CS 514
APPLIED ARTIFICIAL INTELLIGENCE
PROJECT 4

Otto Group Product Classification Challenge using decision trees

https://www.kaggle.com/c/otto-group-product-classification-challenge

PROJECT BY

JASMINE

CS 514
APPLIED ARTIFICIAL INTELLIGENCE
PROJECT 4

INDEX

CS 514
APPLIED ARTIFICIAL INTELLIGENCE
PROJECT 4

Requirements:

To run the source code, you must have the below software installed in your machine.

Software	Download link
Anaconda	https://www.continuum.io/
Python 2.7.14	https://www.python.org/downloads/release/python-2714/
sklearn	https://pypi.python.org/pypi/scikit-learn/0.15.2
numpy	http://www.scipy.org/scipylib/download.html
OS interfaces	https://docs.python.org/2/library/os.html

PROJECT BY

JASMINE

CS 514

APPLIED ARTIFICIAL INTELLIGENCE
PROJECT 4

Usage Manual:

Use my code that is provided to you by the professor.After downloading the entire source code and the data folders, store it in any location. The python scripts are customized to automatically adjust and find the data files subject to both the 'script' and the 'data' folder are under the same parent folder.

In the constants.py file, change the PROJECT_PATH to the location of the project folder. A data folder is created for you .Download the 3 data files(train.csv, test.csv and sample.csv) and copy them into this folder. In terminal change directory to the project folder and run the project. You would find the output saved in ensemble file(in submission mode).

Explanation about the project:

The sklearn.calibration.CalibratedClassifierCV(The base_estimator is fit on the train set of the cross-validation generator and the test set is used for calibration. The probabilities for each of the folds are then averaged for prediction. In case that cv="prefit" is passed to __init__, it is assumed that base_estimator has been fitted already and all data is used for calibration.)is used in this project.

The project has 3 modes-cv, holdout, submission that the user can set(by setting the variable Mode and saving the file and running it).

In the cv mode returns scores and predictions. The mean Log loss achieved is 0.468. It generates a file that gets stored in blend folder.

In the holdout mode(The data set is separated into two sets, called the training set and the testing set. The function approximator fits a function using the training set only. Then the function approximator is asked to predict the output values for the data in the testing set.The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model.) returns the score ie the loss score. In this mode, the Log loss of 0.463 is achieved.

In the submission mode, output file is generated which is stored in ensemble folder. The submission file includes the outputs that were generated on the runs.

PROJECT BY

JASMINE