Tópicos Avançados em IA
Aula #5

Definição do Projeto

Professor: Eduardo R. Hruschka
Estagiário PAE: Luiz F. S. Coletta
(luizfsc@icmc.usp.br)

Sumário

- **Definição do projeto**
  1. Desenvolvimento de algoritmo de Aprendizado de Máquina (AM);
  2. Pré-processamento dos dados;
  3. Calibrando/avaliando algoritmos de AM;
  4. Softwares para mineração de dados.

Definição do Projeto

- **Alunos se organizarão em duplas**

1. Cada dupla desenvolverá um algoritmo de Aprendizado de Máquina (AM).

2. Apresentarão um seminário sobre o projeto (em torno de 12-15 minutos).

Definição do Projeto

- **Desenvolvimento de algoritmo de AM**
  - Disponibilização de base de dados para a mineração
    - Problema de classificação:

	A1	*A2*	*A3*	…	*Am*	*Classe*
X1	*2*	*36*	*2*	…	*48*	*1*
X2	*4*	*18*	*?*	…	*18*	*2*
X3	*1*	*18*	*4*	…	*21*	*2*
X4	*?*	*6*	*4*	…	*10*	*1*
X5	*3*	*6*	*2*	…	*7*	*1*
X6	*1*	*24*	*?*	…	*24*	*2*
X7	*?*	*24*	*?*	…	*15*	*1*
X8	*4*	*18*	*4*	…	*28*	*2*
⋮	⋮	⋮	⋮	⋱	⋮	⋮
Xn	*2*	*18*	*4*	…	*13*	*1*

Definição do Projeto

- **Desenvolvimento de algoritmo de AM**
  - Muitos algoritmos disponíveis para uso
    - Averiguar a caixa de ferramentas!

      - Naive Bayes;
      - Árvore de Decisão;
      - K-NN;
      - Ensembles (agregação de diferentes modelos - supervisionados e não-supervisionados).

Definição do Projeto

- **Pré-processamento dos dados**
  - Dados geralmente nunca estão organizados apropriadamente
    - Valores ausentes/faltantes:

	A1	*A2*	*A3*	…	*Am*	*Classe*
X1	*2*	*36*	*2*	…	*48*	*1*
X2	*4*	*18*	*?*	…	*18*	*2*
X3	*1*	*18*	*4*	…	*21*	*2*
X4	*?*	*6*	*4*	…	*10*	*1*
X5	*3*	*6*	*2*	…	*7*	*1*
X6	*1*	*24*	*?*	…	*24*	*2*
X7	*?*	*24*	*?*	…	*15*	*1*
X8	*4*	*18*	*4*	…	*28*	*2*
⋮	⋮	⋮	⋮	⋱	⋮	⋮
Xn	*2*	*18*	*4*	…	*13*	*1*

Definição do Projeto

- **Pré-processamento dos dados**
  - Valores ausentes;
  - Seleção de atributos (redução de dimensionalidade);
  - Re-escalar/padronizar atributos (normalização).

Definição do Projeto

- **Valores ausentes**

I. Descartar instâncias/atributos

- Interessante se instâncias/atributos não são fundamentais para o problema.

II. Imputar valores a partir de alguma estratégia

- Média/mediana por classe;
- K-NN considerando os outros atributos.

Definição do Projeto

- **Seleção de atributos**
  - Remoção de atributos irrelevantes ou redundantes
    - Filtro;
    - Wrapper;
    - Embedded.

Definição do Projeto

- **Re-escalar/padronizar atributos**
  - Atributos de maior magnitude podem ter maior peso;

I. **Re-escalar** [0,1]

$$l_{ij} = \frac{x_{ij} - \min(\boldsymbol{x}_j)}{\max(\boldsymbol{x}_j) - \min(\boldsymbol{x}_j)}$$

II. **Score-z** (produz distribuição normal padrão N(0,1))

$$z_{ij} = \frac{x_{ij} - \boldsymbol{\mu}_{\boldsymbol{x}_j}}{\boldsymbol{\sigma}_{\boldsymbol{x}_j}}$$

Definição do Projeto

- **Re-escalar/padronizar atributos**
  - Atributos de maior magnitude podem ter maior peso;

I. **Re-escalar** [0,1]

$$l_{ij} = \frac{x_{ij} - \min(\boldsymbol{x}_j)}{\max(\boldsymbol{x}_j) - \min(\boldsymbol{x}_j)}$$

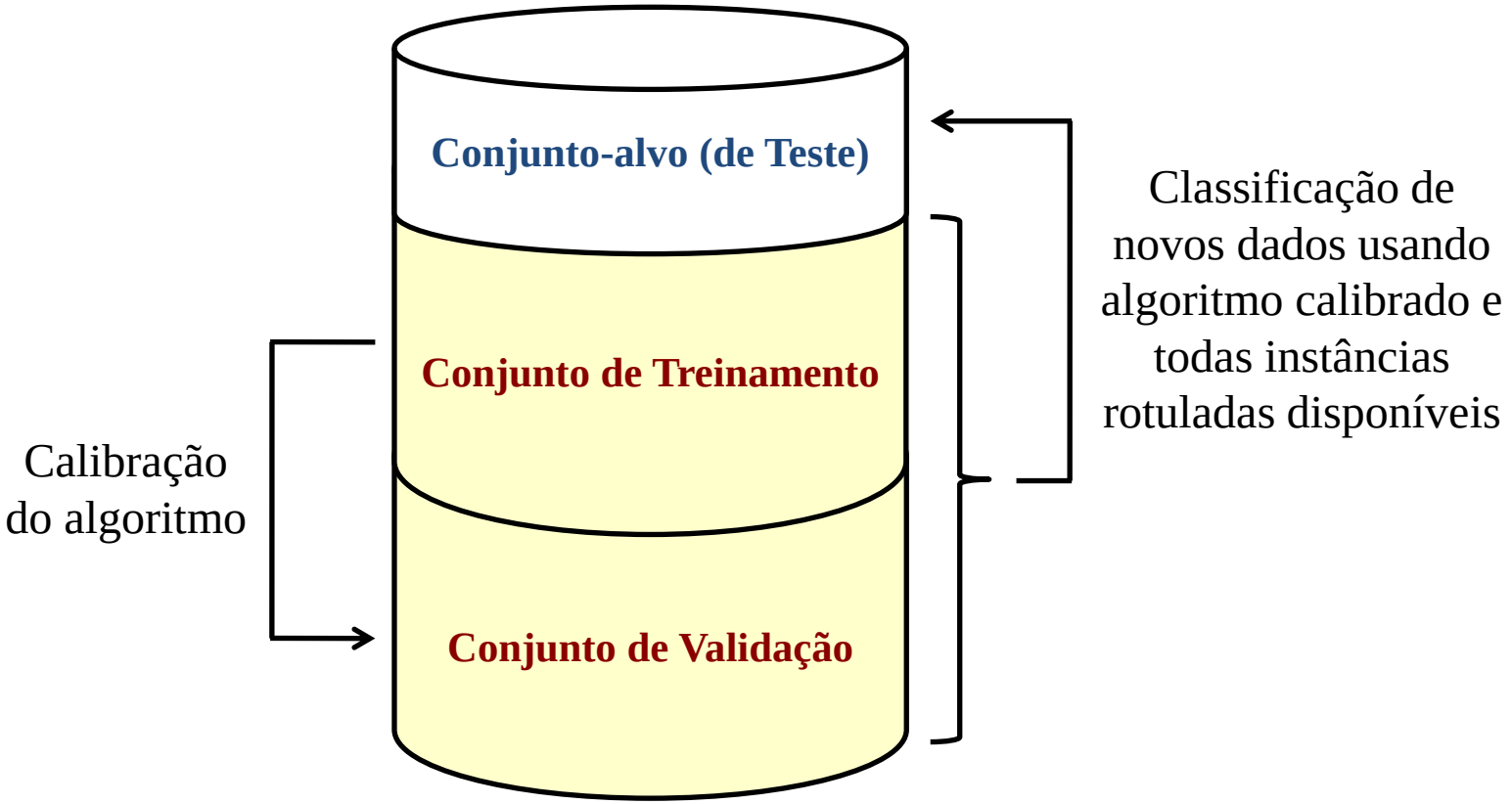
II. **Score-z** (produz distribuição normal padrão N(0,1))

$$z_{ij} = \frac{x_{ij} - \boldsymbol{\mu}_{\boldsymbol{x}_j}}{\boldsymbol{\sigma}_{\boldsymbol{x}_j}}$$

Atenção: pode distorcer a natureza do problema!

Definição do Projeto

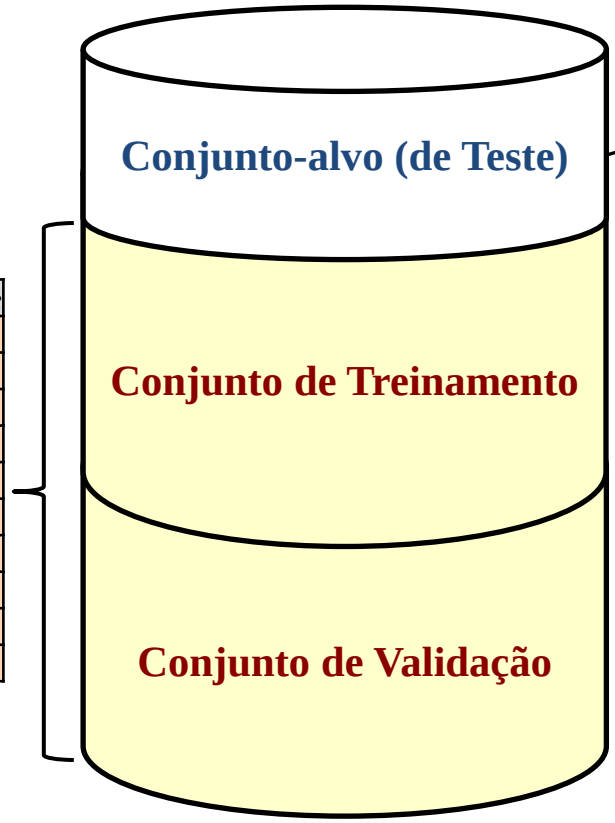
- **Calibrando/avaliando algoritmos de AM**



Definição do Projeto

- **Calibrando/avaliando algoritmos de AM**

	A1	A2	A3	…	Am	Classe
X1	2	36	2	...	48	1
X2	4	18	?	...	18	2
X3	1	18	4	...	21	2
X4	?	6	4	...	10	1
X5	3	6	2	...	7	1
X6	1	24	?	...	24	2
X7	?	24	?	...	15	1
X8	4	18	4	...	28	2
⋮	⋮	⋮	⋮	⋱	⋮	⋮
Xn	2	18	4	...	13	1



	A1	A2	A3	…	Am	Classe
Xn+1	1	18	2	...	21	?
Xn+2	2	36	4	...	30	?
Xn+3	4	6	2	...	11	?
Xn+4	1	18	1	...	9	?
Xn+5	1	18	2	...	3	?
Xn+6	3	6	3	...	25	?
Xn+7	3	24	4	...	13	?
Xn+8	2	24	1	...	33	?
⋮	⋮	⋮	⋮	⋱	⋮	⋮
Xn+p	4	36	2	...	17	?

Definição do Projeto

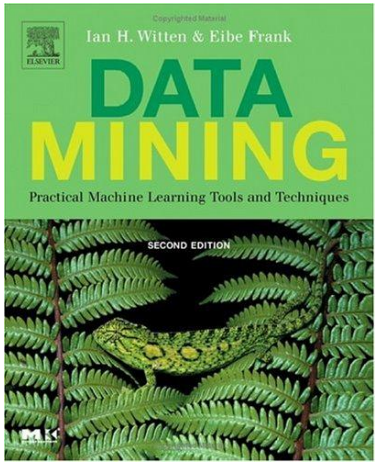
- **Calibrando/avaliando algoritmos de AM**
  - Confiabilidade dos resultados
    - Cross-validation;
    - Leave-one-out;
    - Bootstrap.

Matriz de confusão para classificação binária

		Classes preditas	
		Sim	*Não*
Classes reais	*Sim*	**Verdadeiro positivo**	**Falso Negativo**
	Não	**Falso positivo**	**Verdadeiro negativo**

Definição do Projeto

- **Calibrando/avaliando algoritmos de AM**
  - Leitura recomendada
    - Capítulo 5 em Witten & Frank (2005)



Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005

Definição do Projeto

- **Softwares para mineração de dados**
  - WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
    - Java (open source);
    - Pré-processamento, classificação, regressão, clustering, regras de associação e visualização.

  - Listagem no KDnuggets
    - http://www.kdnuggets.com/software/index.html.

Definição do Projeto

- **Desenvolvimento do trabalho**
  - Usar conjunto de treinamento disponível;
  - Permitir a predição da classe de instâncias provindas de um conjunto de teste:
    - Gravar em arquivo os rótulos de classes das instâncias deste conjunto de teste (respeitando a ordem).

Definição do Projeto

- **Desenvolvimento do trabalho**

Aula	Data	Assunto
8	04/10	Correção da Prova / Desenvolvimento do Projeto
10	18/10	Desenvolvimento do Projeto
11	25/10	Desenvolvimento do Projeto
12	01/11	Desenvolvimento do Projeto
13	08/11	Desenvolvimento do Projeto / Liberação Arquivo de Teste
14	22/11	Apresentação dos Projetos

Atendimento/Orientação

- **Estagiário PAE**
  - Luiz F. S. Coletta
    - Agendamento: luizfsc@icmc.usp.br

- **Créditos**
  - Material inspirado em notas de aulas dos Profs. Eduardo Hruschka e Ricardo Campello.