TRANSFORMAÇÃO DE DADOS
CONVERSÃO SIMBÓLICO-NUMÉRICO

Cristiane Neri Nobre

Conversão simbólico-Numérico

- Técnicas como redes neurais artificias, SVM e vários algoritmos de agrupamento lidam apenas com **dados numéricos**

- **E o que podemos fazer quando temos dados simbólicos (nominais/categóricos)?**

  - Precisamos transformar os valores para valores numéricos!

E como podemos fazer isso?

Conversão simbólico-Numérico

Caso 1:

Quando o atributo é do **tipo simbólico** e assume apenas dois valores, se os valores denotam a presença ou ausência de uma característica podemos utilizar **um dígito binário**

Exemplos

Gênero:

Masculino – 0

Feminino – 1

Tumor:

Maligno – 0

Benigno – 1

Conversão simbólico-Numérico

Caso 2:

Quando o atributo é do **tipo simbólico** e assume mais de dois valores, a técnica utilizada na conversão depende de o atributo ser **nominal ou ordinal**.

Se não houver uma relação de ordem entre os valores do atributo, a inexistência de uma relação de ordem deve continuar para os valores numéricos gerados.

Ou seja, a diferença entre quaisquer dois valores numéricos deve ser a mesma.

Uma forma de conseguir isso é codificar cada valor nominal por uma sequência de *c* bits, em que *c* é igual ao número de possíveis valores ou categorias.

Conversão simbólico-Numérico

Na codificação 1-de-c, também denominada canônica ou topológica, cada sequência possui apenas um bit com o valor 1 e os demais com o valor zero.

A diferença entre as sequências é definida pela posição que o valor 1 ocupa nelas.

Para definir a diferença entre dois valores, pode ser utilizada a distância de Hamming.

Conversão simbólico-Numérico

Nesta codificação, cada posição da sequência binária corresponde a um possível valor do atributo nominal.

Por exemplo, se a sequência binária possui 4 bits, o primeiro corresponde ao primeiro valor, o segundo bit ao segundo valor e assim por diante.

Como apenas um dos bits pode assumir o valor 1, o bit que assumir esse valor sinaliza a presença do valor nominal correspondente àquele bit.

Conversão simbólico-Numérico

Exemplo de codificação 1-de-*c*

Atributo nominal	Código $1-de-c$
Azul	100000
Amarelo	010000
Verde	001000
Preto	000100
Marrom	000010
Branco	000001

Observem que são utilizados 6 bits! Não é um número inteiro!

Isso é o chamo de **binarizar o atributo.**

O que significa isso?

Que um único atributo com 6 opções de resposta vão virar 6 entradas na base de dados.

Conversão simbólico-Numérico

É como se você enxergasse:

Azul	Amarelo	Verde	Preto	Marrom	Branco
1	0	0	0	0	0
0	**1**	0	0	0	0
0	0	**1**	0	0	0
0	0	0	**1**	0	0
0	0	0	0	**1**	0
0	0	0	0	0	**1**

Observem que são utilizados 6 bits! Não é um número inteiro!

Isso é o que chamo de **binarizar o atributo.**

O que significa isso?

Que um único atributo com 6 opções de resposta vão virar 6 entradas na base de dados.

Conversão simbólico-Numérico

É como se você enxergasse:

Azul	Amarelo	Verde	Preto	Marrom	Branco
1	0	0	0	0	0
0	**1**	0	0	0	0
0	0	**1**	0	0	0
0	0	0	**1**	0	0
0	0	0	0	**1**	0
0	0	0	0	0	**1**

E veja que a distância de Hamming entre as instâncias é sempre 2.

Conversão simbólico-Numérico

Onde mais podemos utilizar esta codificação?

Em vários outros contextos!

Exemplo:
Onde você nasceu?

- ○ Em casa
- ○ no hospital
- ○ na rua
- ○ na BR

Instância	Em casa	No Hosp	Na rua	Na BR
1	**1**	0	0	0
2	0	**1**	0	0
3	0	0	**1**	0
4	0	0	0	**1**

Conversão simbólico-Numérico

Exemplo:

Qual o seu curso?

- CC
- SI
- Jogos
- ES

Instância	CC	SI	Jogos	ES
1	**1**	0	0	0
2	0	**1**	0	0
3	0	0	**1**	0
4	0	0	0	**1**

Conversão simbólico-Numérico

Veja que isso é interessante apenas para atributos com poucas opções de resposta.

Imagina se você tiver que criar uma codificação desta para representar 193 países?

Isso não seria uma boa opção!

Conversão simbólico-Numérico

Uma alternativa é a representação dos possíveis valores nominais por um conjunto de pseudoatributos.

Os valores dos pseudoatributos podem ser do tipo binário, inteiro ou real

Pseudoatributo e seus possíveis valores

Pseudoatributo	#Valores
Continente	7 (b)
PIB	1 (i)
População	1 (i)
TMA	1 (i)
Área	1 (i)

Conversão simbólico-Numérico

Onde mais podemos ver este tipo de atributos?

1) Qual a cidade onde você mora?
2) Qual o curso você faz (com muitas opções de resposta)?
3) Qual o tipo de droga você já utilizou?

Conversão simbólico-Numérico

Caso 3:

Quando existe uma relação de ordem, o atributo é do tipo ordinal, e a codificação deve preservar essa relação.

Para isso, deve ser utilizada uma codificação em que a ordem dos valores esteja clara.

Conversão simbólico-Numérico

Quando o valor numérico é um número **inteiro ou real**, essa transformação é simples e direta: basta ordenar os valores categóricos ordinais e codificar cada valor de acordo com sua posição na ordem:

Valor ordinal	Valor inteiro
Primeiro	0
Segundo	1
Terceiro	2
Quarto	3
Quinto	4
Sexto	5

Conversão simbólico-Numérico

Se for necessário converter valores ordinais em valores binários, pode ser utilizado o **código cinza** ou o **código termômetro**.

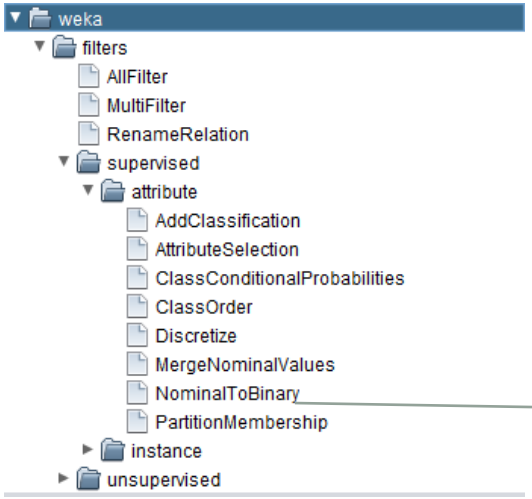
Valor ordinal	Código cinza	Código termômetro
Primeiro	000	00000
Segundo	001	00001
Terceiro	011	00011
Quarto	010	00111
Quinto	110	01111
Sexto	100	11111

Conversão simbólico-Numérico

Como realizar o binarização dos atributos no ambiente WEKA?

- Carregue o arquivo: weather.nominal.arff que fica na pasta **Data** onde o WEKA está instalado
  - Veja que o atributo **Outlook** assume os três possíveis valores: ensolarado, nublado e chuvoso

- Na tela principal do WEKA, vá até a opção **weka/Filters/supervised/atribute/NominalToBinary**



Converte todos os atributos nominais em atributos numéricos binários

Conversão simbólico-Numérico

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Base Original

Atributos transformados

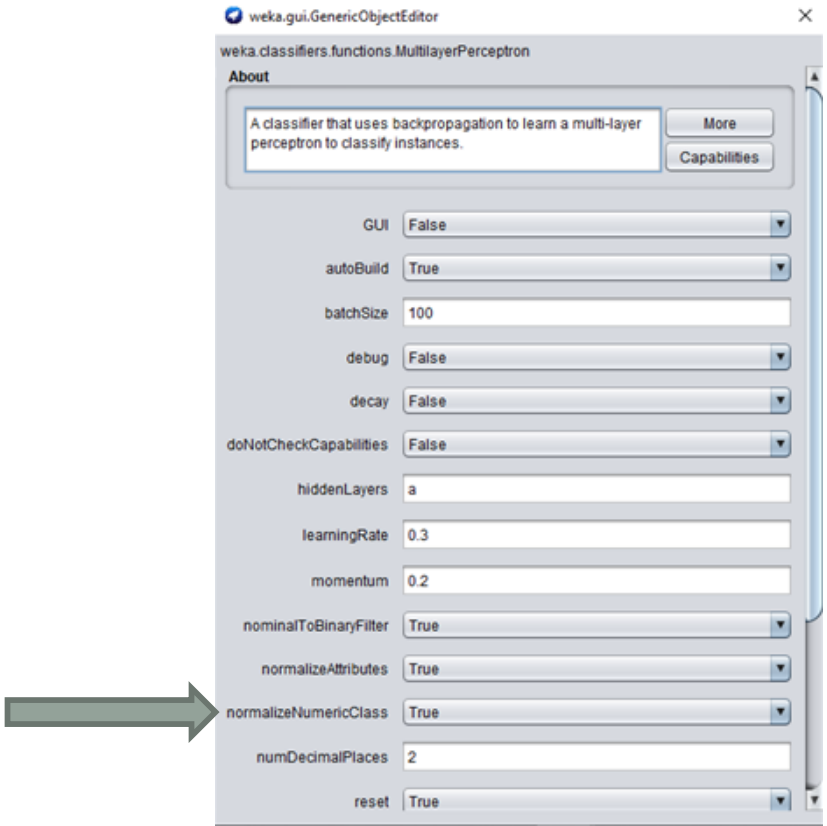
Relation: weather.symbolic-weka.filters.supervised.attribute.NominalToBinary

No.	1: outlook=sunny	2: outlook=overcast	3: outlook=rainy	4: temperature=hot	5: temperature=mild	6: temperature=cool	7: humidity=normal	8: windy=FALSE	9: **play**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	no
2	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	no
3	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	yes
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	yes
5	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	yes
6	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	no
7	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	yes
8	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	no
9	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	yes
10	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	yes
11	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	yes
12	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	yes
13	0.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	yes
14	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	no

Conversão simbólico-Numérico

Mas veja, antes de rodar os modelos, se a implementação da técnica já não faz esta conversão automática

- Rede neural, por exemplo, já faz isso automaticamente!



Referências:

- Capítulo 3 do livro (Seção 3.6.1)

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