Homework - 0

Course - Machine Learning (Spring 2022)

Instructor - Ankit Sharma

IIIT Kota

Lab Eval Week: ***Jan 31-Feb 4***

Logistics: *This lab would be conducted in* ***groups of two****. Groups can be found in the 'groups.pdf' attached.*

Part 1 — Setup & Warmup (10 Points)

Exercise 0: Python setup as described in this week's lab. If not you can:

1. Follow the tutorial here: http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial0.html
2. See the slides attached ('Python Tutorial.pptx').
3. Check Google Colab: https://research.google.com/colaboratory/

Exercise 1: Get yourself comfortable with Python basics Matplotlib. For that:

1. Follow the tutorial here: http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial1/tutorial1.html
2. See the slides attached ('Python Tutorial.pptx').
3. https://docs.python.org/3.6/tutorial/

Exercise 2: Get your hands dirty with Matplotlib, Numpy & Pandas — Python libraries. For that:

1. Follow the tutorial here: http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial2/tutorial2.html
2. Documentation on Pandas. https://pandas.pydata.org/
3. Documentation on matplotlib. https://matplotlib.org/
4. Documentation on Numpy. https://numpy.org/

Part 2 — Data Handling (90 Points)

Instructions:

1. Please go through this:
   - data Exploration tutorial:http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.html ,
   - read the chapter 2 (title: 'Data') of the textbook (https://www-users.cse.umn.edu/~kumar001/dmbook/index.php) it is based on,
   - see the slides for this chapter attached ('data.pdf'),
   - while reading chapter/slides focus on the topics mentioned in steps 4-8 below,
   - and play with the notebook associated with this tutorial:http://www.cse.msu.edu/~ptan/dmbook/tutorials/tutorial4/tutorial4.ipynb .
2. The above tutorial is based on specific datasets. But for this homework each student group will use a different dataset from the UCI repository, etc.. Please choose the dataset as per the below table. The list of imbalanced dataset below is taken from : https://www.kaggle.com/general/46744 , please read this post if that helps understanding the context of datasets better.

Group ID	Dataset Name with Missing Values	Group ID	Dataset Name with Missing Values	Imbalanced Dataset
1	https://archive.ics.uci.edu/ml/datasets/Arrhythmia	8	https://openmv.net/info/class-grades	https://www.kaggle.com/crawford/emnist
2	https://archive.ics.uci.edu/ml/datasets/Auto+MPG	9	https://openmv.net/info/food-consumption	https://www.kaggle.com/mlg-ulb/creditcardfraud
3	https://archive.ics.uci.edu/ml/datasets/Credit+Approval	10	https://openmv.net/info/kamyr-digester	https://www.kaggle.com/uciml/bioassay-datasets

4	https://archive.ics.uci.edu/ml/datasets/Dermatology	11	https://openmv.net/info/raw-material-properties	https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000
5	https://archive.ics.uci.edu/ml/datasets/Echocardiogram	12	https://openmv.net/info/travel-times	https://www.kaggle.com/mlg-ulb/creditcardfraud
6	https://archive.ics.uci.edu/ml/datasets/Horse+Colic	13	https://openmv.net/info/class-grades	https://www.kaggle.com/crawford/emnist
7	https://archive.ics.uci.edu/ml/datasets/Mushroom	14/15	https://openmv.net/info/food-consumption	https://www.kaggle.com/crawford/emnist

3. Perform the analysis mentioned in the tutorial (step 1) using your dataset. The tutorial is a guide for analysis and it is expected that you understand the concept and perform the data handling which is tuned/relevant to your dataset. Therefore, getting some domain understanding of your dataset should be a good starting point.
4. **Standardization and Normalization (10 points):** You should follow the various strategies to deal with standardization/normalization provided in:
    - (a) [5 points] the tutorial (step 1), and
    - (b) [5 points] sections 6.3.1 & 6.3.2 from sklearn: https://scikit-learn.org/stable/modules/preprocessing.html. You should be able to understand the various strategies and apply them to your data.
5. **Missing Values (20 points):** You should follow the various strategies to deal with missing values provided in:
    - (a) [5 points] the tutorial (step 1),
    - (b) [7 points] from this link: https://www.kaggle.com/alexisbcook/missing-values and
    - (c) [8 points] sklearn: https://scikit-learn.org/stable/modules/impute.html . You should be able to understand the various strategies and apply them to your data. Also, It would be good to explore more python libraries or codes which provide other kinds of missing value imputation techniques for

example using the pandas data-frame as the input.

6. **Discretization (20 points):** You should follow the various strategies perform discretization provided in
   - (a) [10 points] the tutorial (step 1), and
   - (b) [10 points] section 6.3.5 from sklearn: https://scikit-learn.org/stable/modules/preprocessing.html . You should be able to understand the various strategies and apply them to your data.
7. **Sampling (20 points):** You should follow the various strategies to perform sampling strategies provided in
   - (a) [8 points] the tutorial (step 1), and
   - (b) [12 points] sampling for **imbalanced** data: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook where you can use the imbalanced data sets from the table above. You should be able to understand the various strategies and apply them to your data.
8. **Dimensionality Reduction (20 points)**: You should follow the various strategies to perform dimensionality reduction provided in
   - (a) [5 points] the tutorial (step 1),
   - (b) [10 points] section 6.5.1 & 6.5.2 from sklearn: https://scikit-learn.org/stable/modules/unsupervised_reduction.html and
   - (c) [5 points] t-SNE: https://lvdmaaten.github.io/tsne/ (https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html). You should be able to understand the various strategies and apply them to your data.

Deliverables:

1. A python notebook:
   - with fairly detailed description of the various statistics and methods used.
   - Your descriptor should indicate the reason for your choice of those stats/methods for your dataset
   - and any interesting findings in your dataset using the employed techniques.
   - Make use of ‘text sections’ in notebooks to add the above descriptions
2. Upload the above notebook along with any other relevant files like readme files etc if needed as part of your submission to this assignment. If each group member has made his/her own notebook, try to combine the notebooks into a single notebook. If not possible upload all the notebooks and other files. Although you have worked in groups, but each member of the group should return the classroom assignment by uploading same material.
3. Demonstrate the working notebook during lab evaluation to the instructor. Without instructor evaluation no marks will be provided even if you have submitted the assignment.