Assignment 7B

Computer Science Department, University of Crete

MACHINE LEARNING - CS 577, Fall 2022

Deadline: Friday, 8/1/2023, 23:55 on e-learn (`https://elearn.uoc.gr`).

Deliverable files: Submit a zip/tar etc. file containing:

- a report in PDF with ALL answers to theoretical tasks, observations and figures produced by scripts.
- **all** Python script files written by you in the scope of the assignment.

NOTE: 10% penalty on your assignment grade will be applied in failing to follow the above instructions for deliverables or not explaining your code/ having comments!

Exercise 1 - Causal-Based Feature Selection (Programming) [100 points]

In this exercise you will implement the feature selection algorithm "Forward-Backward Feature Selection" and you will apply your code in a real-world dataset, widely used in the literature of machine learning, namely "Breast Cancer". You will write your code in the **Assignment7b.py** file that we give you. There, you can find:

1. the real-world dataset that is provided from sklearn
2. the library that you need to install for the conditional independence test. The test that you will use is called Kernel-Based Conditional Independence Test[1]. To install the library run the command *pip install causal-learn* in your environment.
3. a function that performs the Conditional Independence test called *stat_test* that takes as input dataset D (ndarray), the indexes of the features (list of integers), the index of target variable T_{idx} (int) and the selected variable indices S (list of integers).

[1] https://arxiv.org/pdf/1202.3775.pdf

4. the signature of the forward and backward functions that you need to implement. In Assignment7b.py, fill in the code for the forward and backward functions, following the algorithms and .

Algorithm 1 Forward Selection

Input: Dataset **D**, Target Variable **T**, Variables **V**, Significance Threshold **a**
Output: Selected Variables **S**

R $\leftarrow$ V $\setminus$ S //Remaining Variables
while S changes **do**
    //Identify V^* with min p-value given S
    $V^* \leftarrow \arg\min_{V \in D}$ pvalue$(T; V|S)$ ▷ In the provided python code, the cond. ind. test function takes as input the indexes of the variables, not the variables per se.
    //Remove V^* from R
    $R \leftarrow R \setminus V^*$
    if pvalue$(T; V^*|S) \leq a$ **then**
        $S \leftarrow S \cup V^*$
    end if
end while
return S

Algorithm 2 Backward Selection

Input: Dataset **D**, Target **T**, Selected Variables **S**, Significance Threshold **a**
Output: Selected Variables **S**

while S changes **do**
    // Identify V^* with max p-value given $S \setminus V$
    $V^* \leftarrow \arg\max_{V \in S}$ pvalue$(T; V|S \setminus V)$ ▷ In the provided python code, the cond. ind. test function takes as input the indexes of the variables, not the variables per se.
    //Remove V^**if independent*
    if pvalue$(T; V^*|S \setminus V^*) > a$ **then**
        $S \leftarrow S \setminus V^*$
    end if
end while
return S

Model Selection and Evaluation

1. Stratify and Split D into 80% train and 20% test. Remember that D in the provided code contains both the features and the target variable (last column)

2. Perform a 5-Fold Cross validation optimizing for ROC AUC.

3. Apply the forward-backward feature selection, for different significance thresholds, $a = \{0.05, 0.01, 0.005\}$.

4. For the classification models, run at least 5 configurations of SVM and Random Forest (5 for each one) with hyper-parameters of your choice.

5. Report the ROC AUC in the 20% hold-out set for the best configuration

6. Report the selected feature subset of the best configuration

7. Rerun the classifiers without feature selection and report the performance of the best configuration in the held-out set

Final Model and Observations

1. Report the final model and the selected features. Show the selected feature ids in your report

2. Run only the forward backward using alpha=0.05 on all data and store the selected set S. Now, run only the backward phase with set S as input and store the S'. What do you observe? If the sets are different, give a possible explanation.