PGP in AI/ML
Feature Engineering – Project – 1

Submission Date: 23 59hrs on 5-07-2019

Total Marks: 12

The aim of this project is to find most important features in the given dataset using different algorithms and compare them. You will use the filter and wrapper based methods explained in Module-2 of the course and make comparison of how the feature selection affects the performance of the model.

Dataset Description: This is a Glass Identification Data Set from UCI https://archive.ics.uci.edu/ml/datasets/glass+identification).

The problem is to predict the type of class on the basis of the chemical analysis. The study of classification of types of glass was motivated by criminological investigation. At the scene of the crime, the glass left can be used as evidence (if it is correctly identified!).

The data set has 214 instances and 9 features and a class label (which can take 7 different category of values). The following are the 9 features:

[1]	RI	refractive index
[2]	Na	Sodium
[3]	Mg	Magnesium
[4]	Al	Aluminum
[5]	Si	Silicon
[6]	K	Potassium
[7]	Ca	Calcium
[8]	Ba	Barium
[9]	Fe	Iron

You are given three files:
- discretized_glass_features.csv (Use for Chi-Squared Test)
- glass_features.csv (Use for Information Gain)
- glass_target.csv

Problem Statement:

Find the order of most significant to least significant features using the following feature selection algorithms:

1. Filtering Methods
    - Chi-Squared (Categorical Variables)
    - Information Gain
2. Wrapper-based Method - Recursive Feature Elimination (You may use K-Nearest Neighbor classifier as the underlying classification algorithm)

Contents of Project Report:

1. The report must contain 3 sections each dedicated to a selection algorithm.
2. Each section should have
    a. Order of most significant to least significant features **[2M]**
    b. Value of Chi-squared/Information Gain (in case of filtering methods) or RMSE value with the selected set of features (for wrapper based method) **[2M]**

Submission Details: Analysis of each selection algorithm - id_ass2.docx

Contact Details:

Any queries must be raised in the discussion forum of the corresponding assignment folder only.

1. Himanshu - himanshuchawla437@gmail.com
2. Mohit - mohitrathore15@gmail.com
3. Kartheek - f20160015@hyderabad.bits-pilani.ac.in
4. Shristy - f20160688@hyderabad.bits-pilani.ac.in

a) Agam - f20160484@hyderabad.bits-pilani.ac.in