Programa de Disciplina
Especialização em Ciência dos Dados

Módulo: III

Disciplina: Introdução ao Aprendizado de Máquina e Clusterização

Carga Horária: 15 horas (8h teóricas, 7h práticas)

Ofertante: Departamento de Estatística – DEEST/UFOP

Objetivo:
Situar o estudante no contexto de aprendizagem de máquina. Introduzir ao aluno técnicas clássicas de Clusterização de modo que seja capaz de resolver problemas usando as ferramentas disponíveis de forma crítica. Estimular no aluno questionamentos que o capacitem a aprender novas técnicas de agrupamento.

Ementa:
Contextualização de aprendizado de máquina, suas definições e objetivos. A análise de agrupamento (clusterização). Técnicas hierárquicas aglomerativas: método da ligação simples (single linkage), método da ligação completa (complete linkage), método da média das distâncias (average linkage), método do centroide (centroid method), método Ward. Técnicas hierárquicas Técnicas não hierárquicas para a construção de conglomerados (Clusters): k-médias, Fuzzy c-means.

Conteúdo Programático:

1. Introdução ao aprendizado de máquina: contextualização, definições e objetivos.
2. Análise de agrupamento (cluster):
    a) Técnicas hierárquicas aglomerativas para a construção de conglomerados (Clusters):
        i. Distância euclidiana, método da ligação simples (single linkage), método da ligação completa (complete linkage), método da média das distâncias (average linkage), método do centroide (centroid method), método Ward, outras medidas de similaridade e dissimilaridade.
    b) Métodos para encontrar o número “g” de clusters da partição final:
        i. Comportamento do nível de fusão, comportamento do nível de similaridade, coeficiente R2, Pseudo F, correlação semiparcial, pseudo T2.
    c) Técnicas hierárquicas e seleção de variáveis.
    d) Técnicas de agrupamento não hierárquicas para a construção de conglomerados (Clusters):

i. K-médias, Fuzzy c-means.

e) Outras técnicas de agrupamento.

Metodologia:
Neste módulo será utilizado o processo de Ensino-Aprendizagem convencional. Será feita a exposição teórica do conteúdo com concomitante utilização do software adotado para a disciplina. O objetivo é a construção do conhecimento baseado na compreensão crítica das técnicas utilizadas na área do conhecimento do módulo. A cada tópico apresentado um trabalho prático (aprendizagem ativa baseada em problema) é proposto para fins de fixação do conteúdo ministrado.

Atividade Prática Proposta:
Após a apresentação teórica de cada um dos métodos os alunos deverão, de forma individual, implementá-lo e testá-lo. Ao fim do módulo os alunos (grupos de 4) deverão escolher um problema da organização para solução.

Softwares:

1. R (<https://cran.r-project.org/bin/windows/base/> , última versão);
2. R-Studio (<https://www.rstudio.com/products/rstudio/download/> , última versão);
3. Python (<https://www.python.org/downloads/> , última versão);
4. Anaconda (<https://www.anaconda.com/download/#windows> , última versão);
5. Microsoft Office (2010 ou superior).

Bibliografia:

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 2. ed. New York: Springer, 2011. 738 p. ISBN 978-038731-073-2.

DASH, M. et al. Feature selection for clustering - a filter solution. In: *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* [S.l.: s.n.], 2002. p. 115–124.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. 2. ed. New York: Springer, 2009. 745 p. Corr. 9th printing 2017. ISBN 978-038784-857-0. Disponível em: <http://statweb.stanford.edu/~tibs/ElemStatLearn/>. Acesso em: 15 abr. 2020.

JAMES, G. et al. *An Introduction to Statistical Learning, with Applications in R*. 1. ed. New York: Springer, 2013. 426 p. Corr. 7th printing 2017. ISBN 978-146147-137-0. Disponível em: <http://www-bcf.usc.edu/~gareth/ISL/>. Acesso em: 15 abr. 2020.

MINGOTI, S. A. *Análise de dados através de métodos de estatística multivariada*: Uma abordagem aplicada. 1. ed. Belo Horizonte: UFMG, 2007. 297 p. ISBN 978-857041-451-9.