ETAPAS DE PRÉ-PROCESSAMENTO

DADOS INCONSISTENTES E REDUNDANTES

Cristiane Neri Nobre

Dados inconsistentes

- Dados **inconsistentes** são aqueles que possuem valores **conflitantes em seus atributos**

Em que situação isso ocorre?

- Dados inconsistentes são muitas vezes produzidos no processo de integração de dados (ex: diferentes conjuntos de dados podem usar escalas diferentes para uma mesma medida (metros e centímetros)

- Duas instâncias iguais com classificações diferentes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Doente**
19	F	87	Espalhadas	39,0	6	Doente
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Saudável**

Dados redundantes

- Dados **redundantes** podem se referir tanto a **instâncias** quanto a **atributos**
  - Uma **instância é redundante** quando ela é **muito semelhante** a uma outra instância do mesmo conjunto de dados

Sexo	Esco	Fetária	Renda	Area	GostaA	GostaB
F	1	2	4	5	S	N
F	1	2	4	7	S	S

Dados redundantes

- No caso extremo, instâncias da mesma base de dados podem ser literalmente iguais

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	**F**	**67**	**Inexistentes**	**39,5**	**4**	**Doente**
49	M	92	Espalhadas	38,0	2	Saudável
18	**F**	**67**	**Inexistentes**	**39,5**	**4**	**Doente**
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Dados redundantes

- Um **atributo é redundante** quando seu valor para todas as instâncias pode ser deduzido a partir do valor de um ou mais atributos

Data de Nascimento	Idade
2010	10
2014	6
2007	13

NumVendas	Valor por Venda	VendaTotal
100	6	600
30	4	120
20	80	1600

Dados redundantes

- No caso extremo, quando o atributo possui o mesmo valor que um outro atributo para cada uma das instâncias da base de dados

Idade	Sexo	Peso	Manchas	Temp.	# Int.	# Vis.	Diagnóstico
28	M	79	Concentradas	38,0	2	**2**	Doente
18	F	67	Inexistentes	39,5	4	**4**	Doente
49	M	92	Espalhadas	38,0	2	**2**	Saudável
18	M	43	Inexistentes	38,5	8	**8**	Doente
21	F	52	Uniformes	37,6	1	**1**	Saudável
22	F	72	Inexistentes	38,0	3	**3**	Doente
19	F	87	Espalhadas	39,0	6	**6**	Doente
34	M	67	Uniformes	38,4	2	**2**	Saudável

Int = Número de internações

Int = Número de vezes que o paciente esteve no hospital

Dados redundantes

- A redundância de um atributo está relacionada à sua **correlação** com um ou mais atributos do conjunto de dados.

- Dois ou mais atributos estão correlacionados quando apresentam um perfil de variação semelhante para as diferentes instâncias.

Aluno	Horas de Estudo	Nota
1	20	9,5
2	12	2,5
3	14	3,6
4	15	6,7
5	18	5,2
6	9	1

Corr>0.90

Dados redundantes

- Se a correlação ocorrer entre um atributo de entrada e um atributo de classificação, este atributo de entrada terá uma grande influência na predição do valor do atributo rótulo

Quebrou regra do tipo	Quebrou regra?
A	Sim
B	Sim
Não quebrou regra	Não
C	Sim
Não quebrou regra	Não

Dados redundantes

Qual a causa da redundância?

- Problemas na coleta, na entrada, no armazenamento, na integração ou na transmissão de dados

O fato é que bases reais têm muita redundância e inconsistência!

Dados redundantes

- Investigue no WEKA os filtros que ajudam na eliminação de redundância e inconsistência

Alguns filtros:

- Remove
- RemoveMisclassified

Referências:

- Capítulo 3 do livro (Seções 3.5.2 e 3.5.3)

- Katti Faceli et al. Inteligência Artificial, Uma abordagem de Aprendizado de Máquina, LTC, 2015.