **p** e **q**
    - padronização score-z !
  - Calcular produto interno

$$p'_k = (p_k - \mu_{\mathbf{p}}) / \sigma_{\mathbf{p}}$$
$$q'_k = (q_k - \mu_{\mathbf{q}}) / \sigma_{\mathbf{q}}$$
$$\text{correlação}(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}' \bullet \mathbf{q}'}{n}$$

Correlação

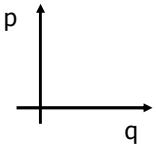
- Valor no intervalo [-1, +1]
  - Correlação (**p**, **q**) = +1
    - Objetos *p* e *q* têm um relacionamento linear positivo perfeito
  - Correlação (**p**, **q**) = –1
    - Objetos *p* e *q* têm um relacionamento linear negativo perfeito
  - Correlação (**p**, **q**) = 0
    - Não existe relacionamento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

Avaliação Visual de Correlação



Scatter plots de um par de instâncias p e q, cada uma com 30 atributos

Similaridade de –1 a 1



Exercício

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

p = [1 -3 0 4 1 0 3]
q = [0 1 4 -2 3 -1 4]

Notas Finais

- Existem outras medidas de similaridade e dissimilaridade além das que vimos nessa aula
- Além disso, existem situações em que as instâncias são descritas por atributos de diferentes tipos e converter todos em um único tipo pode não ser apropriado
  - Nesses casos, existem técnicas para cálculo de (dis)similaridade envolvendo atributos mistos
  - Essas técnicas, no entanto, estão além do escopo deste curso