Annual salary prediction

Dušan Antić SW81/2019, Mladen Vasić SW28/2018

1. Motivation

We would like to be able to predict if a person's salary would exceed 50 000 dollars based on their background information. We could gauge how good our current salary is based on what is our own background, or we could also ask for a raise if our solution predicts a better salary than we currently have.

2. Research questions

Our goal is to predict salaries based on following fields:

- Age
- Workclass
- Final weight
- Education
- Marital status
- Occupation
- Relationship
- Race
- Sex
- Capital gain
- Capital loss
- Hours per week
- Native Country

Dataset contains roughly 32 500 rows.

3. Related work

We did not find any other works based on this very same dataset that we used.

4. Methodology

Dataset already came split up with ratio of 66/33% for training/test sets and they each came in their own separate files, then we loaded both sets using command line arguments. During runtime, our solution removes rows with NaN values ("not a number" values, in this dataset "?"), and then it performs label encoding for non-numeric columns. After encoding is done, the "salary" column is deleted.

After preprocessing, we used support vector classification implemention found in scikit learn libary. We instantiatied the SVC class using default settings and we fit it on our training set, and with that instance we predicted our values.

5. Discussion

For the evaluation measure we selected micro F1 score. There were attempts to optimize our solution by using linear support vector classification, we also tried removing various columns in various different combinations and we never achieved a better result than ≈ 0.79084

6. References

Link the dataset used in this solution

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/?fbclid=IwAR0io31SiSmpHFjfP48y1IVSzPR-oPcVYuwGQUSqQPDjfJsPq9fu2YF1r3M