Machine Learning (BITS F464) - Assignment 2

Fisher linear discriminant, Naïve Bayes and logistic regression

Maximum Marks: 25

Submission Deadline: 0900Hrs 06/07/2018

This assignment comparison of classification algorithms taught in course. You need to compare classifier built using Naïve Bayes, Fisher linear discriminant, and logistic regression.

Dataset Description:
The dataset consist of 39797 instances. An instance is a news article and 61 attributes of each news article has been collected. You can find dataset and description of attributes from http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity . From 61 attributes 58 attributes are predictor attributes, 2 are non-predictor attributes and 1 target attribute. News article's popularity is based on number of times it has been shared. So number of shares is the target attribute. Here as this is a classification task you have to convert number of shares as binary attribute using decision threshold of 1400, i.e. you have to set target value attribute as 1 if number of shares > 1400, otherwise set to 0. Do this conversion before applying any algorithm or pre-processing.

Problem Description:
You have to perform following steps:
- You have to divide the given data into training and testing data as 70-30 split. That is you have to randomly select 70% of data as training data and 30% as testing data.
- On training data you have to build a classification model using inbuilt packages of your preferred language. As describe earlier you have to build models using following algorithms: Naïve Bayes, Fisher linear discriminant, and logistic regression.
  You have to write code in C/C++/Java for implementation of Fisher linear discriminant.
  You may use any of the packages like R / Python for Naïve Bayes, and logistic regression.
- After building the model you have to predict popularity of the news articles in testing data.
- Measure following parameters: recall, precision, accuracy. Description of these parameters can be found from https://en.wikipedia.org/wiki/Precision_and_recall

 Languages Allowed: Any language (preferably R, Python or Matlab)

Evaluation Criteria:
Evaluation will be done based on your approach and result parameters. All group members should have contribution in the assignment. Evaluator can ask anyone to explain any part of the assignment.

Report:
Report should contain following things:
1. ID and names of team members
2. Language used
3. Complete description of the approach taken to solve the problem
4. Results (recall, precision, accuracy) values for different algorithms
5. Conclusion about which algorithm is working better than others and why

Submission should contain zip file of following documents:
1. Source code
2. Report