Prairie View A&M University

ELEG 6163
Statistical Learning for Big Data

Fall 2019 – Project

Choosing one project and complete the requirements.

Project 1: Building a binary classifier with logistic regression to judge if the patient has diabetes.

- **Download the data (Pima Indians Diabetes) (https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f)**
- **Splitting the data into training and testing sets with splitting rate 0.8**
- **Processing missing value on the training and testing sets to generate the preprocessed training set D1 and the preprocessed testing set D2**
- **Selecting 5 features on D_1 with *chi2* to conduct D_3**
- **Building a classifier on D_3**
- **Testing the classifier on D_2 with the 5 selected features and evaluating the testing results with evaluation metrics, namely, precision, recall, and F-score, where the testing means to perform prediction on D_2**
- **Comparing the performance between the model built on the raw data and that built on the preprocessed data and providing some reasons to explain the performance differences**

Project 2: Performance comparison on three classifiers built on the same dataset

- **Download the data (Pima Indians Diabetes) (https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f)**
- **Splitting the data into training and testing sets with splitting rate 0.6**
- **Building 3 classifiers with logistic regression, decision tree, and neural network on the training set**
- **Testing these 3 classifiers on the testing set and evaluating the testing results with evaluation metrics, namely, precision, recall, and F-score**
- **Comparing the performance of these 3 classifiers and providing some reasons to explain the performance differences**

Project 3: Performance comparison on the same classifier (neural network) built on three datasets

- **Download 3 data sets**
  - **The Cleveland Heart Disease Dataset**
    - ✓ **Data Description: https://archive.ics.uci.edu/ml/datasets/Heart+Disease**

    - ✓ Link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
  - Haberman's Survival Data Set
    - ✓ Data Description: https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival
    - ✓ Link: https://archive.ics.uci.edu/ml/machine-learning-databases/haberman/haberman.data
  - Banknote Authentication Data Set
    - ✓ Data Description: https://archive.ics.uci.edu/ml/datasets/banknote+authentication
    - ✓ Link: https://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt
- Splitting the data into training and testing sets with splitting rate 0.6 on these 3 data sets
- Building 3 classifiers with neural network model on these 3 data sets with the same machine learning model
- Testing these 3 classifiers on the testing sets and evaluating the testing results with evaluation metrics, namely, precision, recall, and F-score
- Comparing the performance of different classifiers

Requirements:

- Submitting the source code (.py files) and the data you use for the project, where the codes have no bugs
- Write the summary of the project with the following parts
  - Subtask description
  - Model description
  - Evaluation methods
  - Result analysis
- Presentation with Slides
  - More than 20 slides
  - 10 ~ 15 minutes
  - Subtask description
  - Model description
  - Evaluation methods
  - Result analysis
  - Submit the slides
- **Please submit all required materials (source codes, data, summary, and slides) within one package before the due.**

Due: 12/02/2019