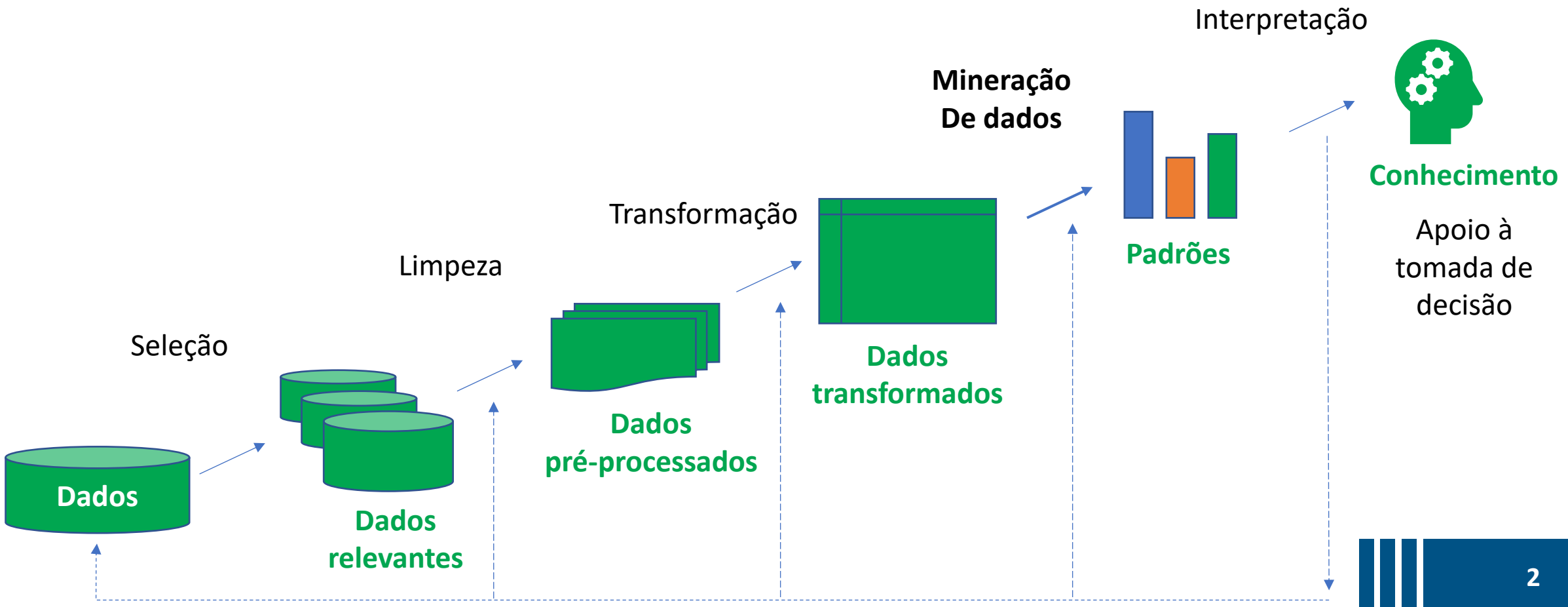
Métricas de avaliação

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

O processo de KDD

"É um processo de várias etapas, não trivial, **interativo** e **iterativo**, para a identificação de **padrões** válidos, novos e potencialmente úteis a partir de um grande conjunto de dados" (FAYYAD, 1996).



1 Seleção

2 Limpeza

3 Transformação

4 **Mineração**

5 Interpretação

- ✓ Selecionar a(s) **tarefa(s)** de mineração de acordo com o problema levantado dentro do projeto de análise de dados:
  - ✓ Associação
  - ✓ Classificação
  - ✓ Agrupamento
  - ✓ **Predição**
  - ✓ ....
- ✓ Escolha do algoritmo dentro da tarefa adotada.
- ✓ Aplicação do algoritmo para construir o modelo.

Introdução

- A predição se encaixa no **aprendizado de máquina supervisionado**
  - Os dados para treinar os algoritmos **precisam** de rótulo (etiqueta, label).
- Os algoritmos podem ser entendidos como funções que, dado um conjunto de **dados rotulados**, constroem um estimador.
  - Entrada: atributos
  - Saída (atributo alvo): classe ou valor
- Predição cujo atributo alvo é do tipo nominal
  - Classificação
- Predição cujo atributo alvo é numérico
  - Regressão

Introdução

- ✓ Exemplo de predição envolvendo um **problema de classificação**
  - ✓ Conjunto de dados referente à planta Iris
  - ✓ Pétalas (P) e Sépalas (S)

Atributo alvo (Classe)

Atributos de entrada

Tamanho (P)	Largura (P)	Tamanho (S)	Largura (S)	Espécie
5,1	3,5	1,4	0,2	Setosa
4,9	3,0	1,4	0,2	Setosa
7,0	3,2	4,7	1,4	Versicolor
6,4	3,2	4,5	1,5	Versicolor
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
6,3	2,3	4,8	1,7	??????

Versicolor?
Virginica?
Setosa?

Introdução

- ✓ Exemplo de predição envolvendo um **problema de regressão**
  - ✓ Conjunto de dados populacional de um país

Atributo alvo
Número

Atributos de entrada

Fertilidade	Agricultura	Educação	Renda	Mortalidade
80,2	17,0	12	9,9	22,2
83,1	45,1	9	84,8	22,2
92,5	39,7	5	93,4	20,2
85,8	36,5	7	33,7	20,3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
87,8	34,5	4	32,7	???

Introdução

- Definição formal da tarefa de predição:

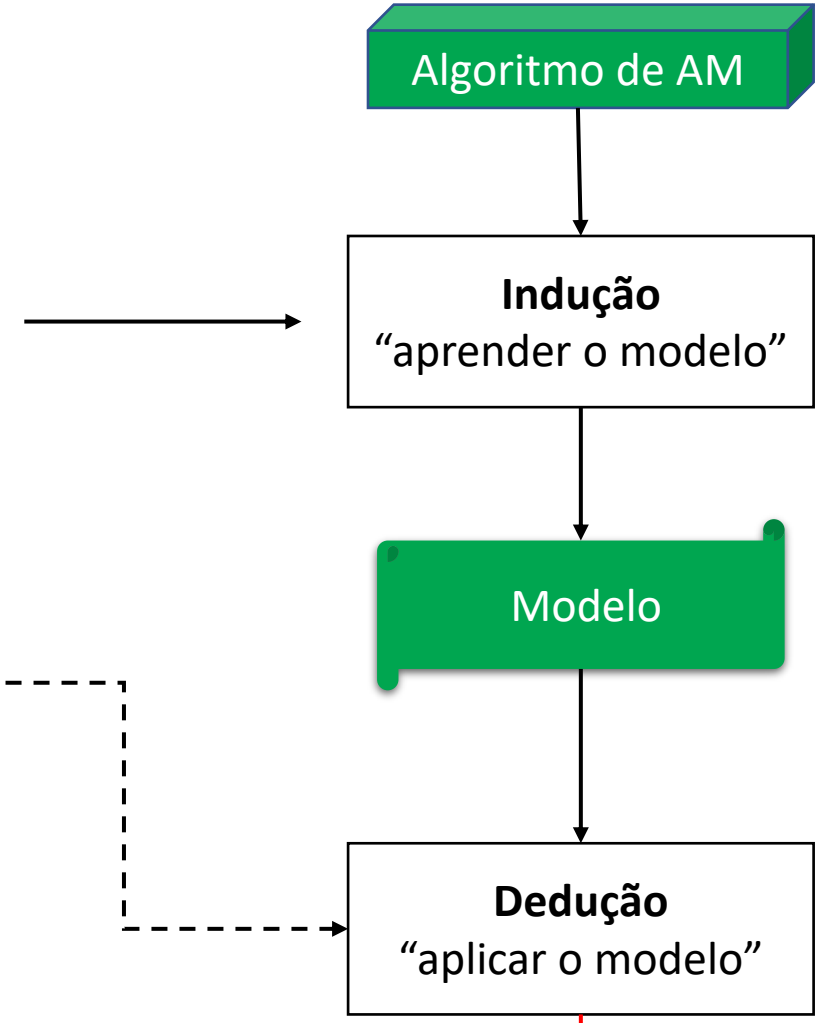
- A partir de um **conjunto de treinamento**, cada objeto de dados é caracterizado por uma tupla (x, y) na qual x é o conjunto de atributos e y é o atributo alvo (classe)

  - **X:** atributo, preditor, variável independente, entrada
  - **Y:** classe, resposta, variável dependente, saída

- Aprender um modelo que mapeia cada conjunto de atributo x em uma das classes.

Conjunto de treinamento

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Não
Não	Solteiro(a)	70K	Sim
Sim	Casado(a)	120K	Não
Não	Divorciado(a)	95K	Sim



Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Não	Casado(a)	55k	?
Sim	Divorciado(a)	80k	?
Sim	Solteiro(a)	110k	?
Não	Solteiro(a)	95k	?
Não	Casado(a)	67k	?

Conjunto de teste

Quão assertivo é o meu modelo em prever as classes?

Avaliação de modelos preditivos

- ✓ Motivação: não é possível estabelecer a priori qual algoritmo de AM se sairá melhor na resolução de qualquer tipo de problema.

- ✓ Necessidade de experimentação
  - ✓ Diversos algoritmos podem ser utilizados para a **indução** de um modelo.
  - ✓ Um único algoritmo pode ser parametrizado de formas diferentes, levando à obtenção de múltiplos modelos para **o mesmo conjunto de dados**.

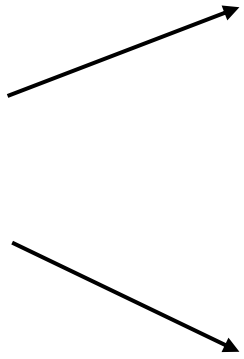
- ✓ Experimentos controlados, seguindo procedimentos que garantam
  - ✓ Corretude
  - ✓ Validade
  - ✓ Reprodutibilidade do experimento
  - ✓ Reprodutibilidade das conclusões obtidas

Avaliação de modelos preditivos

A avaliação de desempenho de um algoritmo de AM é realizada por meio da avaliação do desempenho do preditor gerado por ele na **rotulação** de novo objetos, não apresentados previamente no treinamento.

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Não
Não	Solteiro(a)	70K	Sim
Sim	Casado(a)	120K	Não
Não	Divorciado(a)	95K	Sim



Conjunto de treinamento

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Não
Não	Solteiro(a)	70K	Sim

Conjunto de testes

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Casado(a)	120K	?
Não	Divorciado(a)	95K	?

Entenda **rotulação** como valor categórico (classe) ou contínuo – Classificação ou regressão respectivamente

Avaliação de modelos preditivos

- ✓ Métricas de erro
  - ✓ Classificação
  - ✓ Regressão

Métricas de avaliação para a Classificação

Métricas

✓Classificação

- ✓Dado um conjunto de dados contendo N objetos, sobre o qual a avaliação será realizada, a **taxa de erro** equivale à proporção de exemplos desse conjunto classificados incorretamente pelo algoritmo e é obtida pela **comparação** da **classe conhecida** x_i com a **classe predita** y_i .

$$erro = \frac{1}{n}\sum_{i=1}^{n}(x_i \neq y_i)$$

- ✓Valores entre 0 (menor erro) e 1 (maior erro)
- ✓O complemento dessa taxa corresponde à taxa de acerto ou acurácia do classificador

$$acc = 1 - erro$$

- ✓Valores próximos de 1 são melhores e próximos de 0 são piores.

Métricas

✓Classificação

✓Exemplo 1: Qual é o erro considerando o conjunto de testes abaixo?

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Não
Não	Solteiro(a)	70K	Sim
Sim	Casado(a)	120K	**Não**
Não	Divorciado(a)	95K	**Sim**

Conjunto de testes

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente	Valor real da classe x_i	Valor predito para a classe y_i
Sim	Casado(a)	120K	?	**Não**	Sim
Não	Divorciado(a)	95K	?	**Sim**	Não

$$erro = \frac{1}{n}\sum_{i=1}^{n}(x_i \neq y_i)$$

$$erro = \frac{1}{2} \times 2$$

$$erro = 1$$

$$erro = 100\%$$

$$acc = 0\%$$

Métricas

✓Classificação

✓Exemplo 2: Qual é o erro considerando o conjunto de testes abaixo?

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Não
Não	Solteiro(a)	70K	Sim
Sim	Casado(a)	120K	**Não**
Não	Divorciado(a)	95K	**Sim**

Conjunto de testes

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente	Valor real da classe x_i	Valor predito para a classe y_i
Sim	Casado(a)	120K	?	**Não**	Não
Não	Divorciado(a)	95K	?	**Sim**	Não

$$erro = \frac{1}{n}\sum_{i=1}^{n}(x_i \neq y_i)$$

$$erro = \frac{1}{2} \times 1$$

$$erro = 0{,}5$$

$$erro = 50\%$$

$$acc = 50\%$$

Métricas

Classificação

- Outra alternativa para visualizar o desempenho de um classificador é utilizar uma **matriz de confusão ou tabela de contingência**. Ela ajuda a visualizar o número de predições corretas e incorretas em cada classe. O número K de classes representa a dimensão da matriz, ou seja, teremos sempre uma matriz quadrada: K x K.
  - Linhas: representam as classes verdadeiras $\boldsymbol{x_i}$
  - Colunas: representam as classes preditas $\boldsymbol{y_i}$

- A análise da matriz nos permite calcular uma série de outras medidas de quais classes os algoritmo apresenta maior dificuldade de predição.

- Para exemplificar, vamos assumir a existência de duas classes, uma positiva (+) e outra negativa (-). Logo, teremos uma matriz 2 x 2.

Métricas

- **Classificação**
  - Matriz de confusão =

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

  - Verdadeiros Positivos (**VP**): números de objetos da classe + classificados corretamente.
  - Verdadeiros Negativos (**VN**): número de objetos da classe – classificados corretamente.
  - Falsos Positivos (**FP**): número de objetos da classe negativa que foram classificados incorretamente como sendo da classe positiva.
  - Falsos Negativos (**FN**): número de objetos da classe positiva que foram classificados incorretamente como sendo da classe negativa.
  - Total de exemplos: N = VP + VN + FP + FN

Métricas

- **Classificação**
  - **Taxa de erro na classe positiva (+):** proporção de objetos da classe positiva classificados incorretamente pelo preditor. Também conhecida como taxa de falsos negativos (TFN).

$$TFN = \frac{FN}{VP + FN}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Taxa de erro na classe negativa (-):** proporção de objetos da classe negativa classificados incorretamente pelo preditor. Também conhecida como taxa de falsos positivos (TFP).

$$TFP = \frac{FP}{FP + VN}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Taxa de erro total:** soma dos valores da diagonal secundária dividida pelo total de elementos da matriz.

$$erro = \frac{FN + FP}{N}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Taxa de acerto total ou acurácia:** soma dos valores da diagonal principal dividida quantidade de valores total da matriz.

$$acc = \frac{VP + VN}{N}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Precisão:** proporção de objetos positivos classificados corretamente entre todos aqueles preditos como positivos pelo preditor.

$$precision = \frac{VP}{VP + FP}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Revocação ou sensibilidade:** taxa de acerto na classe positiva. Também é conhecida como taxa de verdadeiros positivos (TVP).

$$TVP = \ recall = sens = \frac{VP}{VP + FN}$$

	Classe +	Classe -
Classe +	**VP**	**FN**
Classe -	**FP**	**VN**

Métricas

- **Classificação**
  - **Precisão e revocação**: considerações
  - A precisão pode ser vista como uma medida de exatidão do modelo.
  - A revocação pode ser vista como uma medida de completude à precisão.
  - Por isso, essas medidas são combinadas em uma nova medida: a medida-F
    - Média harmônica ponderada da precisão e da revocação

$$medidaF = \frac{(w+1) \times recall \times precision}{rev + w \times precision}$$

    - Medida F1 considera que precisão e revocação tem a mesma importância

$$medidaF_1 = \frac{2 \times precision \times recall}{rev + precision}$$

Métricas

- **Classificação**
  - E como ficam os VP, FP, VN, FN para casos com mais de duas classes? Como fazer os cálculos das medidas?
  - Os cálculos são feitos classe a classe. Por exemplo, para calcular a precisão, consideramos a classe em questão como positiva enquanto as demais são vistas como negativas.

Métricas de avaliação para a Regressão

Métricas para regressão

- O erro é calculado a partir da distância entre o valor de $\boldsymbol{x_i}$ (valor conhecido) e do valor predito $\boldsymbol{y_i}$ para o atributo alvo. As medidas de erro para problemas de regressão mais conhecidas são:
  - Erro quadrático médio (Mean squared error – MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

  - Distância absoluta média (Mean absolute error – MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

  - Seus valores são sempre não negativos. Valores mais baixos correspondem a melhores modelos.

Métricas para regressão

Exemplo 1: Qual é o MAD considerando o conjunto de testes abaixo?

Conjunto de dados original

Valor imóvel	Valor do automóvel	Renda anual	Poder de compra
499k	50k	125K	4
100k	20k	100K	5,1
200k	30k	70K	7
345k	35k	120K	**4,7**
500k	80k	95K	**3**

Conjunto de testes

Valor imóvel	Valor do automóvel	Renda anual	Pode de compra	Valor real da classe x_i	Valor predito para a classe y_i
345k	35k	120K	?	**4,7**	3
500k	80k	95K	?	**3**	3

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |x_i - y_i|$$

$$MAD = \frac{1}{2}(4{,}7 - 3) + (3 - 3)$$

$$MAD = \frac{1}{2}(4{,}7 - 3) + (3 - 3)$$

$$MAD \cong 1{,}41$$

Dúvidas?

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br