RWTH AACHEN UNIVERSITY

The First Assignment of IDS 2018-2019

Introduction

This assignment guides you through the analysis of a real-life data set using the techniques and tools provided in the course. This assignment is used to test the understanding of the material in lectures 1-9. It is necessary to follow the assignment in the given order since the result of certain questions might depend on answers to previous steps.

About the data set:

The dataset used in this assignment has 48,842 records and a label indicating a salary of <=50k or >50k USD. There are 11 attributes:

- `Age`
- `workclass: type of employer`
- `education: highest level of education attained`
- `education-num: highest level of education attained`
- `marital-status`
- `occupation`
- `relationship`
- `race`
- `sex`
- `hours-per-week: hours worked per week`
- `native-country`

As it can be seen, for education there are two columns (numerical and categorical); you should choose the right one based on what you want to do (in some cases you need to use categorical and in some other cases you need to use numerical).

The data set is split into two seperate files. Use the *adult.data-3.csv* for the exercises in the two first sections **"Getting to know the data"** and **"Building models to perform predictions".** Use the *adult.test-3.csv* file for the exercises in the third section **"Analysing the quality of the prediction models"**. It is important that you use the data sets as specified. Do not mix them up!

Exercises

Getting to know the data

1. Identify the precise type of each attribute. Note that the right answers are like: "numerical-continuous-ratio", just Numerical or Categorical is not enough.
2. Removing outliers:
    a. Explore into *age* and *hours-per-week* and identify outliers (Boxplot returns whiskers). After identifying outliers, remove them (do it just once on the data set). Now you should have two data sets (cleaned and original).
    b. Draw a Boxplot of the cleaned data set for *hours-per-week*. Is there still any outlier? If so, explain why?
3. Basic visualization:
    a. Visualize mean and median of age, and *hours-per-week* per *sex* by separate plots for the cleaned and the original data sets (there should be 8 plots, 4 plots for the original data set and 4 plots for the cleaned data set).
    b. Explain how mean and median of *age* per *sex* change when you remove the outliers.
4. Distribution:
    a. Explore into distribution of *age* in the original data set. Does it have any well-known distribution (normal, uniform, skewed, ...)? If so, what are the main statistical features (mean, median, and mode) of this specific type of distribution.
    b. Explore into distribution of *age* and *hours-per-week* together. Explain how the data is distributed with respect to these two attributes in the original data sets (using Jointplot)?

Building models to perform predictions

5. Consider all the categorical attributes except label as descriptive features and label (income) as target feature.
    a. Train two decision trees (one based on Entropy and another based on Gini).
    b. What are the best attributes (based on Gini and Entropy) for splitting the trees in the third round of ID3?
    c. Prune the tree which is made based on Entropy by 7000 as minimum number of samples. Identify which value of the first node/attribute is chosen to be split in the second round of ID3 and explain why.
6. Train two logistic regression classifier for predicting the label (income) based on the original dataset.
    a. Create two sets of independent variables: (1) containing the attributes *age*, *marital-status* and *sex* and (2) containing the attributes *work-class, education* and *hours-worked.* Now you should have three data sets: (1), (2) and the initial dataset containing all
    b. Again, inspect the attributes and their data types of your sets (1), (2) and (3). Which attributes are suitable as an input for the logistic regression and which need to be modified first? Why? Modify the attributes that need preprocessing using one-hot encoding.
    c. Train three logistic regression classifiers based on the two data sets (1), (2) and (3). Document which parameters and functions you used.

d. Interpret the three resulting models and compare them. Which model do you recommend and why?
e. Note! (use the model based on all data (3) for the evaluation section)

7. Now you want to make sure that you have the most suitable classifier model for your problem. Try to make a neural network for the original dataset:
    a. What are the inputs of your network?
    b. What are the possible number of input pattern for your network(just including categorical attributes)?
    c. Train your network with all the inputs to predict income attribute:
        i. First, with default parameters and return the parameters.
        ii. Second, try to find the optimized number of hidden layers and nodes. (Start with default number and then at least go with one number above and one number below the default)
        iii. Third, try to train your model with one linear activation function and one non linear activation function, name the functions and explain if there is any difference in your networks and why?(you can use evaluation metrics to show that which activation function works better for this data set)
        iv. (your explanation, code, parameters and the metrics to evaluate your code are needed as a result)
        v. Which model do you recommend to be used as your classifier and why?(with respect to the number of hidden layer and activation function(linear or nonlinear))
    d. Note! (use the model with default parameters of hidden layer and activation function for the next section)

Analysing the quality of the prediction models

8. You have been given an additional file: a validation dataset, to assess the performances of your models. You will then test the validation data on some of the trained models to obtain a confusion matrix for some of the algorithms, calculated comparing the predicted vs true label in validation data; you will then also validate the data with some common metrics obtainable from the confusion matrix itself: precision, recall, accuracy and F1 score.
   Note: do not consider any decision trees for this part.

At this point you will be able to answer the following questions:

a. Considering the metrics, what is the best model and why?
b. Does any model suffer from underfitting or overfitting?
c. Comment on some possible validation techniques in the case a separate test set is not available.

Deliverables

The deadline for the assignment is **Sunday 09/12/2018 23:59**. You will need to hand in your submission via moodle. Note that there is no extension for the deadline and also late submissions will not be considered. Your submission should include a **jupyter notebook**, which presents your results and also contains the python code used to obtain the results.

Report requirements:

- Use the provided jupyter notebook

Grading

Participation in the assignment is one of the prerequisite for taking the written exam. The two assignments and the exam form a whole and it is not possible to retake parts of the course, i.e., the results of the assignments expire after the exam. Furthermore, assignments can only be redone in the next academic year.

The grade of this assignment counts 20% towards the final grade. There are three main sections in you assignment and one related to your report style:

1. The first section is about knowing your data and accounts for 25 percent of your assignment grade;
2. The second section for modeling accounts for 40 percent;
3. The third section including evaluation part accounts for 25 percent
4. As a data scientist showing results is as important as what you have done; therefore, the report style determines 10 percent of your assignment grade.

- please note that correctness of your code, its result and also accuracy of your explanation are important.