Inteligência Artificial

Aula 24- Aprendizagem de Máquina: Pré-processamento [1]

Sílvia M.W. Moraes

Faculdade de Informática - PUCRS

October 19, 2017

[1] Este material não pode ser reproduzido ou utilizado de forma parcial sem a permissão dos autores.

Sinopse

- Nesta aula, introduzimos em **aprendizagem de máquina.**
- Este material foi construído com base no material sobre Data Mining dos professores Rodrigos Barros, Duncan e Renata de Paris e também nos capítulos:
  - 1,2 e 3 - Inteligência Artificial: Uma abordagem de Aprendizagem de Máquina: Facelli e outros.
  - 10 do livro Inteligência Artificial: Luger
  - 18 do livro Artificial Intelligence – a Modern Approach: Russel & Norvig

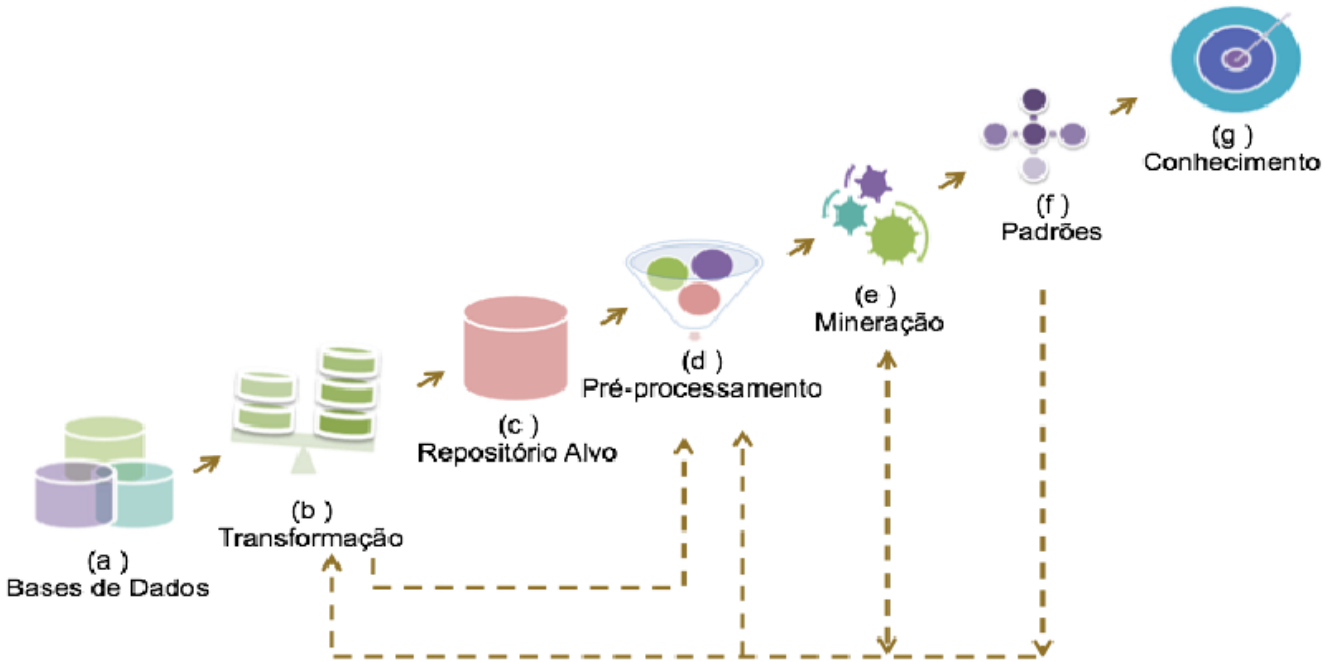
Sumário

Aulas anteriores

- Agente Reativos e Cognitivos
- Solução de Problemas: Algoritmos de busca
- Planejamento Clássico
- Introdução à Raciocínio Probabilistico
- Introdução a Aprendizagem de Máquina

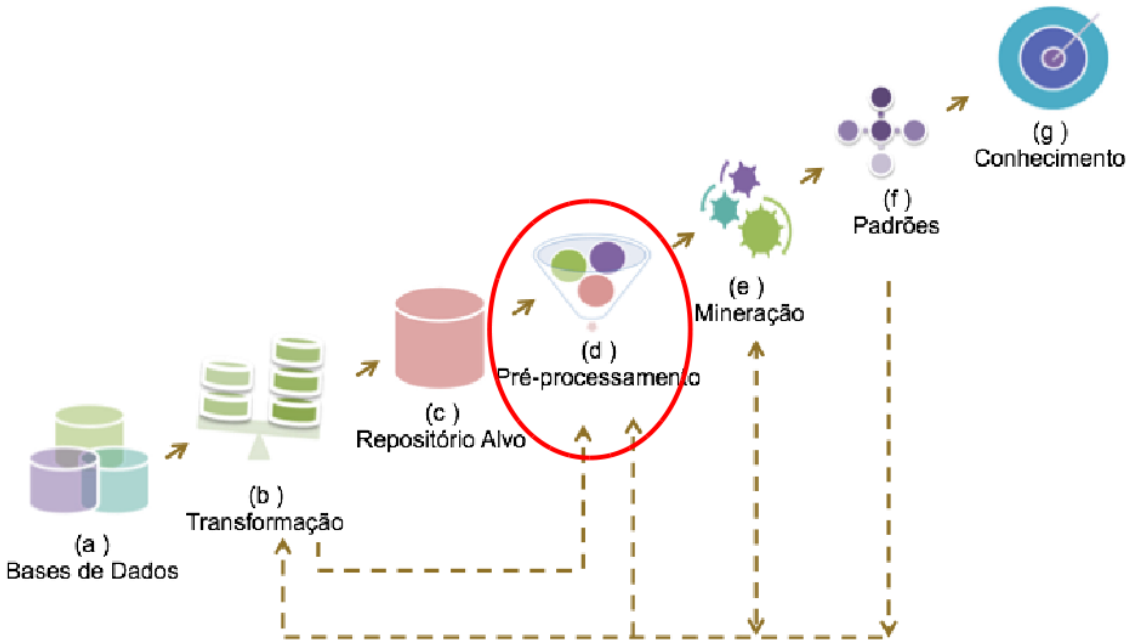
Processo de Descoberta de Conhecimento

- **Knowledge Discovery in Databases (KDD)**: consiste em uma série de passos bem definida cuja meta é transformar dados em conhecimento.



Processo de Descoberta de Conhecimento

- **Knowledge Discovery in Databases (KDD):**
  (d) Pré-processamento :
  - Etapa de **ajuste fino dos dados** para atender ao objetivo da aprendizagem de máquina
  - **~85% de todo o processo** (trabalhosa, mas valiosa)



Pré-processamento dos Dados

- É fundamental para a **qualidade dos resultados**
- Inclui usualmente:
  - **Limpeza de dados**
  - **Transformação dos dados**
  - **Redução de dimensionalidade**

Limpeza dos Dados

- Muitos **dados do mundo real** são potencialmente **incorretos** (falha no instrumento de leitura, erro humano ou de máquina, erro de transmissão). Os dados podem ser:
  - **Incompletos**: falta de valores de atributos, falta de certos atributos de interesse ou contendo apenas dados agregados. Por exemplo, Ocupação = "" (dados em falta) .
  - **Ruidosos**: contendo ruído, erros ou outliers. Por exemplo, Salário = "- 10" (um erro)
  - **Inconsistentes:** contendo discrepâncias em códigos ou nomes. Exemplos: Idade = "42" e Aniversário = "03/07/2010"; ora a é classificação "1, 2, 3" ora é "A, B, C"; "01 de janeiro" como o aniversário de todos? CEP de todos 90000-000?

Limpeza dos Dados: Dados Incompletos

- Como **lidar com dados faltantes**?
  - **Ignorar a tupla:** geralmente feito quando o rótulo da classe está faltando (ao fazer a classificação) ou o atributo é irrelevante
    - não é eficaz quando o % de valores em falta por atributo varia consideravelmente
  - **Preencher manualmente:** tedioso ? inviável?
  - **Preencher automaticamente:** (uso de alguma heurística é usual)
    - uma constante global: por exemplo, "desconhecido"
    - média : a média ou mediana do atributo para todas as amostras pertencentes à mesma classe ou moda, em caso de valor simbólico (uma boa opção)
    - valor mais provável: baseado em inferência (uso de uma fórmula bayesiana ou árvore de decisão)

Limpeza dos Dados: Dados Ruidosos

- Como **lidar com ruídos?**
  - **Encestamento**
    1. Classificar os dados e organizá-los em cestas ou faixas (de frequência igual)
    2. Suavizar o ruído, substituindo os valores pela média ou mediana dos valores pertencentes à mesma faixa de valor.
  - **Agrupamento:** detectar e remover outliers (atributos que não formarem grupos)
  - **Regressão:** Ajustando os dados por meio de funções de regressão e por classificação, no caso de dados simbólicos.
  - **Distância:** técnicas baseadas em distância verificam a que classe pertencem objetos mais próximos de cada objeto x. Se x for de outra classe, ele pode ser um ruído. Borderlines devem ser eliminados.

Limpeza dos Dados: Dados Inconsistentes

- Podem ser **resultantes do processo de integração de bases**.
  - escalas diferentes para uma mesma medida (m, cm)
  - codificação diferente para representar um atributo relacionado a tamanho (pequeno e grande; médio e enorme).

Limpeza dos Dados: Dados Inconsistentes

- Como **lidar com inconsistências?**
  - Podem ser identificados pelo cálculo de **correlação** (mede o quanto duas variáveis tendem a mudar juntas) e análise de **covariância** (mede a relação linear entre duas variáveis).
    - coeficiente de correlação amostral $= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$

      entre os atributos x e y ($\bar{x}$ e $\bar{y}$ são médias)
    - covariância $\sigma = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

Limpeza dos Dados: Dados Inconsistentes

- O que fazer com dados inconsistentes ?
  - **Eliminar dados redundantes**: tuplas cujos atributos possuem os mesmos valores (ou muito próximos).
  - **Eliminar atributos redundantes** (atributos que podem ser deduzidos a partir de outros). Ex: idade e data de nascimento; quantidade de vendas, valor por venda e venda total.

Transformação de Dados

- Algumas técnicas em aprendizagem de máquina só trabalham com um tipo de dado: apenas númerico ou apenas simbólico.
- As transformações pode ser:
  - **Conversão Simbólico-Numérico**
  - **Conversão Numérico-Simbólico**
  - **Normalização**
  - **Simplificação**

Transformação de Dados: Simbólico-Numérico

- **Conversão Simbólico-Numérico**
  - Necessário para redes neurais, SVM e alguns algoritmos de agrupamento.
  - Atributo nominal
    - de 2 valores: um dígito binário é suficiente.
    - de mais de 2 valores:
      - se houver relação de ordem (ordinal), deve ser preservada;
      - o mesmo vale para a ausência de ordem(nominal).
      - uso de sequência binárias de comprimento c , onde c corresponde à quantidade de valores.

Transformação de Dados: Simbólico-Numérico

- **Conversão Simbólico-Numérico**
  - Exemplo - nominal: 100(Azul),010(Verde), 001(vermelho).
  - Exemplo - ordinal: 00 (primeiro), 01(segundo), 10 (terceiro) e 11 (quarto).

  No caso de cadeias binárias muito longas, uma alternativa é a representação de pseudos-atributos (binários, inteiros ou reais). Ex: pais representados por continente, pib, população e área.

Transformação de Dados: Numérico-Simbólico

- **Conversão Numérico-Simbólico**
  - Técnicas que trabalham com dados qualitativos: algoritmos de classificação e associação.
  - Algumas **estratégias**:
    - **Larguras iguais**: divide o intervalo original de valores em subintervalos com mesma largura. (outliers podem prejudicar essa estratégia)
    - **Frequências iguais**: divide o intervalo original por frequência (pode gerar subintervalos de tamanhos bem diferentes).
    - Uso de algum **algoritmo de agrupamento**
    - **Inspeção Visual**

Transformação de Dados: Normalização

- **Normalização**
  - Recomendada quando os limites de **valores de atributos distintos são muito diferentes;**
  - **Evita que um atributo predomine sobre outro;**
  - A normalização pode ser **por amplitude ou distribuição:**
    - **Distribuição**: muda a escala de valores de um atributo. Ex: ordena os valores e substitui seus valores pela sua posição no ranking. (Valores: 9,8,7,2,7; substitui por 4,3,2,1,2)
      Se todos os valores forem distintos, a distribuição é uniforme.

Transformação de Dados: Normalização

- **Normalização**
  - **Amplitude**: pode ser por reescala ou padronização. (padronização lida melhor com outliers)
    - **Reescala**: define uma nova escala, com limites mínimo (*min*) e máximo(*max*) novos para todos os atributos $valor_{novo} = min + \frac{valor_{atual} - menor}{maior - menor}(max - min)$, onde *menor* é o menor valor na escala atual;idem para *maior*.
    - **Padronização:** define um valor central e um valor de espalhamento comuns a todos os atributos. $valor_{novo} = \frac{valor_{atual} - \mu}{\sigma}$, onde μ é a média e σ é a covariância.

Transformação de Dados: Simplificação

- **Simplificação**: transformação para um valor mais facilmente manipulável. Ex: idade ao invés de data de nascimento.

Redução de Dimensionalidade

- Muitos problemas possuem um número elevado de atributos (ex: textos e imagens)
- As técnicas com esse fim seguem as abordagens:
  - **agregação**
  - **seleção de atributos**

Redução de Dimensionalidade: Agregação

- **Agregação:** combina os atributos originais por meio de funções lineares ou não lineares.
  - **Análise de Componentes Principais** (Principal Component Analysis): técnica bem conhecida que correlaciona estatisticamente os exemplos, reduzindo a dimensionalidade do conjunto de dados original pela eliminação de redundâncias.
  - Obs: Essa técnica leva a perda dos valores originais. Em várias aplicações (áreas de biologia, finanças, medicina, etc), os valores originais são importantes para a interpretação dos resultados. Por isso, técnicas de seleção de atributos são mais usadas.

Redução de Dimensionalidade: Seleção de Atributos

- A **seleção de atributos** busca um **subconjunto ótimo de atributos** para o problema.
- Ela **permite**:
  - identificar atributos importantes;
  - melhorar o desempenho dos algoritmos de aprendizagem;
  - reduzir exigência de memória e processamento;
  - eliminar atributos irrelevantes e ruídos;
  - simplificar o modelo gerado e, consequentemente, sua compreensão;
  - facilita a visualização dos dados;

Redução de Dimensionalidade: Seleção de Atributos

- A **seleção de atributos não é uma tarefa trivial, pois pode existir**:
  - número muito grande de exemplos;
  - número muito grande de atributos;
  - relações complexas entre atributos, que dificultam a descoberta de relações entre eles.

Redução de Dimensionalidade: Seleção de Atributos

- Existem várias técnicas que visam selecionar atributos, as mais simples são **baseadas em ordenação**
  1. Ordena de acordo com algum critério (exemplo frequência)
  2. Seleciona
     - por **Ranking**: escolhe os n primeiros melhor classificados.
     - por **Relevância**: escolhe todos os atributos cujo valor está acima de um limiar n.