Machine Learning Analytics RSM8413 – Fall 2022

Group Assignment 3 Due date: 2022-11-27

Deliverables (to be submitted on Quercus):
1. Report including a detailed description of your findings. At most 10 pages + appendices.
2. All Python source code either in a Jupyter Notebook (*.ipynb) or a Python file (*.py). One file!
3. Team<X>predictions.txt (please see Part Two below).

Include an Executive Summary (at the beginning) describing your most salient findings. Explain all steps and results clearly and cogently, so that a reasonably intelligent though statistically naïve manager could understand it. You need to include all graphics in your report. Your narrative should be clear and concise, accompanied by supporting evidence in the form of graphics and tables. All tables and graphics should be well formatted (e.g., tables should not run over from one page to another).

The Case:
The data set was extracted from data provided by the US Census Bureau. The task is to find the set of demographic characteristics that can best predict whether or not the individual has an income of over $50,000 per year. For each record, there are fourteen attributes, and one output variable. Further information on the data set follows below. The training data set is USCensusTraining.csv and contains all the fields for 25,000 records. The test data set is USCensusTest.csv; it contains 7561 records, and is missing the income field.

Part One:

Please perform the following tasks on the training data set USCensusTraining.csv:

a. Generate a neural network to predict income using the other attributes. You may need to ignore one of the attributes. Describe the topology of the resulting network and draw a detailed picture of the network topology, labeling the input and output layers.

b. Which are the most important variables for predicting income?

c. What is the predicted accuracy? What does this number mean?

d. Compare the predicted income with actual income. Which error is the model more prone to making? Is this type of error more protective of, say, banks or loan applicants?

e. Which occupations are associated with predicted income over $50,000? Which education levels? Which ages? Is this intuitive? Construct graphs of the top three categorical predictors, and their relationship to predicted income. Make sure you fine-tune these graphs, and arrange the graphs so that they are truly helpful.

f. Construct a histogram of one numeric variable which is important in the model, with an overlay of income. (You may wish to use “normalize” to increase contrast.) Then construct a histogram of one numeric variable which is not important in the model, with an overlay of income. Do the histograms support the findings of the neural network?

Part Two:

With this assignment, we begin to graduate to the big leagues of data mining modeling. In addition to the usual material you submit, we will grade you on how accurate your predictions are! (Hey, you are getting to be pretty sophisticated data analysts. 😊)

The last section of your code submission should present your best model. This part of your code should write your 7561 predictions of income for the test data set into an ASCII file named Team<X>predictions.txt (where <X> stands for your team number). The ASCII file should contain only one variable, zeroes and ones (0’s indicate a prediction of income <=50K; 1’s indicate a prediction of income >50K). Please also submit this file. Your overall percentage of correct predictions compared to the best submission will then count out of 1.5 points toward the assignment grade.

You must use neural network modeling for Part Two of course, but feel free to back up your neural net model with whatever appropriate tools you care to apply, including CART, for example. Use whichever neural network methods and settings you prefer. Show which is your best model, comparing among your best candidate models.

Hint: You will need to partition USCensusTraining.csv into training and validation sets, if you want to verify the accuracy of your predictions. Spend some time tweaking your best models. Remember to balance accuracy against overfitting.

Attribute information:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- demogweight: continuous. (Explanation: Scalar measure of socio-economic status of the case's district)
- 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.