Análise de Clusters e Métodos Matriciais

Prof.ª Laura Moraes

Setembro 2020

Diretrizes:

- **Data de entrega: 21/10/2020, até às 23h59**.
- Os trabalhos valem **100%** da nota.
- Os trabalhos são **individuais**. Trabalho iguais não serão aceitos.
- O trabalho deve ser entregue **pelo e-class**. Os resultados das análises (plots e textos) deverão estar no **formato PDF ou HTML, sem o código**. O código deverá ser enviado separadamente.

Você usará o R ou outra ferramenta a sua escolha para minerar os dados de interesse. Estes podem ser dados de um problema do seu trabalho atual, algo de interesse pessoal, etc. Você projetará a tarefa de mineração de dados, explorará os dados e descreverá seus resultados. O objetivo é que você obtenha uma experiência prática o mais realista possível, dadas as restrições do que aprendeu.

Ao escrever sua pesquisa, pense em você como um analista contratado por uma empresa (grande ou pequena) que queira entender o estado da arte, usando mineração de dados, do problema em questão. Este guia serve como um roteiro para o que deve ser apresentado no trabalho.

1. **Entendimento do problema**: identifique, defina e motive o problema que você está abordando. Como uma solução de mineração de dados resolverá este problema?
2. **Entendimento dos dados**: descreva os dados (e fontes de dados) que darão suporte à mineração de dados para solucionar o problema. Faça alguns plots com descrições estatísticas. Analise correlações entre variáveis e suas distribuições. Essa é uma ótima oportunidade para levantar pontos de atenção, como dados faltantes, categorias mal-definidas, desbalanceamento entre categorias e variáveis altamente correlacionadas, Que hipóteses e conclusões você pode tirar olhando as descrições estatísticas?
3. **Preparação dos dados**: especifique quais pré-processamentos são necessários para a análise. Inclua codificações de dados categóricos, normalizações e redução de dimensionalidade. Compare o dataset antes do pré-processamento e depois.
4. **Modelagem**: escolha **dois algoritmos de clusterização** a serem utilizados. Quais parâmetros podem ser variados nesses algoritmos afim de se obter o melhor resultado?
5. **Avaliação**: descubra o melhor número de clusters e descreva interpretações para eles. Avalie a diferença de interpretação entre os algoritmos.
6. **Relatório**: documente através de textos e gráficos a sua análise.