Data Preprocessing Project

New Attempt

Due Oct 10 by 11:59pm **Points** 100 **Submitting** a file upload **Available** Aug 26 at 12am - Nov 25 at 11:59pm

Dear students,

Please see the following instructions. This is a data preprocessing project assignment. You may choose a language of your choice. Data Preprocessing currently involves 80% of the efforts in any commercial data science or data engineering project in the industry.

Python is very popular and used extensively in this course. However, you are permitted to choose an alternate language (such as Julia) only if you have the experience and confidence, and if you will not need any code samples given to you in the alternate language you choose over Python. In general, the entire course depends mainly on Python and is used extensively for all course demos and code samples (tutorials). You may refer to these artifacts and piggyback your code on top of existing tutorials so that you are successful in your class projects.

1. Find the tutorial (jupyter notebook) on Data Preprocessing on CANVAS.

Files->labs-->Tutorial_4_Data_Preprocessing.ipynb

2. Clearly understand each step of the tutorial by executing the tutorial in Python. Export to PDF and submit the PDF file for this part.

3. Next you have two choices. Use choice (b) only if choice (a) is not possible.

3(a) You may choose a data set which is already unclean and fix it with data pre-processing.

[[[[

for 3(a) check to see if below link has unclean datasets:

The following link "**perhaps**" has unclean datasets ready for preprocessing. Check it out. If not, you may use the link to choose a dataset and use simple techniques to make it unclean first.

For Example, one of the links is (London Air dataset download link):

https://www.londonair.org.uk/london/asp/datasite.asp?CBXSpecies1=COm&CBXSpecies2=NOm&CBXSpecies3=NO2m&CBXSpecies4=NOXm&CBXSpecies5=O3m&day1=1&month1=jan&year1=2018&dagraph (https://www.londonair.org.uk/london/asp/datasite.asp?CBXSpecies1=COm&CBXSpecies2=NOm&CBXSpecies3=NO2m&CBXSpecies4=NOXm&CBXSpecies5=O3m&day1=1&month1=jan&year1=2018&day2=1&month2=j

]]]]]

3 (b) You can reengineer any dataset on CANVAS so that you can change it from good to bad by clever use of instructions in the language you are programming in. So, it may be all about becoming more proficient in the language. Then apply data preprocessing.

Apply cleaning (preprocessing) tasks on your data. Make sure you cover all techniques of data pre-processing.

The Rubric will be as follows:

Testing the data preprocessing Jupyter notebook tutorial4 on CANVAS:
Export tested program with result data into a PDF file. 40% (these are just like bonus points)

Remaining 60% (steps 3(a) or 3(b) :

Detailed breakup of remaining 60%:

1. Choosing domain area, finding dataset, preparing data:	total 15% :
(a) Choosing the domain area of value.	5%
(b) Finding the data set in any domain.	5%
(c) If it is too clean, use any technique to modify the CSV file manually or programmatically for pre-processing.	5%
2. Documenting summary in PPT and a short report :	total 45% :
(a) Documenting or describing a comprehensive set of techniques for preprocessing.	5%
(b) Showing all steps in proper order for preprocessing and why you think the order you used is important.	20%
(c) Developing and documenting human insights with human interpretation on preprocessed data and possible effect on predictions.	10%

(d) Splitting dataset into two sets of data randomly: training and test data. (this is for practice)
Calculate Means and Standard Deviations on both partitions of the dataset.
Document your ratios for splitting and results. 5%
Comparing training and test sets and developing comparative intuition on the
meaning of your results of splitting,
with respect to statistical parameters: 5%

**** 5% extra points for fixing any imbalanced dataset issues with clear documentation ****
Imbalanced dataset is a dataset for example, where there are disproportionate records of each kind say, 80 percent records on women and 20 percent records on men.
Finally: Submit the jupyter file or if you are using any other language, submit code and export your code into PDF file and submit both files.
Submit all documentation with it (as a separate report).
Also mention each team member's contribution.
CAVEAT:

It is possible that your data set is not rich enough to perform all preprocessing steps on a single data set. In this case, it is okay for you to use different data sets so you may apply partial set of pre-processing steps on each data set and thus complete all kinds of preprocessing operations across all the data sets.

In any case, you may describe how you would order the pre-processing steps on a single dataset, so that the operations are consistent with each other, such as dependencies, because of the ordering.

CSC177-rubric-1

Criteria	Ratings	Pts
Test using in-class Jupyter notebook (tutorial4 on data preprocessing)		40 pts
Chose domain area		5 pts
find dataset for preprocessing		5 pts
Make dataset ready for preprocessing		5 pts
document techniques for preprocessing		5 pts
arrange steps for preprocessing in correct order and explain why		20 pts
document insights after preprocessing		10 pts
split dataset using the split function and calculate mean and standard deviation		5 pts
compare the split datasets and develop intuition on datasets		5 pts
		Total Points: 100