Programming Assignment 0: A Sample Classification Problem

Due September 5, 2022

In the repository you have access to tabular data with ten features $x_j \in \mathbb{R}$, 100,000 rows, and a binary label $y \in \{0, 1\}$, saved as .csv file under pa0/data/pa0_train.csv. You are tasked with constructing a model $\tilde{y} : \mathbb{R}^{10} \to \{0, 1\}$ so that $\tilde{y}(x) = y$ for as many pairs $(x, y) \in \mathbb{R}^{10} \times \{0, 1\}$ *as possible*. There is starter code in pa0/code/ which you may and should use to guide your development.

You must submit code you use to construct this model and and output (.csv is fine)* reporting performance metrics of your choice (auc, other statistical metrics, etc.). All submissions should include, at the very least, accuracy (percentage of correct predictions). You may use any library you're already familiar with and any kind of model (I recommend sklearn and logistic regression), and should not spend too much time optimizing your model. You should focus your effort, instead, on interpreting results and commenting on performance.

*Collate your data into dataframe and call method 'to_csv(path).'