CS 412 project documentation

Tathagata Ganguly (tgangu2)

(a) what is your data and task?

My data is Young people data from Kaggle. I have chosen the task to predict the tendency of young people to predict how likely they would be willing to spend money on good food.

(b) what ML solution did you choose and, most importantly, why was this an appropriate choice?

I have done preprocessing at first by removing NaN values of non-categorical columns with mean of the rest of the data in the column.Then Replaced the NaN values of categorical columns with mode of the rest of the data in the column.After that I implemented the special Label Binarizer to implement One-hot encoding for categorical data columns. Then I performed feature Engineering using 'selectKBest' library. Then I divided the data into test ,development and training data and used three different types of classifier to train. Atlast I found out random forest classifier to be performing better than the other two.

(c) how did you choose to evaluate success?

I chose to evaluate success based on the accuracy ,root mean square error and f1-score.

(d) what software did you use and why did you choose it?

I chose to use python scikit-learn package as it has all the machine learning libraries implemented for preprocessing, classification. Only parameter tuning is required to be done from our end.

I have used Jupiter notebook as the parts of the code can be partially run and only that portion of the code would be interpreted and not everything. This saves a lot of time as sometimes it might be time consuming.

(e) what are the results?

```
Accuracy :
RandomForestClassifier
0.484848484848
LinearSVC
0.411764705882
LogisticRegression
0.393939393939
```

(f) show

some examples from the development data that your approach got correct and some it got wrong: if
you were to try to fix the ones it got wrong, what would you do?
In future, I would try to scale the data better and should perform ensemble methods like “bagging” and boosting” more efficiently.

GitHub Link for the project: https://github.com/bhepuganguly/IML_Project.git

Few links I used:

Label encoder: https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/

FillNaN :https://machinelearningmastery.com/handle-missing-data-python/

Feature selection : https://machinelearningmastery.com/feature-selection-machine-learning-python/

Label_Binarizer : https://stackoverflow.com/questions/31947140/sklearn-labelbinarizer-returns-vector-when-there-are-2-classes