UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL - UFRGS
INSTITUTO DE INFORMÁTICA - DEPTO INFORMÁTICA TEÓRICA
BIOLOGIA COMPUTACIONAL- 2019

LISTA DE EXERCÍCIO IV - Parte 2

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 aquivo deve identificar a tarefa, exemplo "e8-1a.py" referente ao item "1a" da tarefa. Arquivos corrompidos serão desconsiderados.
- Além do código fonte deve-se entregar um um único arquivo `PDF` apresentando o pseudocódigo do algoritmo desenvolvido e os resultados encontrados.
- Data de entrega: 26.11.2019 (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. Ao aplicar a técnica de agrupamento k-means (`Trab IV - Parte I`) ao conjunto de dados de treinamento de pacientes com leucemia aguda foi possível identificar dois grupos em que as 72 amostras foram divididas. Ao Analisar as amostrar, associadas a cada um dos grupos após a aplicação do k-means, foi possível observar uma predominância de rótulos `ALL` e `AML` em cada um dos grupos. Sabe-se que dentre os 7129 genes existe um conjunto menor de genes que pode aumentar predominancia de samples de mesmo rótulo em cada grupo (todas as amostras pertecerem a um mesmo tipo de Leucemia `ALL` ou `AML`). Nesta tarefa você precisa encontrar um conjunto reduzido de genes que separe o conjunto de amostras e que aumente a predominância de rótulos `ALL` e `AML` em um dos grupos.

   (a) Baixe o arquivo de treinamento disponível no link: `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.

   (b) Crie uma população inicial aleatoria de 50 individuos cada um composto por 3572 genes escolhidos aleatóriamente no conjunto original de 7129 genes.

   (c) Utilize o algorimo k-means implementado no `Trab IV - Parte 1` para agrupar as 72 amostras considerando os genes selecionados para cada um dos 50 individuos. Escolha k=2. Calcule a correspondência entre os aglomerados (*clusters*) obtidos e o diagnóstico `ALL/AML` (label original). Calcule a aptidão de cada individuo considerando a taxa de acerto em realizar a separação das classes. Exemplo: Se para um individuo X com 3572 genes o agrupamento possibilitou que todos os dados de `ALL` e `AML` estejam perfeitamente separados em duas classes então o indivudo tem aptidão 100%. Implemente uma abordagem baseada em Algoritmo Genético (discutido em aula) para eveluir a seleção de genes e encontrar o individuo de maior aptidão. O Algoritmo Genético deve ser executado por 100 gerações. O algoritmo deve ser executado em sextuplicata sendo calculado a média e o desvio padrão da aptidão do indivíduo encontrado na última geração.

   (d) Organize uma apresentação descrevendo o algoritmo implementado e discutindo os resultados obtidos. Faça um gráfico demonstrando a convergência ao longo das 100 gerações (considere a repetição que obteve melhor aptidão ao final).