Lab Assignment 3

CIS660/EEC 525

Sunnie Chung

Designing and Building a Prediction Model for Bike Buyer Data with a Classifier

Choose any two classifiers covered in class and apply to your Bike Buyer data Set

Plan your experiment with:

1. Determine Data preprocessing methods required to apply for each of your classifiers

2. For each classifier,
   2-1. Compare the accuracy of the classifier with two different sets of input parameters if applicable
   Or

   2-2. Compare the accuracy of the classifier with two different data preprocessing methods.

   Optional for Extra Credit
   2-3. Experiment for Feature Selection with PCA tools or Your Own Experiment (See Below for an example)

3. Compare the accuracy of each test of the classifiers

4. Discuss about your results:
   - Why your inducted model is different for the same training data as you change the parameter values or the classifier.
   - Why a certain parameter setting or a classifier shows with better accuracy than the others that you tried.
   - Anything you observed

Data:

- Use your data VTargetMail from Lab1 that you selected and preprocessed for the Training and Test Sets for Lab3
- For EEC 525 Students, you can choose your data source from any sensor data set or machine generated signals

Phases:

1. Determine Data preprocessing methods to apply for each of your classifiers

For example, Discretization for Decision Tree
Vectorization of a record for SVM
Normalization for Neural Network

2. Design your Data Analytic Experiment with Two different Classifiers of Your Choice.

Choose any two different classifiers covered in class, for example, Decision Tree, Naïve Bayesian, SVM, Neural Network, K Nearest Neighbor, or any other classifier to compare the Accuracy of the results from your classifier.

2-1. Experiment to Find the Best Parameter Setting for your Classifier.

For Example:

Decision Tree Classifier: C5 for GainRatioSplit, CART for GiniSplit on the same set of data with different parameter settings as follow:

- Measure: Entropy, GINI
- Different Minimum Support Thresholds
- Different Complex Penalty Degrees on the Number of Splits

Neural Network:

Test with two Different Topologies: The number units of a hidden layer, The number of hidden layers

SVM: Test with different Kernel functions

K Nearest Neighbor: Test with two different K values and distance metrics

Or alternatively

2-2. For Naïve Bayes, NN or SVM, Experiment with two different Data Transformation Methods

For Continuous and numeric Attributes,

1) Data set as floating point without Discretization and Binarization
2) Data set with Discretization and Binarization

For Extra Credit

2-3 Experiment for Feature Selection with either

1) Feature Significance Analysis with PCA tools

Or

2) Your Own Experiment as follow:

Simple Experiment for Feature Selection Methodology

2-2-1. Pick the best parameter setting and data transformation from Phase 2.

2-2-2. Apply Your Classifier with the best parameters set to each different feature sets from your input file to see if there is any significant difference in the result for each iteration. (See Below for an example)

3. Validate your result with your Test Set to compare the Accuracy of your models for each classifier with different Parameter settings or different transformation method.
4. Discuss about your results:
    - Why your inducted model is different for the same training data as you change the parameter values or the classifier.
    - Why a classifier shows better accuracy than the others for a certain parameter setting or with a different transformation method.
    - Any observations you made
5. Extra Credit to Anyone that Gets Best Top 5 Accuracy of the Class

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

Extra Credit: Experiment for Feature Selection with either

Feature Significance Analysis with PCA tools
(See the PCA Example in Lecture Note Section for this)
Or
Your Own Experiment as follow:
Simple Experiment for Feature Selection Methodology

1. Simple Experiment for Feature Selection Methodology to choose the best feature set:
   1-1 Pick the best Model with the best parameter setting from Phase 1 and 2.
   1-2 Apply your Model with the best parameters to different input sets (created with different combinations of feature sets from your VTargetMail input file to see if there are any significant differences in the result of each feature set in terms of Accuracy.

   See Lecture Note Slide 75 on DW or Slide 34 in J Han's Chap 3 Data Exploration as below:

How to Determine the Prediction Power of an Attribute?

- Ex. A customer table **D:**
  - Two dimensions **Z**: *Time* (*Month*, *Year*) and *Location* (*State, Country*)
  - Two features **X**: *Gender* and *Salary*
  - One class-label attribute Y: *Valued Customer*
- Q: "Are there times and locations in which the value of a customer depended greatly on the customers gender (i.e., Gender: predictiveness attribute **V**)?"
- Idea:
  - Compute the difference between the model built on that using **X** to predict Y and that built on using **X – V** to predict Y
  - If the difference is large, **V** must play an important role at predicting Y

56

Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

34

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 each 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
6. Put a Note in the Cover Page if you did the Extra Credit Experiment in your Lab3 Report.