Machine Learning 2019 assignment

Students: 1073978, 1741606, 1619542, 1325952, 1653526

Computer Science

Coms3007

1. Description of the dataset:

- What are the attributes?

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.

- What are the targets?

The targets are the income classes (>50K, <=50K).

- How many data points do you have?

48842

- Sample data points from the dataset:

39, State-gov,77516, Bachelors,13, Never-married, Adm-clerical, Not-in-family, White, Male,2174,0,40, United-States, <=50K

50, Self-emp-not-inc,83311, Bachelors,13, Married-civ-spouse, Exec-managerial, Husband, White, Male,0,0,13, United-States, <=50K

38, Private,215646, HS-grad,9, Divorced, Handlers-cleaners, Not-in-family, White, Male,0,0,40, United-States, <=50K

53, Private,234721, 11th,7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male,0,0,40, United-States, <=50K

28, Private,338409, Bachelors,13, Married-civ-spouse, Prof-specialty, Wife, Black, Female,0,0,40, Cuba, <=50K

37, Private,284582, Masters,14, Married-civ-spouse, Exec-managerial, Wife, White, Female,0,0,40, United-States, <=50K

49, Private,160187, 9th,5, Married-spouse-absent, Other-service, Not-in-family, Black, Female,0,0,16, Jamaica, <=50K

52, Self-emp-not-inc,209642, HS-grad,9, Married-civ-spouse, Exec-managerial, Husband, White, Male,0,0,45, United-States, >50K

31, Private,45781, Masters,14, Never-married, Prof-specialty, Not-in-family, White, Female,14084,0,50, United-States, >50K

42, Private,159449, Bachelors,13, Married-civ-spouse, Exec-managerial, Husband, White, Male,5178,0,40, United-States, >50K

37, Private,280464, Some-college,10, Married-civ-spouse, Exec-managerial, Husband, Black, Male,0,0,80, United-States, >50K

- State what you are trying to predict with the data?

Prediction task is to determine whether a person makes over 50K a year or not.

- How were the inputs/targets normalised?

?

- How was the data pre-processed?

See attached preprocessingNotebook.ipynb

- How was the data split into training/validation/test data?

The training and validation data is in the dataProccesed2.csv. The full set of the data was spilt into 2/3 for training and validation and 1/3 for test data.

- List of classification algorithms applied to the data:

Naïve Bayes ?

- Details of each implementation

- Error on the test

- How and why hyper-parameters were selected

- Presentation of errors in the form of confusion matrix (for e.g.)

Decision Trees ?

- Details of each implementation

- Error on the test

- How and why hyper-parameters were selected

- Presentation of errors in the form of confusion matrix (for e.g.)

Logistic Regression ?

- Details of each implementation

- Error on the test

- How and why hyper-parameters were selected

- Presentation of errors in the form of confusion matrix (for e.g.)

1. ?