15.851–8 – Mineração de Dados

Lista 4

Rafael Izbicki

Lista em Trios.

Exercício 1. Considere o banco (real) `dados_covid.csv`. O objetivo é prever se um paciente que chega a um hospital e é submetido a um exame de sangue simples possui COVID ou não.

(a) Leia o banco de dados e faça um histograma (ou uma estimativa contínua de densidade) de cada variável por grupo. Quais variáveis parecem diferencias mais os grupos? Seu gráfico deve ser profissional.

(b) Divida o conjunto de dados em treinamento e teste. Você pode escolher as proporções.

(c) Pensando que o objetivo deste problema é fazer uma triagem de que pacientes podem ter COVID (para serem analisados mais a fundo), quais métricas você considera mais importantes para fazer a avaliação de um modelo preditivo? Justifique.

(d) Ajuste os seguintes métodos preditivos de classificação *na sua versão probabilística*: a. Logística com penalização lasso (com tuning parameters escolhidos via CV), b. KNN (com tuning parameters escolhidos via CV), c. Árvore (com poda), d. Floresta, e. Rede Neural (com arquitetura escolhida por você) e f. XGBoost (em que você deve otimizar ao menos 2 tuning parameters).

(e) Qual dos métodos acima apresentou melhores resultados? Compare seu desempenho via a curva ROC.

(f) Com base na sua resposta para o item (c), defina um corte para cada um dos classificadores obtidos e compare ele segundo as métricas que achar razoáveis para esse problema.

(g) Utilize PDP's para interpretar cada um dos métodos