Lab 5
Classification with Machine Learning

CIS492/593 Big Data

Sunnie Chung

Classification with Machine Learning

Designing and Building a Prediction Model with a Machine Learning (ML) Classifier

Choose any ML classifier(s) covered in class and apply to your choice of the data set from the three data sets in the Lab5 section: Adult Data set, Wine Selection Data Set, or Patient Data set for your classification goal as specified for each data set in the Lab5 section.

Plan your experiment with:

1. Determine Data preprocessing methods required for your data set to apply for your classifier
2. Display your result in Confusion Matrix and Calculate in Accuracy, Recall, Precision, MacroF1
3. Do 5-Fold Cross Validation (k= 5) Compare the accuracy of each test of the classifier. Your Overall Accuracy is Ave of the five model accuracy from 5 runs of your classifier.

Data Sets and Description of Classification:

- **Predicting Wine Quality in Scale 1 – 10 From Chemical Properties of Wine**
  http://archive.ics.uci.edu/ml/datasets/Wine+Quality
  http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

- **Predicting Whether or not Income > 50K From an Adult's Census Information**
  http://archive.ics.uci.edu/ml/datasets/Adult
  http://archive.ics.uci.edu/ml/machine-learning-databases/adult/

- **Predicting Whether or not the patient has a breast cancer or not From the Patient data set**

Phases:

1. Determine Data preprocessing methods to apply for each of your classifiers:]
   For example, Discretization for Decision Tree

   Normalization for SVM, Neural Network
   Vectorization of a record for SVM, ANN

2. Design your Data Analytic Experiment with Your Choice of Classifier.
3. Validate your result with your Test Set to compare the Accuracy of your models

Available Platforms:

You can use any data analytic systems/tools of your choice. Some of those systems/tools are in the followings:

- R
  https://www.r-project.org/
  http://www.rdatamining.com/

- Python has the most recent Machine Learning Library and data analytic Algorithms

- SQL Server Analysis Services (SSAS) Data Tools: You can use R in 2016 Data Tool
  https://msdn.microsoft.com/en-us/library/mt604845.aspx
  or Stand Alone R Server
  https://msdn.microsoft.com/en-us/library/mt674874.aspx
  https://msdn.microsoft.com/en-us/library/mt671127.aspx

- Any available Classifiers as Open Source:
  For example, C5 or CART for Decision Tree
  Download C5 and CART at:
  http://www.rulequest.com/see5-info.html
  http://www.salford-systems.com/downloadspm

- Other useful data mining tool sites

  http://www.cs.waikato.ac.nz/~ml/weka/

  http://www.kdnuggets.com/software/classification-decision-tree.html

  http://www.salford-systems.com/downloadspm

Submission:

1. Screen Captures of your Installation/Setting up Procedure and document the related Source info (Which software, Link to the Site, Which Classifier Algorithm, etc).
2. Document your experiments with all the steps for your classifier
3. Document your models if applicable with each the different parameter settings or different transformation methods and the result in Accuracy
4. Report your discussion, observation, findings on Your Results
5. Grade will be based on completion of the required tasks and Accuracy (Performance) of your classifiers