IBI5031 - Aprendizagem de Máquina para Bioinformática

Quarta Lista de Exercícios

Docente: Marcelo da Silva Reis[1]

[1] Laboratório de Ciclo Celular, Instituto Butantan

Programa de Pós-Graduação Interunidades em Bioinformática da USP
São Paulo, 27 de novembro de 2020

Instruções

- Esta lista de exercícios é para ser resolvida de forma estritamente individual;
- A resolução deverá ser feita no Jupyter notebook - utilizar um único caderno (arquivo) para todos os códigos desta lista;
- É importante que o código no caderno Jupyter seja escrito forma clara e organizada (intercale códigos com explicações utilizando o Markdown); no caso de funções, explique em um cabeçalho os argumentos que a função recebe, o que ela faz e o que ela devolve;
- A entrega do caderno Jupyter deverá ser feita no eDisciplinas, página oficial desta edição de IBI5031, até o dia do prazo final;
- Prazo final de entrega: **8 de dezembro**;
- Bom trabalho!

Questões

Nesta lista retomaremos os trabalhos com o conjunto de dados "Parkinsons", disponível no UCI Machine Learning Repository:

archive.ics.uci.edu/ml/datasets/parkinsons.

Todavia, agora vamos aplicar técnicas de Aprendizagem Não-Supervisionada para analisar esse conjunto de dados.

Questão 1. *(5 pontos)* Utilize o algoritmo de agrupamento k-médias (*k-means*), disponível no scikit-learn e que é apresentado em detalhes em:

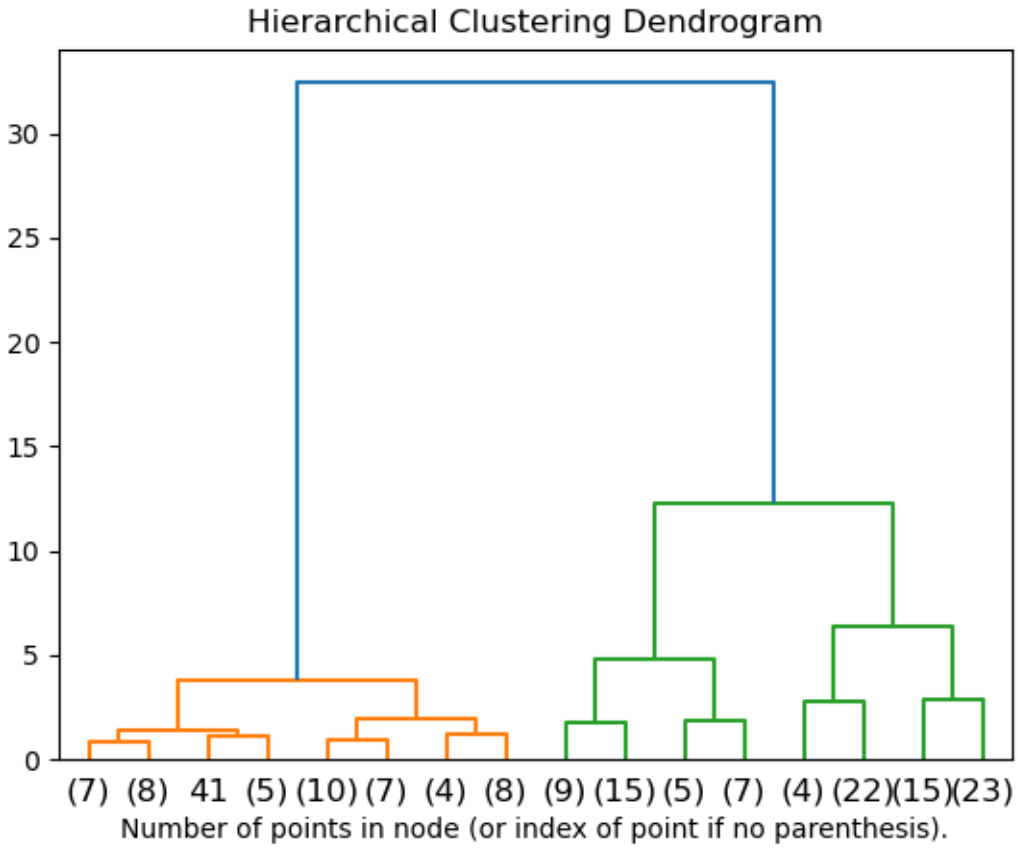
scikit-learn.org/stable/modules/clustering.html#k-means

para agrupar as características dos 197 pontos (não agrupe as classes!). Como sabemos que esse conjunto de dados é de classificação binária, inicie escolhendo agrupar os pontos em duas classes. Compare o resultado obtido aqui com as classes reais e também com o seu resultado da classificação feita por SVM feita na Lista 3 (considere apenas o melhor resultado de classificação obtido na Lista 3). Monte uma tabela para comparar os erros e acertos das duas metodologias.

Questão 2. *(5 pontos)* Com os métodos de agrupamento hierárquico disponíveis no scikit-learn:

scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

faça uma aglomeração hierárquica das características dos 197 pontos. Teste realizar aglomerações com diferentes critérios, e plote dendogramas para visualizar as hierarquias de seu cluster. Para um ou mais critérios, o procedimento foi capaz de separar (total ou parcialmente) os pontos em dois grupos, cada um correspondendo a uma das duas classes reais? Em outras palavras, foi possível obter um resultado como o da figura abaixo, com cada uma das subárvores em laranja contendo pontos de apenas uma das duas classes? (e.g., "com Parkinson" à esquerda e "sem Parkinson" à direita?)


Hierarchical Clustering Dendrogram
30
25
20
15
10
5
0
(7) (8) 41 (5) (10) (7) (4) (8) (9) (15) (5) (7) (4) (22) (15) (23)
Number of points in node (or index of point if no parenthesis).