2018 년 2 학기 기계학습 실습 과제

개요

- 두 개의 public domain dataset 을 사용하여 4 개 이상의 classifier 를 사용하여, 각 classifier 의 성능을 비교 분석하고 그 dataset 에 대해 최대한 모든 분석을 한다. 본 과제의 목적은 단순히 프로그램을 돌린 결과만 제시하는 것이 아니라, 그 dataset 에 대한 본질을 알아내는 것이다. 각 학생이 가지고 있는 데이터 분석력을 최대한 보여야 한다.

마감

- 10 월 17 일 (수) 밤 11 시 59 분
- 블랙 보드에 업로드

Datasets

- UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/)에서 제공하는 Car Evaluation DataSet(https://archive.ics.uci.edu/ml/datasets/Car+Evaluation) 및 Air Quality Dataset(https://archive.ics.uci.edu/ml/datasets/Air+Quality)
- 제공된 Data set 을 그대로 사용하기 어려울 경우 수정해야할 필요가 있음. (Air Quality Dataset 은 원래 Regression 용도로 만들어진 Data 이기 때문에, classification 으로 사용하기 위해서는 attribute 를 하나 정해서 class(nominal) 형태로 변환해 줄 필요가 있음.

Tools

- 다양한 툴 사용 가능
    - Weka(Default), Orange, RapidMiner 등등의 GUI 도구
    - 가능하다면 scikit-learn 등의 라이브러리를 이용하여 코딩하여도 가능

Analysis and Evaluation

- 각각의 Dataset 에 대하여 ID3, C45, NaïveBayes, Logistic Regression, MLP 를 적용시킨 모델 학습 및 학습 결과 분석(Logistic Regression 및 MLP 는 실행은 해보지만, 상세한 분석은 하지 않아도 됨)
- 필요한 경우 전처리(Preprocess)를 적용
- 분석 결과를 사용함으로써, dataset 자체에 대한 분석을 시도하도록 하여야 함.
- 별도의 test set 을 생성하지는 않고, training set 을 학습 평가의 척도로 사용
- zeroR, oneR 의 결과를 baseline 으로 사용하여 비교

Report

- 실험 요약
- 데이터에 대한 설명
- 실험 설계 및 방법. 진행 내용 구체적으로. 전처리가 필요한 경우, 전처리 과정에 대한 설명 필요
- 각각의 Dataset 에 대하여 비교 분석 결과
- 결론