Pré-processamento

Prof. Jefferson T. Oliva

Aprendizado de Máquina e Reconhecimento de Padrões (AM28CP)
Engenharia de Computação
Departamento Acadêmico de Informática (Dainf)
Universidade Tecnológica Federal do Paraná (UTFPR)
Campus Pato Branco

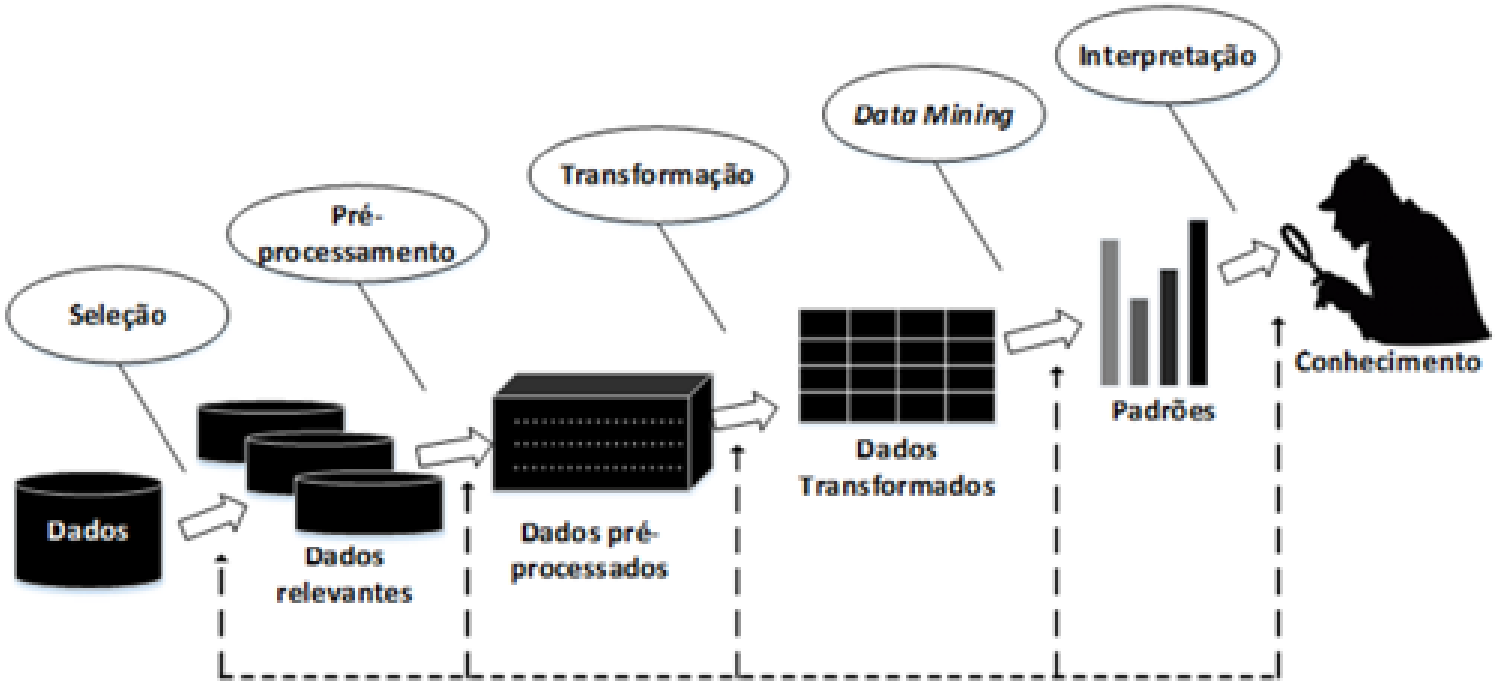
DAINF
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CÂMPUS PATO BRANCO

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Sumário

- Amostragem de Dados
- Limpeza de Dados
- Transformação de Dados
- Desbalanceamento

Introdução


Interpretação
Data Mining
Transformação
Pré-
processamento
Seleção
Conhecimento
Padrões
Dados
Transformados
Dados pré-
processados
Dados
relevantes
Dados

Introdução

- Pré-processamento é uma das fases essenciais de um projeto de aprendizado de máquina e é aplicado para diversas finalidades
  - Amostragem
  - Limpeza
  - Lidar com desbalanceamento
  - Transformação

Introdução

- No pré-processamento também inclui:
  - Redução da dimensionalidade*
  - Seleção de atributos*
- Maldição da dimensionalidade
- A etapa de pré-processamento pode ser a mais demorada na mineração de dados

Sumário

Amostragem

- Alguns métodos de aprendizado podem apresentar dificuldades para lidar com número grande de exemplos, como o k-vizinhos mais próximos (KNN - *k-nearest neighbors*)
- Quanto maior é a quantidade de dados, mais tempo será necessário para treinar um modelo preditivo
- Também, um conjunto muito pequeno pode não ser suficiente para o treinamento de um modelo eficiente
- Se as amostras não forem representativas, o modelo resultante pode não representar o problema real

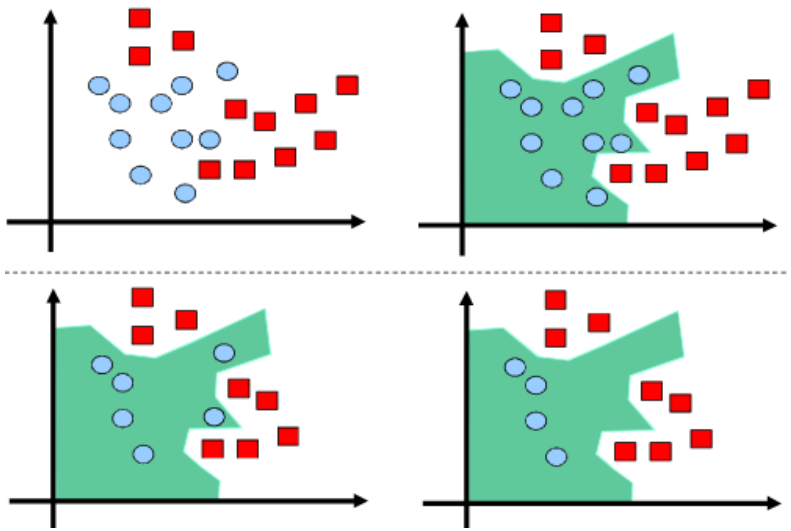
Amostragem

- É aconselhável que, além do conjunto de amostras não seja grande, os dados obedeça a mesma distribuição estatística do conjunto a partir do qual foi gerado
  - Por exemplo, a média de cada um dos atributos das amostras selecionadas deve ver próxima quando comparada à média do conjunto original

- Métodos de amostragem estatística
  - Aleatória simples
  - Estratificada
  - Progressiva

Amostragem

- Amostragem aleatória simples
  - Cada elemento é selecionado de forma aleatória
  - Amostragem com reposição
  - Amostragem sem reposição



Amostragem

- Amostragem estratificada
  - As amostras são selecionadas de forma balanceada por classe
    - Por exemplo: caso um conjunto original tenha 70% de dados de classe e 30% de classe B, após a amostragem, o conjunto de exemplos selecionados terá a mesma proporção (70/30)
    - Pode acarretar em modelos tendenciosos (favorecimento da classe majoritária)
- Amostragem progressiva
  - Começa com um conjunto de pequeno de amostras
  - No decorrer da aplicação, o conjunto aumenta progressivamente

Limpeza de Dados

Limpeza de Dados

- Geralmente, os dados não são gerados em formato adequado para a aplicação de métodos de aprendizado de máquina

- Para o melhor desempenho de algoritmos de aprendizado de máquina, é geralmente desejável que os dados estejam "limpos"
  - Entra lixo, sai lixo

  - Correção de problemas nos dados
    - Ruídos
    - Anomalias (*outliers*)
    - Valores ausentes
    - Valores inconsistentes
    - Valores redundantes

Limpeza de Dados

- Ruídos
  - Possuem valores (erros) diferentes do esperado
  - Podem ser causados por problemas nos equipamentos de coleta, transmissão e/ou armazenamento
  - Outras causas: falha humana, má fé na coleta, falhas no processo de medição
  - Alguns métodos lidam bem com dados ruidosos, outros podem apresentar dificuldades ou incapacidade para processar dados ruidosos
  - Podem ser tratados por meio de filtros, como passa-banda
- *Outliers*
  - Diversas aplicações têm o objetivo de encontrar anomalias (*e.g.* fraudes em cartões de crédito)

Limpeza de Dados

- Valores ausentes
  - Há diferentes causas para valores ausentes
    - O atributo não foi considerado importante durante a coleta de dados
    - Desconhecimento do valor durante o preenchimento
    - Distração
    - ...
  - Formas de lidar com valores ausentes
    - Ignorar valores ausentes
    - Descartar exemplos (linhas)
    - Descartar atributos (colunas)
    - Preenchimento de valores (para cada atributo com valores ausentes): média, mediana, moda, heurísticas, etc

Limpeza de Dados

- Valores ausentes

Idade	Sexo	Peso	manchas	Temperatura	Diagnóstico
18	M	90	não	38,4	Doente
?	F	65	sim	36,0	Saudável
45	F	72	sim	37,5	Doente
20	?	53	não	36,3	Saudável
25	F	48	não	37,9	Doente
29	F	55	sim	39,2	Doente
36	M	70	não	35,8	Saudável
?	F	63	não	40,0	Doente
40	M	82	?	36,0	Saudável

Limpeza de Dados

- Valores ausentes

Idade	Sexo	Peso	manchas	Temperatura	Diagnóstico
18	M	90	não	38,4	Doente
30	F	65	sim	36,0	Saudável
45	F	72	sim	37,5	Doente
20	F	53	não	36,3	Saudável
25	F	48	não	37,9	Doente
29	F	55	sim	39,2	Doente
36	M	70	não	35,8	Saudável
30	F	63	não	40,0	Doente
40	M	82	não	36,0	Saudável

Limpeza de Dados

- Valores inconsistentes
  - Dados podem conter valores inconsistentes (*e.g.* CEP inválido)
  - Se for o atributo alvo (classe), pode levar a exemplos ambíguos
    - Valores iguais para os mesmos atributos, mas as classes são diferentes
  - Caso não seja possível fazer correções das inconsistências, os exemplos incorretos devem ser descartados

40	M	82	não	36,0	Saudável
40	M	82	não	36,0	Doente

Limpeza de Dados

- Dados redundantes
  - Um exemplo é considerado redundante quando é muito similar a um outro exemplo, ou seja, os atributos possuem valores muito próximos e a classe dos exemplos é a mesma
  - Na limpeza, os exemplos redundantes devem ser removidos

Transformação de Dados

Transformação de Dados

- Tem a finalidade de mudar o tipo de um atributo
  - Simbólico para numérico
  - Binarização
  - Numérico para simbólico
  - Normalização
  - ...

Transformação de Dados

- Conversão de valores simbólicos
  - Alguns métodos de aprendizado de máquina trabalham apenas com valores numéricos
    - Exemplo: k-vizinhos mais próximos
    - Valores simbólicos precisam ser convertidos para numéricos
    - A conversão depende da ordenação dos valores
    - Também, essa operação depende da quantidade de valores (*e.g.*, se forem 2, que caracteriza como binário, basta converter símbolo para 0 e o outro, para 1), conforme o exemplo abaixo

      $sim = 1$, $não = 0$

Transformação de Dados

- Conversão ordinal para numérico
  - Caso o atributo seja ordinal e não esteja no formato numérico, o mesmo pode ser convertido de acordo a ordem de valores, como no exemplo abaixo
    - baixo (1), médio (2) e alto (3)

- Binarização
  - Valores consecutivos diferem em 1 bit
  - Codifica cada valor por um número binário
  - Exemplo de codificação termômetro para 4 bits: 0000 (0), 0001 (1), 0011 (2), 0111 (3) , 1111 (4)
    - O tamanho da sequência de bits depende da quantidade de valores diferentes que o atributo possui

Transformação de Dados

- Discretização de valores
  - Transformar valores numéricos em intervalos (ou categorias)
  - Para isso é necessário a definição de um número de categorias
  - Em seguida, mapear os valores dos atributos
  - Exemplo de discretização para o atributo "idade", cujo valor seria em anos:
    - 0-11: infância
    - 12-20: adolescência
    - 21-40: adulto jovem
    - 41-65: meia idade
    - $>$ 65: idoso

Transformação de Dados

- Normalização
  - Faz com que o conjunto de valores de um atributo assuma uma propriedade
  - Diversos métodos podem ser aplicados, por exemplo:
    - Re-escala (*min-max*): onde todos os valores são convertidos para o intervalo [0, 1]

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

    - Padronização (normalização z): caso o conjunto de valores tenha uma distribuição Gaussiana, produz valores cuja média é 0 e desvio-padrão é 1

$$x' = \frac{x - \overline{x}}{\sigma}$$

Transformação de Dados

- Outras abordagens para a transformação de dados
  - Extração de características
  - Redução de dimensionalidade

Dados Desbalanceados

Dados Desbalanceados

- É dito que uma base de dados está desbalanceada quando o número de exemplos varia em diferentes casos
- O desbalanceamento é natural em alguns domínios
  - Exemplo: detecção de fraudes bancárias
- Também, a causa do desbalanceamento pode ser problemas na coleta de dados
- Diversas técnicas de aprendizado de máquina têm dificuldades em lidar com o desbalanceamento de classes
  - Tendência: favorecer a(s) classe(s) majoritária(s)
- Alternativa: balanceamento artificial
  - Selecionar a mesma quantia de exemplos para cada classe

Dados Desbalanceados

- Outras alternativas para lidar com o desbalanceamento:
  - Aumentar a quantidade de exemplos para as classes minoritárias por meio da geração de dados sintéticos
    - SMOTE (*synthetic minority over-sampling technique*): novos dados são criados por meio de interpolações entre os exemplos da classe minoritária
    - DASYN (adaptive synthetic sampling): é uma variação do método SMOTE na qual exemplos sintéticos são gerados em regiões onde a densidade da classe minoritária é baixa
  - Induzir modelos de classificação unária: onde apenas a(s) classe(s) majoritária(s) é(são) considerada(s) no treinamento
    - Caso haja mais de uma classe majoritária, as mesmas podem ser consideradas uma única classe (detecção de anomalias/*outliers*)
    - A fronteira de decisão é criada apenas para a(s) classe(s) de interesse, na qual, caso um novo exemplo seja classificado fora dessa limitação, o mesmo é considerado anomalia/*outlier*

Referências


BISHOP, C. M.
Pattern Recognition and Machine Learning.
Springer, 2006.


DE CARVALHO, A. P. L. F.
Métodos Baseados em Distância. Aprendizado de Máquina.
Slides. Ciência de Computação e Matemática Computacional.
ICMC/USP, 2015.


DUDA, Richard O.; HART, Peter E.; STORK, David G.
Pattern classification.
2nd ed. New York, NY: J. Wiley & Sons, 2001.


RASCHKA, S.; MIRJALILI, V.
Python Machine Learning.
Packt, 2017.