

Tema para Dissertação do 2º ciclo em Engenharia Informática

Título: Imputação de Valores Omissos em Análise Descritiva de Dados

Orientador: Paula Prata
(e-mail: pprata@di.ubi.pt)

Co-orientador: Maria Eugénia Ferrão (Departamento de Matemática)

Contexto

Com o crescimento explosivo de dados disponíveis provenientes de diversas fontes, a garantia da qualidade desses dados torna-se cada vez mais importante [1]. Um problema frequente é a falta de dados para uma ou mais das características em estudo. Definem-se 3 tipos de dados omissos [2] [3], dados omissos completamente aleatórios, dados omissos aleatórios e dados não aleatórios. A análise com dados omissos pode conduzir a resultados com viés e, conseqüentemente, resultar na tomada de decisões erradas.

Objetivos

Pretende-se estudar quais as estratégias existentes para a imputação de valores omissos. Num estudo de simulação, implementar diferentes métodos de imputação, fazendo uso da linguagem de programação R, e analisar o erro dessas imputações. Aplicar as conclusões obtidas para a imputação sobre conjuntos de dados reais.

Tarefas

T1 – Elaborar o estado da arte sobre imputação de valores omissos.
T2 – Estudar bibliotecas para imputação de valores omissos.
T3 – Estudo de simulação.
T4 – Imputação de valores omissos num conjunto de dados reais.
T5 – Escrever dissertação.

Cronograma de Tarefas

[illegible]

Referências

- [1] C. WU, C. WUN, H. CHOU. Using association rules for completing missing data. Proceedings of the 4th International Conference on Hybrid Intelligent Systems, Taiwan, pp. 236-241, 2004.
- [2] Rubin DB. Inference and missing data. Biometrika 63: 581-592, 1976
- [3] Little RJA, Rubin DB. Statistical Analysis with Missing Data(2ndedn.), 2002.