Assignment2 report

Haijin He

Part1:

PART 1: Given the dataset in Assignment 1, train three classifiers of your choice on the data to achieve the highest possible cross-validated accuracy. You may use any library you want. You will turn in a report describing your activity and the results you obtain.

Tool for part1: Orange

Using default parameters

Classifiers:

decision tree: maximum depth 100, do not split subsets smaller than 5, induce binary tree

Naïve Bayes

logistic regression: regulization with Ridge L2, C=1

Random Forest: number of trees 10, do not split subsets smaller than 5

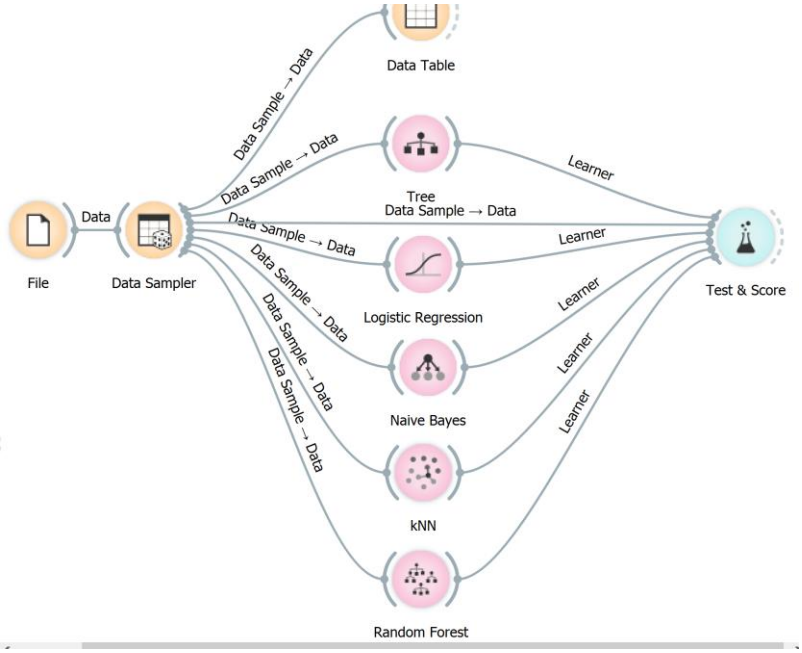
kNN: k=5

The data format is first converted to csv using a Python 3 script converter.py

Data then loaded into Orange and various models are used to evaluation accuracy.

A data Sampler widget is added for fast computation, and for some reason if input data is large decision tree widget will give an error. Data Sampler is set to random sample 30% of input data.

Orange graph:



Result:

The best result is from Random Forest, which have a F1 of 0.999

Method	AUC	CA	F1	Precision	Recall
Random Forest	1.000	0.999	0.999	0.999	0.999
kNN	1.000	0.997	0.997	0.997	0.997
Logistic Regression	1.000	0.996	0.996	0.996	0.996
Naive Bayes	1.000	0.996	0.996	0.996	0.996
Tree	0.996	0.992	0.992	0.992	0.992

Part2:

PART 2: Program, in your preferred language, a hierarchical clustering algorithm to cluster the dataset in Assignment 1. You will measure goodness of your clustering using Rand Index. You may tune the cut-off

parameter to obtain high accuracy. You may use the knowledge that number of clusters is four.

Programming language: Python 3

Source file: hierarchical.py

Hierarchical clustering algorithm: Agglomerative

Distance measure: Euclidean

Distance between clusters: average distance.

Due to the implementation efficiency, only around 400 random samples from assignment 1 is fed to the algorithm and Rand index evaluated.

A plot is generated to show the relationship of number of cluster in the result vs Rand Index.

We can see cluster number 4 is best, with a Rand Index of 0.98

It is interesting that cluster number above 4 only slightly reduced Rand Index. A closer look at the divided clusters showed they largely retains the 4 cluster structure, with other clusters having only one or a few data points. So the observation may be specific to the input data.