

## LISTA DE EXERCÍCIO IV - Parte I

### Instruções:

- A resolução do exercício deve ser feita **individualmente**. Cópias evidentes entre trabalhos não serão aceitas.
- A entrega deve ser online via Moodle (exclusivamente), somente até a data especificada. Não serão aceitos trabalhos atrasados.
- Para cada uma das tarefas deve-se entregar o com código fonte. O nome do arquivo deve identificar a tarefa, exemplo "e4-1a.py" referente ao item "1a" da tarefa. Arquivos corrompidos serão desconsiderados.
- Além do código fonte deve-se entregar um único arquivo PDF apresentando o pseudocódigo do algoritmo desenvolvido e os resultados encontrados.
- Data de entrega: 19.11.2019 (terça-feira) até as 13:00 via Moodle (<https://moodle.ufrgs.br/login/index.php>).

NOME: ..... CARTÃO: .....

Objetivos: Análise de dados de *microarray*, técnicas de clusterização.

1. Realize a leitura do artigo: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Golub Etal (1999). Disponibilizado em: [https://www.dkfz.de/gpcf/fileadmin/downloads/Expression/Golub\\_1999.pdf](https://www.dkfz.de/gpcf/fileadmin/downloads/Expression/Golub_1999.pdf)
2. Baixe o arquivo de treinamento disponível no link: [https://web.stanford.edu/~hastie/CASI\\_files/DATA/leukemia\\_big.csv](https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv). Esta base de treinamento é formada por perfis de expressão gênica de 72 amostras de medula óssea de pacientes com leucemia aguda, cada perfil consiste da expressão de 7129 genes. Os exemplos de treinamento estão rotulados como ALL (*acute lymphoid leukemia*) e AML (*acute myeloid leukemia*), dois tipos distintos de leucemia.
3. Agrupe os dados da base de treinamento por amostra (paciente) usando o k-médias. Escolha k=2 e k=3. Comente sobre a correspondência entre os aglomerados (*clusters*) obtidos e o diagnóstico ALL/AML. Entregar relatório de no mínimo 2 páginas descrevendo o experimento e os resultados obtidos. Para cada um dos respectivos grupos (2 e 3) diga quantas amostras de pacientes ALL e AML aparecem em cada grupo.

Sugestões de pacotes a serem utilizados na execução da tarefa:

- Weka:  
<http://www.cs.waikato.ac.nz/ml/weka/>
  - Python Clustering:  
<https://datasciencelab.wordpress.com/2013/12/12/clustering-with-k-means-in-python>
  - R:  
<https://cran.r-project.org/web/packages/cluster>
- \* qualquer pacote de clusterização pode ser utilizado. Se desejar pode implementar o algoritmo k-means.