Huanying Yeh

Mr. Tivnan

EN.580.140.12.IN21 Statistical Foundations of Machine Learning

11 January 2021

Course Project Proposal

Main Objective:
My project would be a classification task that predicts whether a person's income exceeds $50K/yr based on census data. The dataset is from the UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Census+Income.

Motivation:
Classification models are very helpful for looking for important features of data, finding patterns, and making educated predictions about known new data. It can be applied in numerous ways, such as weather forecast, stock market predictions, and targeted advertising. When having a very large quantity of data, one can use Machine Learning to simplify the categorization process and keep improving the predictions.

Data:
According to the data description of the dataset, the attributes are:

>50K, <=50K.

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.

capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

As seen from above, the dataset covers numerous demographic factors that can determine a person's income. There are 48842 instances of data, which allows for in-depth training. The data are labeled, so we will use supervised learning.

Notes:
This task is one of the homework assignments of an intro-level Machine Learning course taught by professor Hung-yi Lee at National Taiwan University. I've been watching his lectures on YouTube: https://youtu.be/CXgbekl66jc and would like to attempt one of the projects given in the course.