Introduction to Machine Learning Course

Exercise 3 –
The Elections Challenge –Modeling

TA in charge: Tal Daniel

The elections are getting warmer and the campaign managers are pushing us real hard to get some answers to the questions and challenges posed at the beginning of the campaign. After a tough negotiation, we were able to convince them to start with the following issues

- Predict which party will win the majority of votes
- Predict the division of voters between the various parties
- On the Election Day, each party would like to suggest transportation services for its voters. Provide each party with a list of its most probable voters
- Identify the factor (voters' characteristic) which by manipulating you are most likely to change which party will win the election

Mandatory Assignment

First, redo the data preparation task (of the former exercise), but this time on the right set of features (see the first comment below). Specifically, write a script that will implement the following

1. Load the Election Challenge data from the ElectionsData.csv file
    - Can be found at the "The Election Challenge" section in the course site
    - Make sure that you've identified the correct type of each attribute
2. Select the right set of features (as listed in the end of this document), and apply the data preparation tasks that you've carried out in the former exercise, on the train, validation, and test data sets
    - At the very least, handle fill up missing values
        - All other actions, such as outlier detection and normalization, are not mandatory. Still, you are encouraged to do them, as they should provide you added values for the modeling part
    - Whether to use the validation set for pre-processing steps is your call
3. Save the 3x2 data sets in CSV files

Next, write a Python process that can handle the following prediction tasks

- Predict which party will win the majority of votes
- Predict the division of voters between the various parties
- On the Election Day, each party would like to suggest transportation services for its voters. Provide each party with a list of its most probable voters

Such a process should implement and execute the following

1. Load the prepared training set
2. Train at least two models
    a. Each training should be done via cross-validation on the training set, to maximize performance of the model while avoiding overfitting
3. Load the prepared validation set
4. Apply the trained models on the validation set and check performance
    a. It is your call which performance measure to use, and it is possible to check multiple measures
5. Select the best model for the prediction tasks
    a. The model selection is "manual" (not an automatic process), but it should be based on the performance measurements
6. Use the selected model to provide the following
    a. Predict to which party each person in the test set will vote (predict the label of each row in the test set)
    b. Construct the (test) confusion matrix and overall test error

Please submit

1. The Python script file that implements the data preparation part using the right set of features
2. CSV files of the prepared train, validation and test data sets
3. The Python script file that implements the modeling (training and testing) part
4. A CSV file that contain the voting predictions (predicted labels) on the test set
5. A short documentation that
    a. Explains your process and any significant decision/insight you would like to share
    b. Explains the following
        i. Your choice of the two models to train and test their performances
        ii. Your choice of the performance measurements
        iii. Your decision which model to use, to provide the final predictions
    c. Includes answers for the following
        i. Predict which party will win the majority of votes
        ii. Predict the division of voters between the various parties (the percentage of votes per party)
        iii. The (test) confusion matrix and test error for the predicted votes

Non-Mandatory Assignments

The following list includes additional, non-mandatory, assignments. You are highly encouraged to do at least some of them. Each functional implementation of ANY assignment will get a bonus

A. Automate the model selection procedure, i.e. the selection of the best model based on the performance measurements of all the trained models (Step 5 of the mandatory process)
    - Provide a Python script file and a document that explains the process, your insights, and conclusions.
B. It may very well be that "one size doesn't fit all", namely that modeling differently each of the tasks provides better results. Check this paradigm -
    - Use a different modeling procedure (train and test) for each of the three mandatory prediction tasks
    - Compare results with the results obtained using the one model approach
        - Note that "Better results" are not merely a higher accuracy, but also simpler models (why is it important?), stable predictions, etc.
    - Provide a Python script file for each of the tasks and a document that explains the process, your insights, and conclusions.
C. Handle the fourth predication task
    - Identify the factor (voters' characteristic) which by manipulating you are most likely to change which party will win the elections
        - Provide a document that list these factors, the manipulation needed (e.g. increase voters yearly income), and the new winning party
            - you may provide a few such scenarios, each of which results with a new winner
            - handle this task strictly from a technical perspective, meaning please ignore the semantic of the features
        - Provide a Python script file/s that implements such a manipulation
        - Explain how did you identify these key factors

Triplets Mandatory Assignments (Bonus for Pairs)

The following list includes mandatory assignments for triplets. Triplets must submit all the assignments. Pairs are highly encouraged to submit at least some of them, and for pairs it will be considered as bonus assignments

- Implement the Widrow-Hoff (else known as LMS, ADALINE) algorithm and compare its classification performances to the Perceptron algorithm (implemented in scikit learn). The comparison should be conducted on the following data sets
    a. Iris, digits (scikit learn include these data sets) – Learn ONLY binary classification settings, while using the 1-vs-all scheme
    b. Generate two artificial data sets that represent binary classification (you can use scikit learn to generate the data sets) – one where the Perceptron

algorithm converge faster and the other one where the Widrow-Hoff algorithm converge faster

i. "Faster" means using less iterations (a smaller number of training examples)
ii. Explain what are the properties of the data sets that gave an advantage to each of the algorithms

c. You are encouraged to build the comparison in a similar way to the scikit learn example "Comparing various online solvers"
   i. Please note that this example is conducting a multi-class classification, while you need to compare only binary classification

Comments

- The right feature set –
  Since there is a redundancy in the original set of features, the "right" set of features is not unique (meaning, there are a few subsets of features that could have been selected). Still, from now on, please use the following features (a.k.a "the right feature set") at the Election Challenge assignments
  - Avg_environmental_importance, Avg_government_satisfaction, Avg_education_importance, Most_Important_Issue, Avg_monthly_expense_on_pets_or_plants, Avg residancy altitude, Yearly_ExpensesK, Weighted_education_rank, Number_of_valued_Kneset_members
- You may use a "balanced" training set (to handle imbalanced classes while training a model) but the test set should reflect the original data distribution, so that the reported performance measurements will be unbiased

This exercise can be submitted in pairs or triplets!

- **You should submit only one copy but remember to document who are the contributors**
- **"No Couples/Triplets Swapping" during the semester**
  - **At least not without my formal approval**
- **Triplets must submit the bonus assignments that are marked mandatory for triplets**