Mineração de Dados aplicada à Economia utilizando a linguagem R

Clarissa Simoyama David
clarissa.david@ufabc.edu.br

Universidade Federal do ABC

Introdução

Introdução

- Bibliografia utilizada:
  - Introdução à Mineração de Dados com aplicações em R - Leandro Augusto da Silva, Sarajane Marques Peres, Clodis Boscarioli
  - R-Bloggers
  - Datasets: UCI Repository, `raw.githubusercontent.com/vincentarelbundock/Rdatasets`
- Download do R:
  - `http://cran.r-project.org/`
  - `http://www.rstudio.com/`

Convenções

- Instâncias (x_i);
- Conjunto de dados ou *dataset* ($X = x_1, x_2, ..., x_n$);
- Atributos;
- Rótulo ou classe.

Tipos de Dados

- Dados estruturados: dados armazenados em tabelas, em que linhas armazenam as instâncias de um determinado evento e as colunas representam atributos que descrevem a instância;
- Dados não estruturados: textos, imagens, vídeos, sons etc, precisando passar por uma etapa de pré-processamento.

Mineração de Dados

- Descoberta de padrões em bases de dados automaticamente;
- A partir dos padrões descobertos, há condições da geração de conhecimento útil para um processo de tomada de decisão;
- Aplicação de técnicas que recebem como entrada conjunto de dados e devolvem como saída um padrão de comportamento.

Pré-processamento de dados

Problemas:

- Dados com qualidades baixas;
- Ausência de valores;
- Dados ruidosos;
- Redundância;
- Valores inconsistentes.

Limpeza de dados

Objetivo: amenizar a existência de valores faltantes e a existência de valores ruidosos. Soluções para valores faltantes:

- Remoção da instância em que ocorre a falta do valor;
- Preenchimento manual dos valores;
- Preenchimento automático dos valores.

Soluções para valores ruidosos (chamados *outliers*):

- Inspeção e correção manual;
- Identificação e limpeza automática.

Integração de dados

Medidas para integrar dados provenientes de diversas fontes de dados. Os maiores problemas encontrados são:

- Valores inconsistentes: para um mesmo atributo são encontrados valores com discrepâncias em termos de tipo ou de domínio. Solução: remoção da instância com valor inconsistente, correção manual, procedimentos de correção automática;
- Valores redundantes: uso de nomenclaturas diferentes para atributos equivalentes, armazenamento de atributos derivados de uma mesma fonte, inserção de instâncias repetidas são exemplos de valores redundantes. Soluções: análise de correlação de dimensionalidade, PCA, análise de fator.

Transformação de dados

- Grandeza dos dados muito distintas;
- Normalização de dados;
  - Normalização min-max:

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

  - Normalização z-score:

$$v' = \frac{v - media(v)}{\sigma(v)}$$

- Conversão de dados: conversão de valores numéricos para categóricos (discretização) e conversão de valores categóricos para numéricos (codificação).

Exemplo prático para pré-processamento de dados

Análise preditiva

Análise preditiva

- Processo que permite descobrir como as instâncias de um conjunto de dados se relacionam;
- São descritos por características (atributos descritivos) e por seus rótulos associados (classes);
- Número finito de classes existentes no domínio da análise - Classificação (predição categórica);
- Número dentro de um conjunto contínuo de valores possíveis para associação - Regressão (predição numérica);
- Modelo de predição.
- Análise de comportamento em redes sociais, biometria, predição de subida ou queda de ações etc.

Classificação

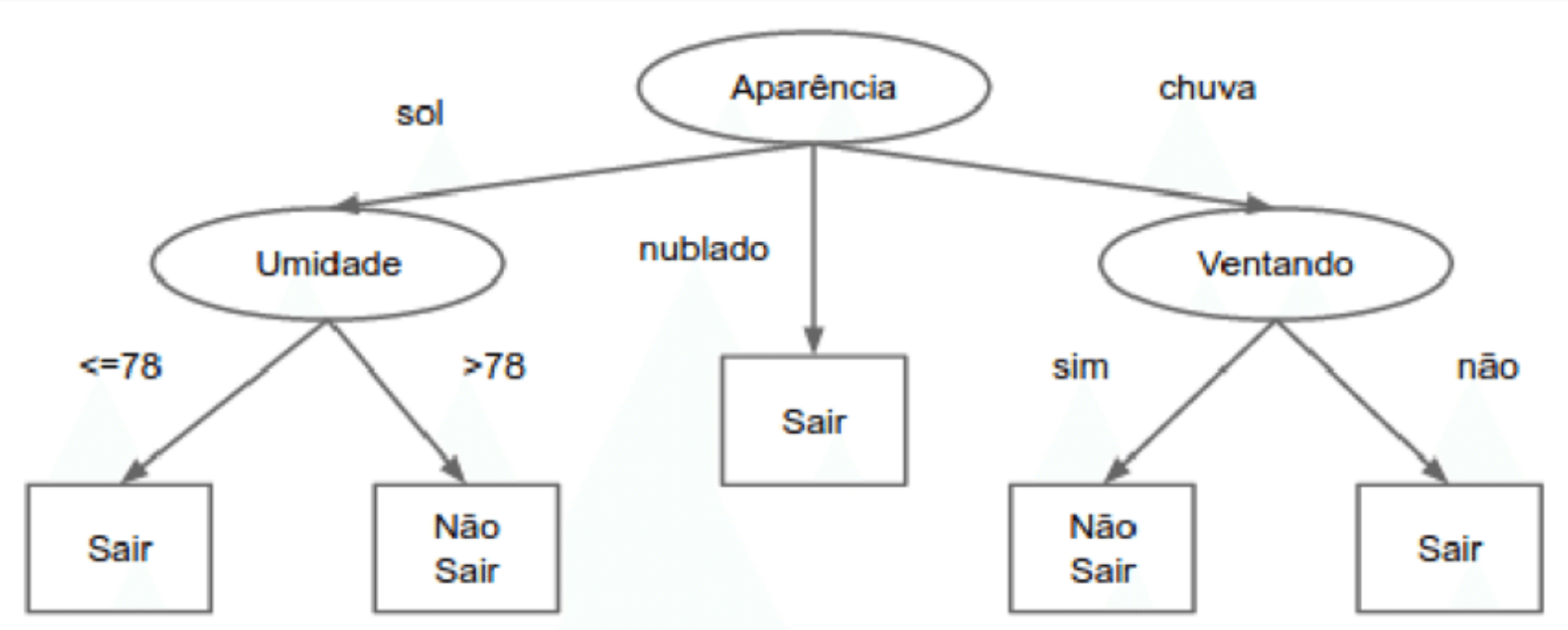
Classificação

- Classe: informação sobre a instância;
- Classificação: o objetivo é, com base em um conjunto de dados previamente rotulado, gerar um modelo capaz de prever um rótulo para uma nova instância de dados;
- Classificação binária: paciente saudável ou não saudável;
- Classificação multiclasse: classificação de serviço (excelente, bom, regular, ruim).

Árvore de decisão

- Árvore de decisão: são estruturas formadas por nós folha que representam classes e nós de decisão que correspondem a testes de valores de atributos;
- É um modelo que pode ser interpretado como regras de "SE ENTÃO SENÃO".

Árvore de decisão



Árvore de decisão

Critérios de seleção de atributos para construção da árvore:

- Ganho de informação - Entropia
- Índice Gini - Impureza (analisa as diferenças entre as distribuições de probabilidade dos valores dos atributos)

Problema que ocorre na construção da árvore: Overfitting (sobreajuste).

Soluções:

- Poda prévia
- Poda posterior

Overfitting

Exemplo prático para árvore de decisão

Regressão

Regressão

- Em classificação, queremos predizer o rótulo para uma instância nova;
- Porém, quando o rótulo é do tipo numérico (contínuo ou discreto), temos um problema de regressão ou predição numérica;
- Regressão linear simples ou não linear simples:
  - Em ambos os casos, são categorizados como simples se há um único atributo;
  - Se a função representa a equação da reta ou plano, é linear;
  - Se a função representa uma equação exponencia, é não linear.
- Regressão multivariada: mais de um atributo.

Regressão linear

- Análise estatística que envolve duas variáveis: a de resposta, o rótulo da instância, e explicativa, o conjunto de atributos descritivos;
- Considera que o valor da variável de resposta pode ser estimado por uma combinação linear das variáveis explicativas:

$$y = a + bx$$

- a e b são coeficientes (pesos) de regressão, especificando o intercepto do eixo y e a inclinação da reta;
- Deve-se encontrar valores para os coeficientes de regressão.

Regressão linear

- Método dos mínimos quadrados para encontrar o melhor ajuste dos valores:

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

$$a = \mu_y - b\mu_x$$

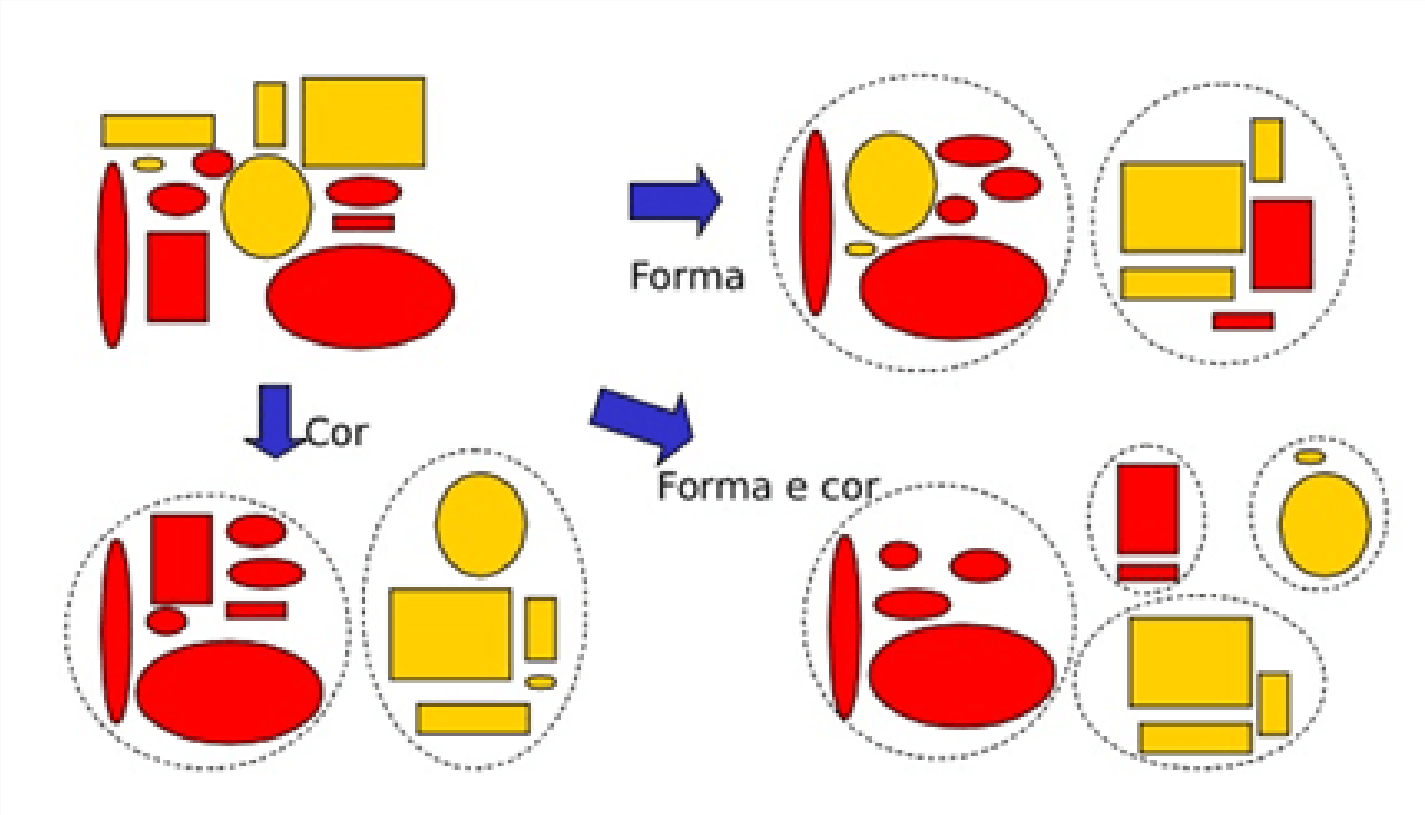
- x_i são as instâncias, y_i são os rótulos, μ é a média dos valores.

Exemplo prático para regressão linear simples

Agrupamento de dados

Análise de agrupamento

Agrupamento de Dados consiste na tarefa de agrupar exemplos de dados com base na similaridade entre eles.



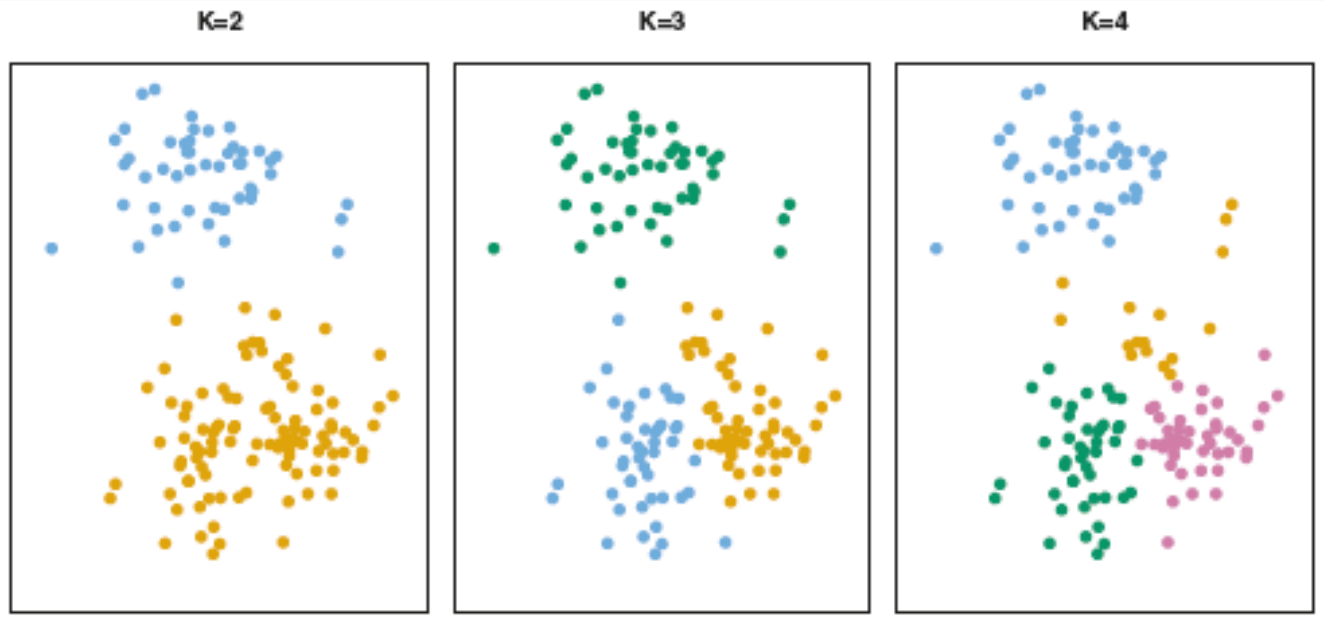
Agrupamento de Dados

- Não possui conhecimento prévio dos rótulos;
- Medidas de distância;
- Similaridade intragrupos maximizada, similaridade intergrupos minimizada;
- Agrupamento hierárquico
- Agrupamento *hard*
  - k-médias
- Agrupamento *fuzzy*
  - *Fuzzy C-means* (FCM)

Agrupamento particional - k-médias ou k-means

- Agrupamento *hard*;
- Particionar os instâncias dentre k grupos diferentes, cada um com seu centróide, no qual cada instância pertence ao grupo mais próximo da média;
- Seleciona quantidade de grupos para uma base de dados (sendo este o valor de k);
- Inicializa os centróides de forma aleatória;
- Cálculo da dissimilaridades entre os instâncias da base de dados e os centróides;
- Atualizado os valores dos centróides dos grupos;
- Novamente a realização do cálculo das dissimilaridades entre instâncias e centróides.

k-médias



A proximidade pode ser calcula de diversas formas, sendo algumas delas a distância euclidiana e correlação.

Agrupamento particional - Fuzzy C-Means

- Mesma ideia do k-means;
- Instâncias pertencem a todos os grupos, com diferentes graus ou níveis de pertinência (possivelmente nulo);
- Matriz de pertinência (matriz U);

$$U = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.1 \\ 0 & 0.3 & 0.5 & 0.9 \end{bmatrix}$$

Exemplo pratico para o agrupamento

Contato:

E-mail: clarissa.david05@gmail.com

Github: github.com/cladavid