Mineração de dados
Data Mining: Processo, toolkits, e padrões

Alvaro Henry Mamani Aliaga
Edith Zaida Sonco Mamani

Descripção da apresentação

- O que é mineração de dados
- Tipos de Atributos
- Aprendizado supervisionado e não-supervisionado
- Algoritmos de apredizado
- O processo da mineração de dados
- Toolkits

O que é mineração de dados

É um processo de análise dos dados para descobrir padrões e constroir modelos preditivos.

Mineração de dados em ingles *“Data mining”*.

Data mining é diferente da *“Data analysis”*.

Tipos de atributos

Tipos de atributos

Numerico

Ordinal

Nominal

Tipos de atributos: Numerico

Podem ser discretos ou continuos.

- Exemplo continuo: A edade de uma pessoa.
- Exemplo discreto: Numero das imagens submetidas por um usuario.

Tipos de atributos: Ordinal

São discretos.

Exemplo: O tamanho de uns tenis.

(small, medium, large)

Tipos de atributos: Nominal

- Não tem ordem.
  - Exemplo: Cor dos olhos.
  - {blue, green, black, brown}

Tipos de atributos

- Os algoritmos trabalham com valores continuos e nominais.
- Continuo pode ser discretizado.
- Valor discreto pode ser convetido a numerico.
- Valores ordinais são discretos.

Tipos de atributos

Attribute type	Description	Discrete/ continuous	Example
Continuous	Takes real values	Continuous	The amount of time spent by a user on the site
Ordinal	There is ordering in the fixed set of values that the attribute can take	Discrete or continuous	Length of session expressed as {small, medium, large}
Nominal	There is no ordering related to the values taken from the fixed set by the attribute	Discrete	Gender of a person {male, female}

Fonte: ALAG, S., Collective Intelligence in Action

Aprendizado supervisionado e não-supervisionado

Aprendizado supervisionado

Precisa de um “conjunto de treinamento”.

- Decision trees.
- Neural Networks.
- Regression.
- Bayesian belief networks.

Aprendizado não-supervisionado

- O algoritmo analisa os dados para formar *cluster-groups* de similares pontos.
  - K-means clustering.
  - Hierarchical clustering.
  - Density-based clustering.
- O aprendizado não-supervisionado é bom para análize dos dados e descoberta de padrões de um jeito automático.

Algoritmos de apredizado

Algoritmos de apredizado

Type of algorithm	Description	Type of input	Type of output	Example
Regression	Builds a predictive model that predicts the output variable based on the values of the inputs	Continuous	Continuous	Regression, neural networks
Classification	Predicts the output value for the discrete output variable	Discrete[a]	Discrete	Decision tree, Naïve Bayes'
Clustering	Creates clusters in data to find patterns	Discrete or continuous	None	k-means, hierarchical clustering, density-based algorithms
Attribute importance	Determines the importance of attributes with respect to predicting the output attribute	Discrete or continuous	None	Minimum description length, decision tree pruning
Association rules	Finds interesting relationships in data by looking at co-occurring items	Discrete	None	Association rules, Apriori

Fonte: ALAG, S., Collective Intelligence in Action

O processo da mineração de dados

O processo da mineração de dados

Modelar e selecionar os atributos

Criar o conjunto de aprendizagem

Normalizar e limpar os dados

Análize dos dados

Avaliar a qualidade do modelo preditivo

Toolkits

Weka. Waikato Environment for Knowledge Analysis.

JDM, Java Data Mining.

Conclusões

- Facilidade em desenvolvimento de aplicações data mining.
- Criar Data mining é muito util nas aplicações web.
- Fazer o datamining similar ao jdbc.
- O uso de uma API, ajuda no desenvolvimento para fazer datamining.

Obrigados