

IBI5031 - Aprendizagem de Máquina para Bioinformática

Terceira Lista de Exercícios

Docente: Marcelo da Silva Reis¹

¹ Laboratório de Ciclo Celular, Instituto Butantan

Programa de Pós-Graduação Interunidades em Bioinformática da USP

São Paulo, 23 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: **4 de dezembro**;
- Bom trabalho!

Questões

1. **Questão 1.** (10 pontos) Entre no repositório UCI Machine Learning Repository e baixe o conjunto de dados "Parkinsons":

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

Este conjunto de dados foi gerado para classificar pessoas como portadoras ou não da doença de Parkinson a partir de padrões vocais; portanto, é um conjunto de dados de classificação binária, com uma amostra de 197 pontos e 23 características.

Utilizando as máquinas de suporte vetorial do scikit learn, cuja documentação está disponível em:

scikit-learn.org/stable/modules/svm.html,

classifique Parkinsons utilizando modelos SVM. Considere o uso de diferentes kernels (transformações não-lineares feitas de forma implícita) disponíveis no scikit learn:

- linear;
- polinomial;
- função de base radial (RBF);
- sigmóide (implementado com a tangente hiperbólica).

Faça seleção de modelos utilizando validação cruzada e diferentes valores de parâmetros para os diferentes kernels (e.g., o parâmetro d que determina o grau do kernel polinomial). Organize seus resultados finais em uma matriz de confusão.