CNG 562 Introduction to Machine Learning

Homework 1

Due April 5th, 2019

Adult Census Income

This data set consists of nearly 50K rows extracted from 1994 US Census Bureau extracted by Ron Kohavi and Barry Becker.

The prediction task is to determine whether a person makes over $50K a year.

There are 55 articles on Google Scholar citing this dataset. For Kohavi's original article, see

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

This part of the homework should be returned as a Jupyter Notebook with the filename <SURNAME>_adult_census_income.ipynb in your GITHUB account.

In your Jupyter notebook, you must at least,

1. Download the data from Kaggle,
2. Explore the data prior to developing of prediction model,
3. Prepare the data for prediction,
4. Train *at least with*
   a) Logistic Regression,
   b) Decision Trees,
   c) Support Vector Machines,
   d) K-Nearest Neighbour Methods.

5. Search for the best performing models, using Grid Search and Cross-Validation.

Ref: https://archive.ics.uci.edu/ml/datasets/Census+Income

Forest Fires

This dataset was presented with Cortez and Morais' article:

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: [Web Link]

The prediction task is to determine the area burnt in the fire.

This part of the homework should be returned as a Jupyter Notebook with the filename <SURNAME>_forest_fires.ipynb in your GITHUB account.

Data is publicly available on Google Sheets:
https://docs.google.com/spreadsheets/d/1kKO6DN2JZFvxu4XVegyGqYZTQayJseBbyfwgPW6R7kw/edit?usp=sharing

In your Jupyter notebook, you must at least,

1. Download the data from Google Sheets,
2. Explore the data prior to developing of prediction model,
3. Prepare the data for prediction,
4. Train *at least with*
    a) Linear Regression,
    b) Decision Trees,
    c) K-Nearest Neighbour Methods.
6. Search for the best performing models, using Grid Search and Cross-Validation.

The original authors have noted that there are many more “small fires”. Please re-iterate the above steps to develop a binary classifier to detect a “small fire”. Here, given a threshold T, all fires which burnt area less than T, should be considered a small fire. You must decide on a good value for the threshold T.

Ref: https://archive.ics.uci.edu/ml/datasets/Forest+Fires