

# **TRANSFORMAÇÃO DE DADOS**

## **CONVERSÃO NUMÉRICO-SIMBÓLICA**

---

Cristiane Neri Nobre

# Conversão numérico-simbólica

- Algumas técnicas de AM foram desenvolvidas para trabalhar com **valores qualitativos**.
- Isso ocorre com uma parcela dos algoritmos de **classificação (Ex: ID3) e de associação (Ex: APRIORI)**
- Alguns destes algoritmos que também podem lidar com dados quantitativos têm o seu desempenho reduzido quando o fazem (**Ex: C4.5**)

# Conversão numérico-simbólica

- Nestas situações, a recomendação é discretizar o atributo.
- Quando um atributo quantitativo é discretizado, o conjunto de possíveis valores é dividido em intervalos, e cada intervalo de valores quantitativos é convertido em um valor qualitativo.
- Em alguns métodos, o usuário pode influenciar a definição dos intervalos, definindo valores para parâmetros como número máximo de intervalos.
  - Métodos paramétricos
- Os métodos não paramétricos definem os intervalos utilizando apenas as informações presentes nos valores dos atributos

# Conversão numérico-simbólica

- Os métodos de discretização podem ser **supervisionados** e **não supervisionados**.
- Nos métodos supervisionados é utilizada a informação sobre a **classe dos exemplos**.
- As técnicas supervisionadas geralmente levam a melhores resultados, uma vez que a definição dos intervalos sem conhecimento das classes pode levar à mistura de classes.
- Uma abordagem supervisionada simples seria escolher pontos de corte que maximizam a pureza dos intervalos.
  - Isso pode ser feito utilizando o conceito de entropia

# Conversão numérico-simbólica

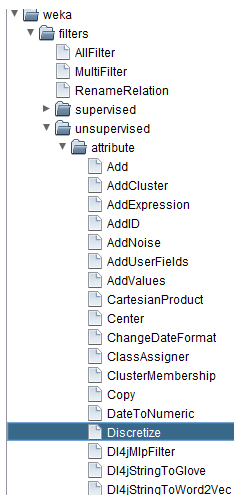
**Algumas das estratégias utilizadas pelos diferentes métodos são apresentados a seguir:**

- **Larguras iguais:** Divide o intervalo original de valores em subintervalos em mesma largura.
  - O desempenho dessa estratégia pode ser afetado pela presença de *outliers*.
- **Frequências iguais:** Atribui o mesmo número de objetos a cada subintervalo.
  - Essa estratégia pode gerar intervalos de tamanhos muito diferentes.
- Uso de um algoritmo de agrupamento de dados
- Inspeção visual

# Conversão numérico-simbólica

Como realizar o discretização dos atributos no ambiente WEKA?

- Carregue o arquivo: weather.numeric.arff que fica na pasta **Data** onde o WEKA está instalado
  - Veja que os atributos **temperatura** e **umidade** assumem valores inteiros
- Na tela principal do WEKA, vá até a opção **weka/Filters/unsupervised/attribute/Discretize**



Discretiza um intervalo de atributos numéricos no conjunto de dados em atributos nominais.

# Conversão numérico-simbólica

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Base Original

Alterando bins=2

Atributos transformados

No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Nominal	Nominal	Nominal	Nominal
1	sunny	'(74.5-inf]'	'(80.5-inf]'	FALSE	no
2	sunny	'(74.5-inf]'	'(80.5-inf]'	TRUE	no
3	overcast	'(74.5-inf]'	'(80.5-inf]'	FALSE	yes
4	rainy	'(-inf-74.5]'	'(80.5-inf]'	FALSE	yes
5	rainy	'(-inf-74.5]'	'(-inf-80....	FALSE	yes
6	rainy	'(-inf-74.5]'	'(-inf-80....	TRUE	no
7	overcast	'(-inf-74.5]'	'(-inf-80....	TRUE	yes
8	sunny	'(-inf-74.5]'	'(80.5-inf]'	FALSE	no
9	sunny	'(-inf-74.5]'	'(-inf-80....	FALSE	yes
10	rainy	'(74.5-inf]'	'(-inf-80....	FALSE	yes
11	sunny	'(74.5-inf]'	'(-inf-80....	TRUE	yes
12	overcast	'(-inf-74.5]'	'(80.5-inf]'	TRUE	yes
13	overcast	'(74.5-inf]'	'(-inf-80....	FALSE	yes
14	rainy	'(-inf-74.5]'	'(80.5-inf]'	TRUE	no

## Referências:

- Capítulo 3 do livro (Seção 3.6.2)
- Katti Faceli et al.  
Inteligência Artificial, Uma abordagem de Aprendizado de Máquina, LTC, 2015.

