INSTITUTO FEDERAL
Santa Catarina

Pré-processamento de Dados

PROF. TIAGO G. MORAES

Roteiro

- Introdução
- Tipos de atributos
- Tipos de ruídos
- Tratamento de dados

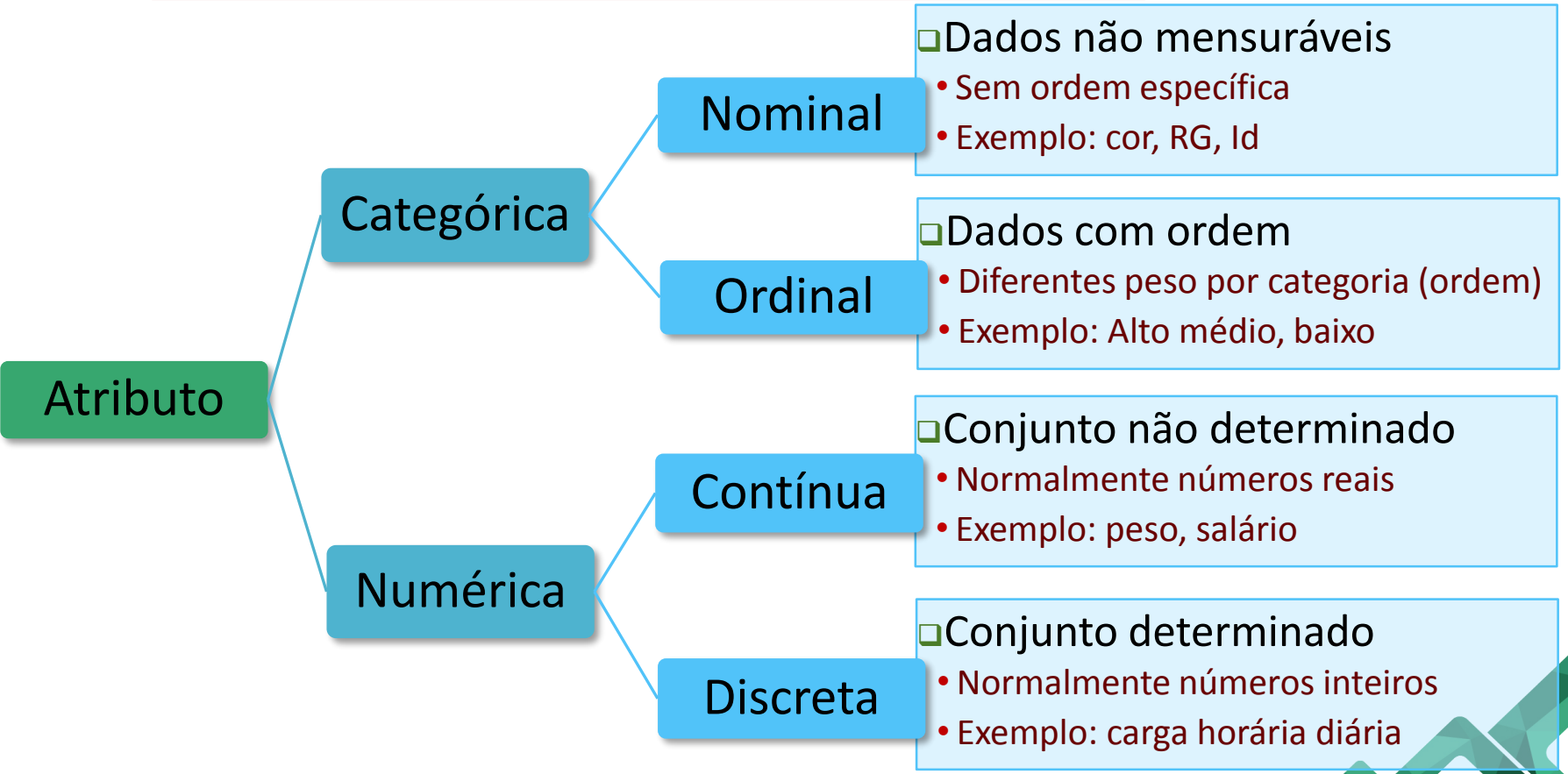
Introdução

- Pré-processamento:
  - Etapa que prepara os dados para um algoritmo de ML (Machine Learning)
  - Alguns modelos:
    - Utilizam tipos de dados específicos
    - Melhoram seus resultados com dados numéricos de mesma grandeza
  - Afeta diretamente a qualidade (percentual de acerto das predições) do modelo criado

Introdução

- É necessário:
  - Entender os tipos de variáveis
  - Entender os tipos de ruído
  - A dependência dos dados
- Tarefas:
  - Interpretação da base
  - Limpeza do ruído(dados faltantes, dados não coerentes)
  - Normalização de dados
  - Mudança de tipos de dados
    - Por exemplo: atributo categórico em numérico
  - Entre outras: Redução de dimensionalidade, obteção de novos atributos

Tipos de atributos



Tipos de atributos

- Classificação
  - atribuir classes para as amostras, baseado em suas características.
- Ruído
  - distorção, falha ou imprecisão que ocorre na aquisição dos dados.
- Classificadores
  - Usados para classificar ou descrever padrões ou objetos a partir de um conjunto de propriedades ou características.

Tipos de atributos

- Exemplo: dados de clientes que solicitaram empréstimo

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	R$12302,25	25	R$20000,00	A	Não
2	R$82002,22	60	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Tipos de atributos

- Exemplo: dados de clientes que solicitaram empréstimo

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	R$12302,25	25	R$20000,00	A	Não
2	R$82002,22	60	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Classe de interesse

Tipos de atributos

- Exemplo: dados de clientes que solicitaram empréstimo

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	R$12302,25	25	R$20000,00	A	Não
2	R$82002,22	60	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim
Cat. Nominal	Num. Contínua	Num. Discreta	Num. Contínua	Cat. Ordinal	Cat. Nominal

Tipos de ruídos

- Valores faltantes
  - Valores nulos, vazios, não preenchidos
- Valores inconsistentes
  - Fogem da realidade: salário negativo
- Variáveis independentes
  - Não causam impacto na classificação desejada
  - Exemplo: Time que torce
- Variáveis com valores discrepantes
  - Dependendo do classificador, algumas variáveis podem influenciar demais na classificação dado a sua grandeza numérica
  - Solução: normalizar, padronizar

Tipos de ruídos

- Exemplo:

Ruído: valor faltante

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	-	25	R$20000,00	A	Não
2	R$82002,22	-60	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Ruído: variável independente

Ruído: valor inconsistente

Tipos de ruídos

- Exemplo:

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	-	25	R$20000,00	A	Não
2	R$82002,22	-60	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Ruído: valores discrepantes

Tratamento de dados

- Solução para valores Faltantes e inconsistentes

  - Eliminar problema:
    - Excluir a coluna (atributo inteiro)
      - utilizar somente se a coluna tem muitos valores inconsistentes
    - Excluir a linha (instância)

  - Substituir por outro valor:
    - Descobrir o valor → com a fonte dos dados (por ser díficil)
    - Substituir por um valor aproximado:
      - Por exemplo: usar a média dos valores do atributo

Tratamento de dados

- Possível Solução
  - Eliminar coluna id
  - Substituir valor de salário nulo e idade negativa

ID	Salário anual	Idade	Valor solicitado	Nota Bom Pagador	Aprovação
1	**R$79935,57**	25	R$20000,00	A	Não
2	R$82002,22	**34**(-60)	R$100000,00	B	Não
3	R$115502,25	45	R$50000,00	A	Sim
4	R$42302,25	32	R$20000,00	A+	Sim

Tratamento de dados

- Solução para Variáveis com valores discrepantes

  - Padronização (melhor, evita problema de *outliers*)

$$X = \frac{x - \text{média(coluna x)}}{\text{desvio padrão (coluna x)}}$$

  - Normalização

$$X = \frac{x - \text{mínimo(coluna x)}}{\text{máximo(coluna x)} - \text{mínimo(coluna x)}}$$

Tratamento de dados

- Usando padronização nas colunas de interesse:

Salário anual	Idade	Valor solicitado
R$79935,57	25	R$20000,00
R$82002,22	34	R$100000,00
R$115502,25	45	R$50000,00
R$42302,25	32	R$20000,00

Salário anual	Idade	Valor solicitado
0	-1,086	-0,729
0,069	0	1,391
1,189	1,327	0,066
-1,258	-0,241	-0,729

Tratamento de dados

- Mudança do tipo de atributos
  - Muitas vezes os modelos necessitam de atributos somente numéricos
  - Mudança: atributo categórico para numérico
- Categórica ordinal:
  - Como existe uma ordem nas categorias, pode-se substituir por números discretos
  - Exemplo: baixo, médio e baixo ➔ 1, 2, 3

Altura
Alto
Muito Alto
Baixo
Baixo

Altura
4
5
2
2

Tratamento de dados

- Mudança do tipo de atributos
  - Muitas vezes os modelos necessitam de atributos somente numéricos
  - Mudança: atributo categórico para numérico
- Categórica nominal:
  - cria-se uma nova categoria para cada possível valor
  - Se preenche com 0's e 1's

Sexo
Masculino
Feminino
Feminino
Feminino

Masculino	Feminino
1	0
0	1
0	1
0	1