

CIS 4526: Foundations of Machine Learning
Homework 2: Decision Trees

Instructor: Alex Pang
Due Date: 3/31 23:59 pm

Overview

The goal of the first homework is to implement Decision Tree and Random Forest yourself. I will give you a script (DT_orig.py) that follows a blog, namely, Implementing a Decision Tree from Scratch, where the author explains how you can implement a Decision Tree from scratch. However, it has deficiency and your job is to make it better. Download 2 python files (DT_orig.py and DT_first_last.py) and a jupyter notebook for testing from Canvas and fill in the missing parts.

What you need to do

Fill in the missing code inside the TODO sections. In particular, add the following parts

1. Add a `_gini` method to calculate gini index so the model can switch between using entropy and gini as the criterion.
2. Make the fit and predict method work with DataFrame instead of numpy arrays.
3. Make the code work for the case where the target variable can be a categorical variable as well as integer variable from encoding the different classification classes.
4. Add `classification_report` and `confusion_matrix` functions (do not call the sklearn versions).
5. Implement the RandomForest model (do not call the sklearn version).
6. Add an additional stopping criteria based on whether the impurity (entropy or gini) is low enough in addition to `max_depth` and `min_samples_split`. The idea is have a parameter to control overfitting.
7. Add reasonable enough comments to each methods.

Zip up the 4 files listed below and submit the zip file to Canvas.

1. breast_cancer.csv
2. DT_orig.py (no change)
3. DT_first_last.py where you replace first and last with your actual first and last name.
4. ML_SP23_HW2.ipynb with your answer inside the notebook

Grading criteria

1. Correctness
2. Good enough comments
3. Elegance
4. Oral explanation (possible)