UNIVERSITY OF SRI JAYEWARDENEPURA

B.Sc. (Special) Degree in Computer Science - Part I

CSC 369 2.0 Machine Learning I

Assignment - 2018

Note: This assignment accounts 25% of the module assessment.

The CSV file "Adult Census Income Binary Classification dataset.csv" consists of an adult dataset obtained from a Census database. In each row contains an individual's annual income results from various factors. Intuitively, it is influenced by the individual's education level, age, gender, occupation, and etc. The objective of this assignment is to develop machine learning models to predict whether an individual's annual income exceeds $50K. The followings are the attributes of the dataset:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. 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.
4. education-num: continuous.
5. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
6. 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.
7. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

8. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
9. sex: Female, Male.
10. capital-gain (income from investment sources, apart from wages/salary): continuous.
11. capital-loss (losses from investment sources, apart from wages/salary): continuous.
12. hours-per-week: continuous.
13. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, OutlyingUS (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.

There are 32,561 individual records in the dataset. For some individuals, some attributes are missing and they are denoted by a question mark ('?').

Design the following machine learning models to the given dataset suitably.

1. k-nearest neighbor algorithm (Use Python 'sklearn' package)
2. Logistic regression (Use 'Keras')
3. Neural network (Use 'Keras')

All coding, results, plots and documentation should be in a single 'IPython' or 'Jupyter' notebook. Create a 'GitHub' account for you if you don't have a one yet and store this file in your 'GitHub' repository. Submit the 'link' of this file on or before the given submission date.

The followings should be clearly addressed when designing the models.

1. How to pre-process the data (normalization, non-linear transformations, feature extraction, coding of discrete inputs and targets, handling of missing data, etc.)?
2. How to handle the 'over-fitting' problem?
3. How to compare the designed machine learning models? Use post-training analysis methods such as accuracy, confusion matrix, precision, recall and F-measure appropriately.

Lecture-in-charge
Dr. TGI Fernando
Department of Computer Science
University of Sri Jayewardenepura
Nugegoda

-------END-------