Faculdade de Tecnologia Bandeirantes - BandTec 1^o semestre de 2015

Ementa Modelagem Descritiva e Preditiva

Pós Graduação em Big Data Prof. Fabrício Jailson Barth

1 Introdução

Esta disciplina é a primeira de uma série de disciplinas que irá tratar sobre o assunto de modelagem descritiva e preditiva. Serão três as disciplinas que irão tratar de forma mais focada este assunto:

- Modelagem Descritiva e Preditiva.
- Análise Preditiva.
- Mineração de Textos.

2 Objetivo

O objetivo da disciplina de Modelagem Descritiva e Preditiva é introduzir os conceitos e ferramentas sobre ciência de dados, processo de descoberta de conhecimento e modelagem descritiva e preditiva, em especial, métodos não supervisionados de modelagem. Esta disciplina terá como seqüência a disciplina de Análise Preditiva, que irá tratar exclusivamente de modelagem e análise preditiva. A disciplina de Mineração de Textos fará uso dos conceitos tratados nas disciplinas de Modelagem Descritiva e Preditiva, mas com foco em dados não estruturados.

3 Ementa e programação aula-a-aula

A ementa e programação aula-a-aula desta disciplina estão organizados da seguinte forma:

- **Aula (1) Conteúdo**: Contexto de Big Data; Ciência de Dados; Processo de Descoberta de Conhecimento (KDD); Métodos preditivos (supervisionados) e descritivos (não supervisionados) para modelagem; Aquisição e organização dos dados; Tipos de Dados; Métodos básicos de análise descritiva e exploratória com R[1] (distribuição, concentração e correlação de variáveis), e; Organização de projetos usando o RStudio[2]. **Referências**: Capítulos 1, 2 e 3 do livro [2], e; artigo [3]. **Dinâmica**: aula expositiva com discussão de conceitos, e; exercícios de análise descritiva e exploratória com R e RStudio.
- **Aula (2) Conteúdo**: Conceito de clustering; Algoritmo K-means; Funções de similaridade; Pré-processamento dos dados para clustering; Técnicas para determinar o número de clusters, e; Interpretação e apresentação dos resultados. **Referências**: Capítulo 4 do livro [2]. **Dinâmica**: aula expositiva com discussão de conceitos e apresentação de exemplos, e; exercício envolvendo a identificação de agrupamentos planos.
- **Aula (3.1) Conteúdo**: Clustering com dados reais. **Dinâmica**: *cada aluno deverá trazer um caso, acompanhado ou não com um dataset, onde acredita que é possível e útil a identificação de clusters. Idealmente, cada aluno deverá fazer a análise destes dados e apresentar os resultados utilizando a linguagem de marcação* R Markdown[3].

[1] http://www.r-project.org/
[2] http://www.rstudio.com/
[3] http://rmarkdown.rstudio.com/

- **Aula (3.2) Conteúdo**: Funções de similaridade para valores categóricos e heterogêneos; Clustering hierárquico, e; Métodos de similaridade para clustering hierárquico. **Referências**: Capítulo 10 do livro [1]. **Dinâmica**: aula expositiva com discussão de conceitos e apresentação de exemplos, e; exercícios envolvendo a identificação de agrupamentos hierárquicos.
- **Aula (4.1) Conteúdo**: Clustering hierárquico com dados reais. **Dinâmica**: *cada aluno deverá trazer um caso, acompanhado ou não com um dataset, onde acredita que é possível e útil a identificação de clusters - pode ser o mesmo caso da aula anterior.*
- **Aula (4.2) Conteúdo**: Principal Component Analysis (PCA). **Referências**: Capítulo 10 do livro [1]. **Dinâmica**: aula expositiva com discussão de conceitos e apresentação de exemplos, e; exercícios envolvendo $\cdots$.
- **Aula (5) Conteúdo**: Conceito de Regras de Associação; Algoritmo Apriori; Medidas de Suporte, confiança, lift e leverage, e; intepretação e análise de regras de associação. **Referências**: Capítulo 5 do livro [2]. **Dinâmica**: Aula expositiva com discussão de conceitos e apresentação de exemplos, e; exercícios envolvendo a identificação de regras de associação.
- **Aula (6.1) Conteúdo**: Regras de associação com dados reais. **Dinâmica**: *cada aluno deverá trazer um caso, acompanhado ou não com um dataset, onde acredita que é possível e útil a identificação de regras de associação.*
- **Aula (6.2) Conteúdo**: *Exercício final da disciplina. O objetivo desta atividade será fazer uma análise descritiva completa fim-a-fim.*

4 Método de avaliação

Durante esta disciplina serão realizados inúmeros exercícios práticos. No entanto, quatro serão os trabalhos avaliados: Clustering com dados reais; Clustering hierárquico com dados reais; Regras de associação com dados reais, e; Análise descritiva completa fim-a-fim.

A Nota Final da disciplina será a média dos trabalhos citados acima.

Referências

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 4th edition, 2014.

[2] EMC Education Services, editor. *Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data*. John Wiley & Sons, 2015.

[3] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59(10):??–??, 9 2014.