Lab Assignment 4

CIS490/590

Sunnie Chung

Classification with Machine Learning

Designing and Building a Prediction Model for Adult Data with a Machine Learning (ML) Classifier to Predict Whether or not Income > 50K for a given Adult's Census Profile Information in your Training set

Plan your experiment with:

1. Choose Your Classifier and determine Data preprocessing methods required to create a training set to apply 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 5 Accuracy

Phases:

1. Design your Data Analytic Experiment with Your Choice of Classifier.

Choose a ML classifier: Decision Tree (DT), Neural Network (NN) or Support Vector Machine (SVM) Covered in class for Classification.

2. Determine Data preprocessing methods to apply your classifier

Transform the Adult Data Set with Correct Data Preprocessing Method for Your Classifier to Create a Training Set and Test Set with a Class Label.

For example,
Discretization for Decision Tree
Vectorization of a record with Binarization (One Hot Encoding) for SVM
Normalization for Neural Network

3. Validate your result with your Test Set to compare the Accuracy of your models

Data Sets and Description of Classification:

- **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/

You can Change Your Classification (Prediction Goal) and Data Set as You Wish.

Available Platforms:

You can use any data analytic systems/tools of your choice. Some of those systems/tools are in the followings. See Lab 4 on ML Section or Classification Lecture Note Section for More Tools.

- Python has the most recent Machine Learning Library and data analytic Algorithms
  https://scikit-learn.org/stable/
  https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing

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

- 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