CS 5834: Fall 2023: Intro to Urban Computing

Homework Assignment 3

Intro to ML

Assigned: Oct 3, 2023
Due: Oct 17, 2023, 11:59pm ET
Total Points: 100

Use the census dataset from the UCI ML repository at:
https://archive.ics.uci.edu/ml/datasets/Census+Income
for this question. Download this data and take some time to understand its format and contents. The ".names" file gives information about the attributes, conventions, etc. and the ".data" and ".test" file give you the training and test datasets, respectively. Note that most information is categorical. Our goal is to setup an ML problem where we predict whether a person makes over 50K a year using information such as age, work class, marital status, education, occupation, relationship, race, gender, hours per week, and native country.

We will use the scikit-learn library in Python for this assignment. Sample code that you can adapt will be provided.

(40 points) Conduct an exploratory analysis of the data. Plot different attributes, their distributions and study how the target class varies w.r.t. these attributes. Make qualitative conclusions about relationships you observe. Form your own understanding of which attributes are likely predictive and explain it.

(60 points) Develop 3 ML classifiers: Naive-Bayes, Logistic regression, and Decision tree classifiers to predict if an individual makes over 50k per year. Use 5-fold cross validation to evaluate your model. Report precision, recall and F-score of the classification. Make qualitative conclusions/interpretations about which classifier works best (and why).

You might find the following links useful:
https://scikit-learn.org/stable/modules/naive_bayes.html
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html