Tutorial 8 Assignment

Semi-Open Machine Learning Project

In this assignment, you will do a ML project based on a dataset of your choice and build a ML model using the 7 steps outlined in the Tutorial 8 LiveScript.

Choose Your Dataset

You should choose a dataset that is similar to the `Car_Advertisement.csv` data that we have used in the tutorial, which should satisfy that:

- It is a binary classification problem, i.e., the target (Y) is a binary result like "Purchased/Not Purchased", "Lived/Death", "Adopted/ Not adopted".
- There are at least 200 entries.
- There are at least 2 numerical features.

Some good resources to find data sets are:

- Kaggle: https://www.kaggle.com/datasets
- The UCI ML dataset archive: http://archive.ics.uci.edu/ml/index.php
- Just Google "[keywords] dataset"

If you run out of ideas finding your dataset, you can choose one from my recommendation list:

- **The Titanic dataset**, where you will predict whether a passanger survived or not based on their age, gender, number of siblings, etc. Download it here: https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html
- *+ Hint: in this problem, "gender" is more likely to play a role in the prediction, so don't delete "gender" like we did in the Car_Advertisement dataset!*
- **The Adult income dataset**, where you will predict whether income exceeds $50K/yr based on census data. Download here:http://archive.ics.uci.edu/ml/datasets/Adult
- *+ Hint: there are a lot of features in this dataset. You don't have to use all of them. To be simple, you can choose only 2 numerical features to build your ML model. If you don't know how to handle some of the more complicated categorical dataset, you can also ignore them.*

Go through the 7 Steps in a LiveScript/Python Jupyter Notebook

Once you select a dataset, you will go through the 7 steps like I did in the tutorial.

1. Frame the problem. Identify the features and targets; Clarify your goal.
2. Get the data. Split into training and testing set, and set aside the testing set for now.
3. Explore the data. Get some insight on the attributes!
4. Pre-processing the data to better expose the underlying data patterns to ML algorithms.
5. Try out different ML models. You need to try at least these 5 models: **discriminant analysis model, KNN, Naive Bayes, SVM-linear, SVM-nonlinear.**
6. Fine-tune your models. Once you are confident about your final model, measure its performance on the test set to estimate the generalization error.
7. Present your solution. Lauch your model.

You can choose either MATLAB Livescript or Python Notebook, depending on your preference.

Feel free to adapt the Livescript / Python Notebook that I provided.

Submit an Exported PDF

Once you are done with the 7 steps, please export your MATLAB Livescript / Python Notebook into a **PDF**, and submit the PDF on Canvas **by Monday March 30 at noon 12:00 pm.**

If you encounter an error when exporting your Livescript to PDF, you can export it as Word and then save the Word document as PDF.

Marking Rubric

This assignment worths 10 points. You will be assessed on:

- Data pre-processing (3'): this corresponds to Step 1-4 in your project. You will be assessed on how well you understand the identify the problem and the technical accuracy of pre-processing.
- ML Models & Interpretation (4'): this correcponds to Step 5-6. You will be assessed on the coding accuracy of the ML models and your through process on selecting the best model
- Present results (3'): this corresponds to Step 7. In this part, you should be able to convey your result to a non-technical staff.

See a detailed marking rubric on Canvas.