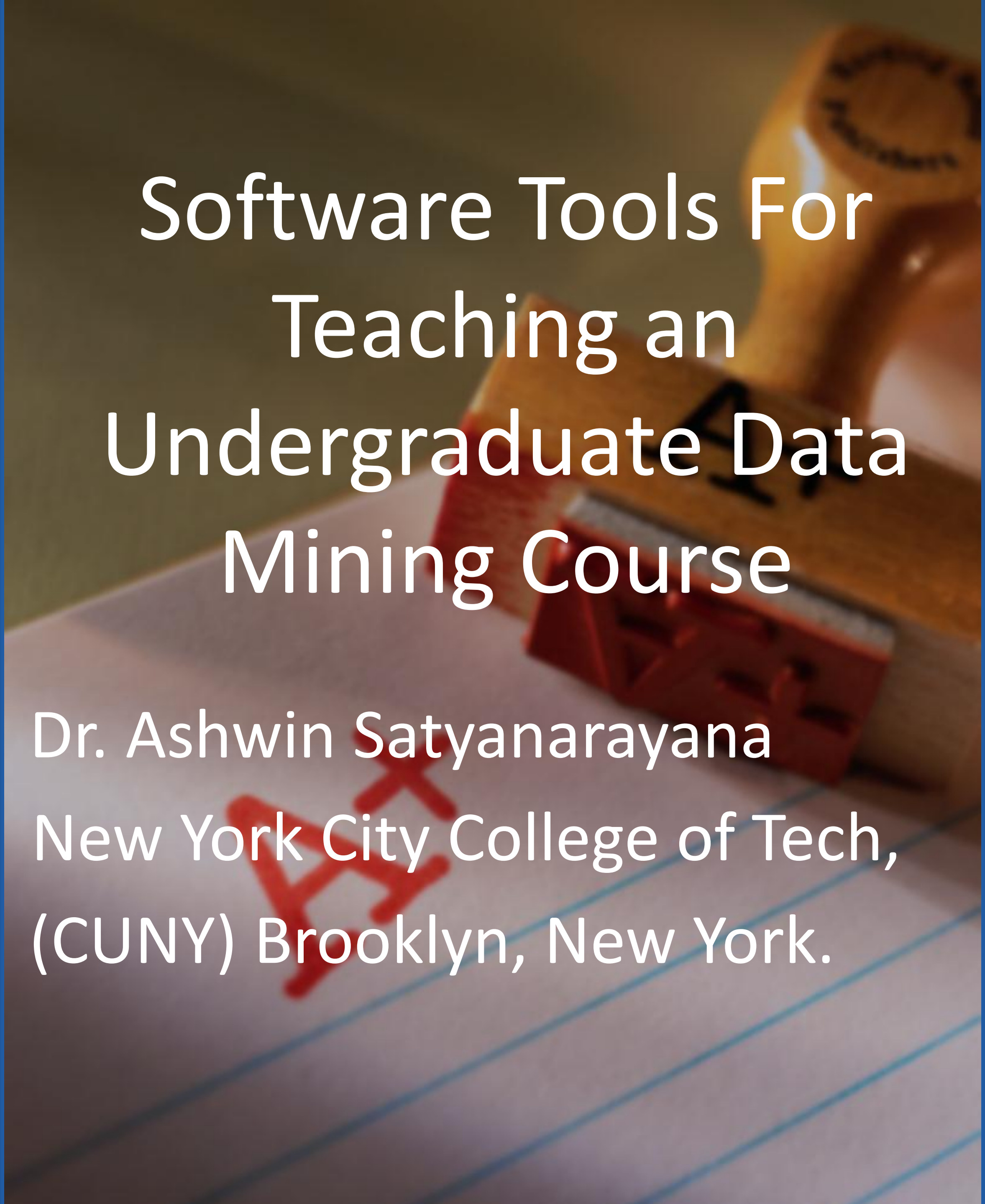


Software Tools For Teaching an Undergraduate Data Mining Course

Author: Dr. Ashwin Satyanarayana

Organization: New York City College of Tech, (CUNY) Brooklyn, New York.

Introduction & Motivation

- Enormous amounts of data are generated every minute. Some sources of data, such as those found on the Internet are obvious.
  - Social networking sites,
  - search and retrieval engines,
  - media sharing sites,
  - stock trading sites, and
  - news sources
- **Data mining**, a growingly popular field in Computer Science, is the transformation of *"large amounts of data"* into meaningful patterns and rules.
- Recent studies have noted the rise of data mining as a career path with increasing opportunities for graduates.

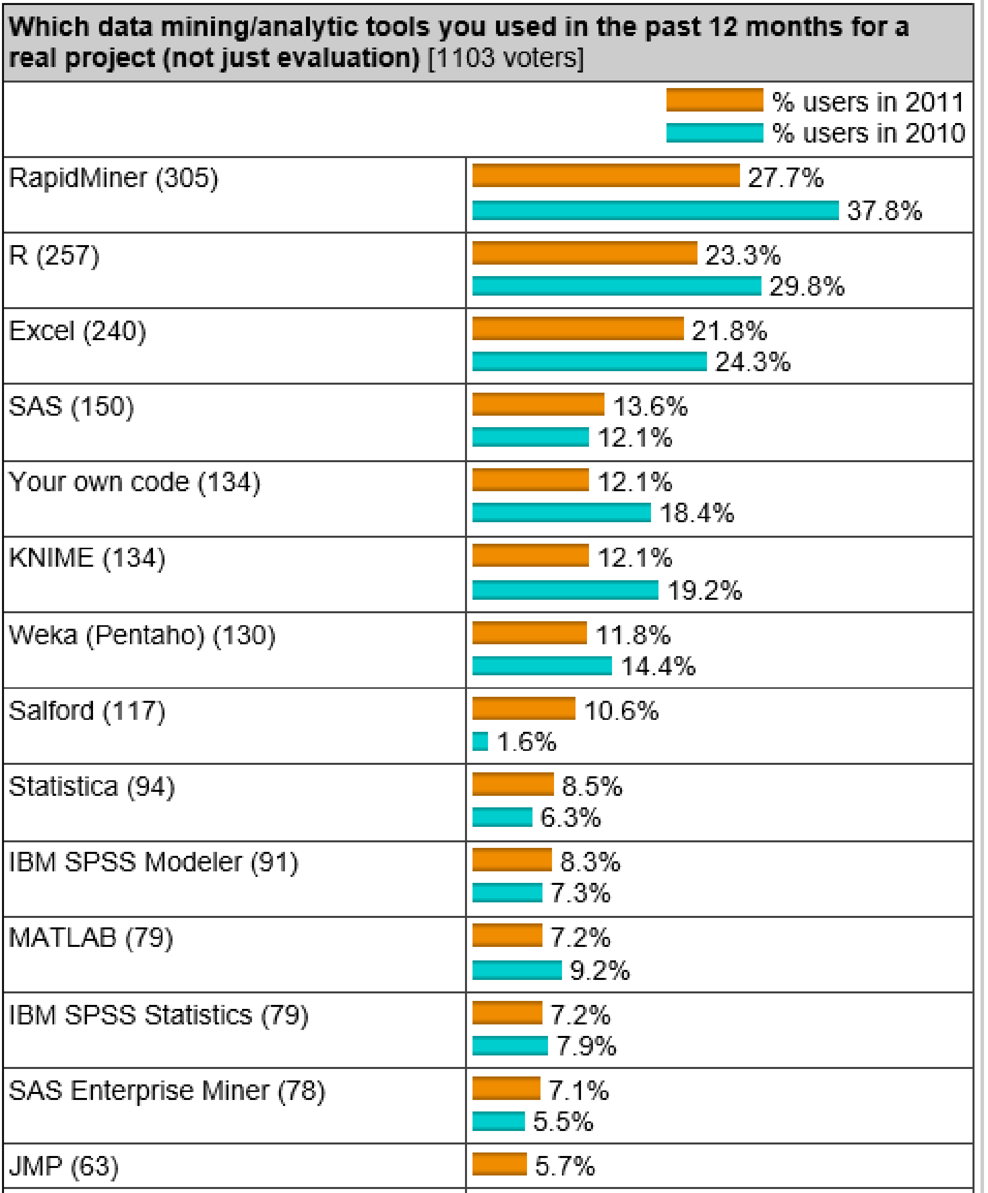
Introduction & Motivation

- We are in a new era in modern information technology - the "Big Data" era. In March, 2012, the U.S. Government announced a "Big Data Research and Development Initaitve" -- a **$200 million dollar** commitment to improve our ability to *"extract knowledge and insights from large and complex collections of digital data."*
- Government agencies such as **NSF, NIH, and DoD** are investing hundreds of millions of dollars toward the development of systems that can help them extract knowledge from their data.
- The career potential for our graduates continue to blossom in this field. A recent study released by Gartner projects that in 2013, ***"big data is forecast to drive $34 billion of IT spending,"*** with a total of $232 billion to be spent through 2016

Challenges for Faculty

- Big Data Sources
- Tools
  - Easy to Download
  - Cheap
  - Fast Learning Curve (for a one semester course)
  - Should work well with Real World "Big Data"
- Commercial or Open Source

Tools: KDNuggets Poll (2011)



Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [1103 voters]

	% users in 2011	% users in 2010
RapidMiner (305)	27.7%	37.8%
R (257)	23.3%	29.8%
Excel (240)	21.8%	24.3%
SAS (150)	13.6%	12.1%
Your own code (134)	12.1%	18.4%
KNIME (134)	12.1%	19.2%
Weka (Pentaho) (130)	11.8%	14.4%
Salford (117)	10.6%	1.6%
Statistica (94)	8.5%	6.3%
IBM SPSS Modeler (91)	8.3%	7.3%
MATLAB (79)	7.2%	9.2%
IBM SPSS Statistics (79)	7.2%	7.9%
SAS Enterprise Miner (78)	7.1%	5.5%
JMP (63)	5.7%	

Commercial Tools

- Pros:
  - Students get Hands On experience with working with real tools used in the industry
- Cons:
  - Expensive
  - Not easily accessible for Students to download at home
- In this paper we discuss 3 commercial tools:
  - SAS Enterprise Miner
  - MATLAB
  - IBM SPSS Modeler

Open Source Tools

- Pros:
  - Free and easy to download for Students
  - Fast learning curve
  - Free manuals and study guides available
- Cons:
  - On Hands On experience with industry tools
- In this paper we will discuss 3 open source tools:
  - RapidMiner,
  - R and
  - WEKA

Commercial Tools

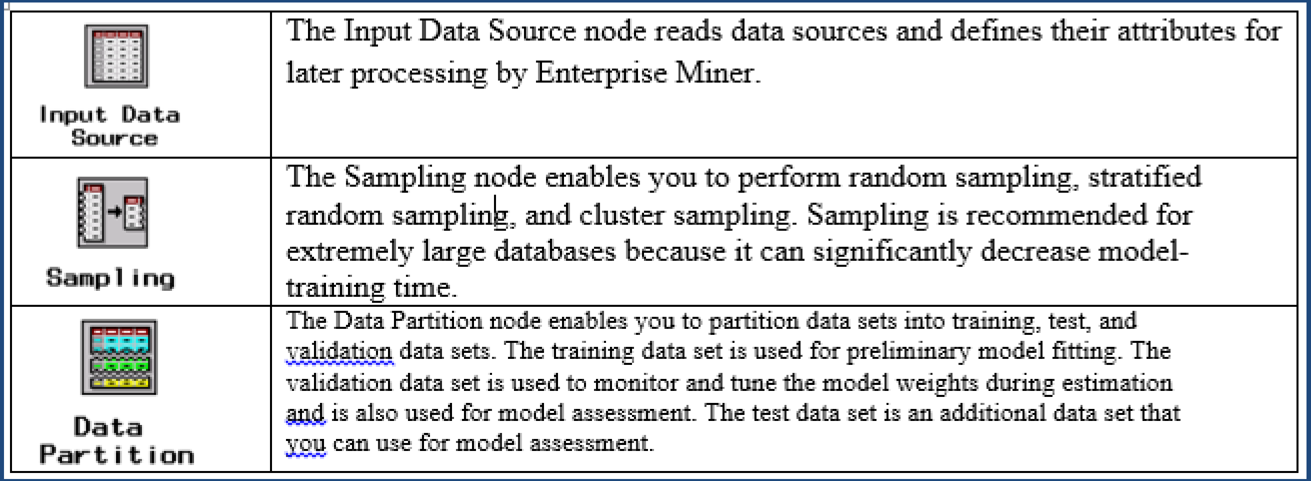
1. SAS Enterprise Miner
2. IBM SPSS Modeler
3. MATLAB

SAS Enterprise Miner

- Interactive Statistical and Visualization tools
- Easy to find trends and anamolies
- Focus on Model Development Process
- Nodes are arranged into the following categories according to SEMMA:
  - Sample *(Identify input data, sample, partition, etc)*
  - Explore *(plot the data, association analysis, etc)*
  - Modify *(prepare the data for analysis)*
  - Model *(fit the predictive model)*
  - Assess *(compare competing predictive models)*

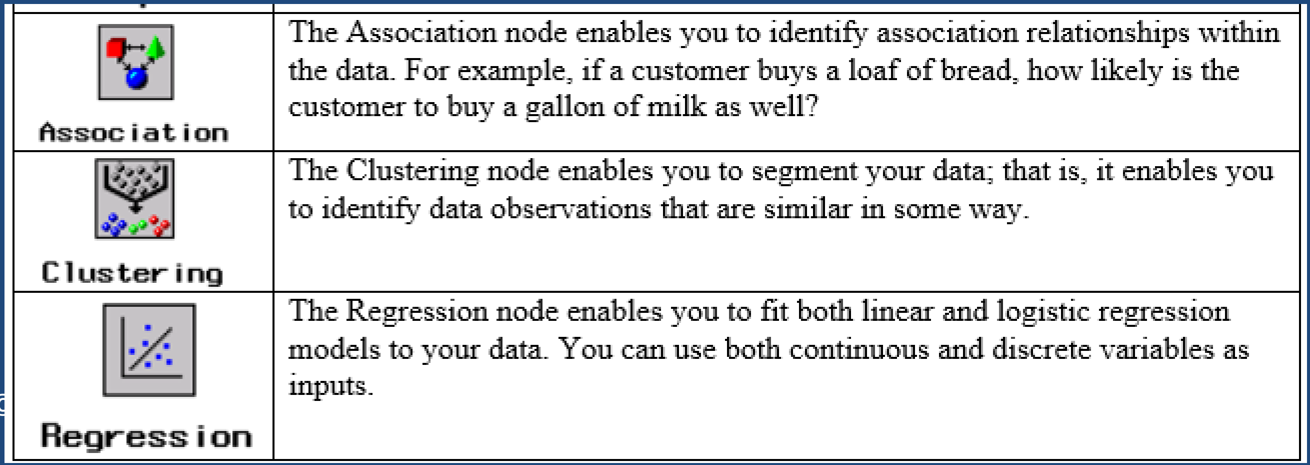
SAS Enterprise Miner (…contd)

- Sample Nodes:



Input Data Source	The Input Data Source node reads data sources and defines their attributes for later processing by Enterprise Miner.
Sampling	The Sampling node enables you to perform random sampling, stratified random sampling, and cluster sampling. Sampling is recommended for extremely large databases because it can significantly decrease model-training time.
Data Partition	The Data Partition node enables you to partition data sets into training, test, and validation data sets. The training data set is used for preliminary model fitting. The validation data set is used to monitor and tune the model weights during estimation and is also used for model assessment. The test data set is an additional data set that you can use for model assessment.

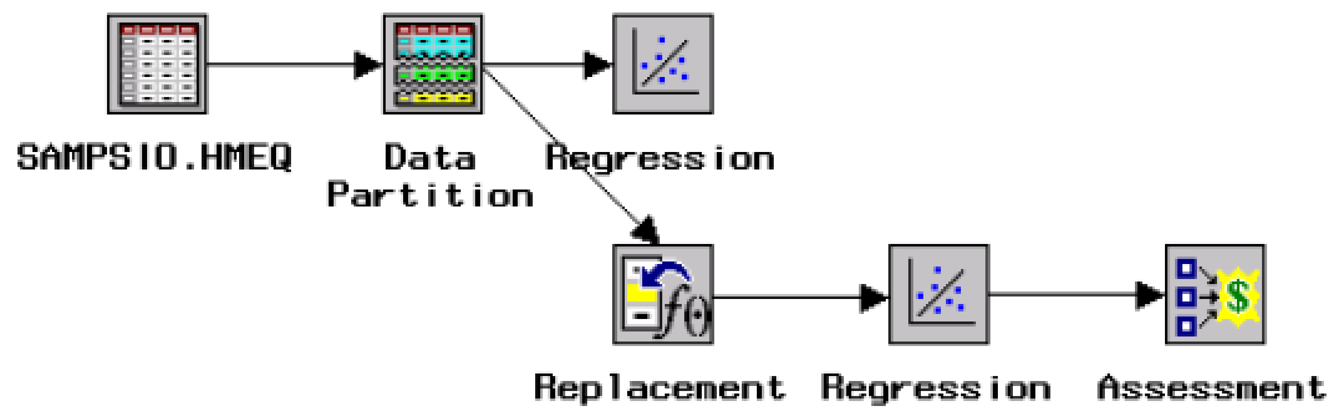
- Model Nodes:



Association	The Association node enables you to identify association relationships within the data. For example, if a customer buys a loaf of bread, how likely is the customer to buy a gallon of milk as well?
Clustering	The Clustering node enables you to segment your data; that is, it enables you to identify data observations that are similar in some way.
Regression	The Regression node enables you to fit both linear and logistic regression models to your data. You can use both continuous and discrete variables as inputs.

SAS Enterprise Miner (...contd)

- Combine Nodes and Create a Model:



- In Summary, SAS Enterprise Miner plays with different aspects of Data Mining:
  - Data Preparation and Investigation,
  - Fitting and Comparing Models and
  - Generating reporting

IBM SPSS Modeler

- Data Mining Software application by IBM
- Visual Interface which allows users to leverage statistical and data mining algorithms without programming
- Offers a variety of modeling methods taken from machine learning, artificial intelligence, and statistics.

IBM SPSS Modeler (...contd)

- Modeling Methods are divided into 3 categories:
  - Classification Models: use the values of one or more **input** fields to predict the value of one or more output, or **target**, fields. (Decision trees, regression, neural networks, etc)

CRT	**Classification and Regression Tree**
QUEST	Quest Node – Binary Classification method for buiding trees

IBM SPSS Modeler (...contd)

– Association Models: Finds patterns in your data where one or more entities are associated with one or more entities.

	Apriori Node
	Sequence Node

– Segmentation Models: Divides the data into segments or clusters of records that have similar patterns.

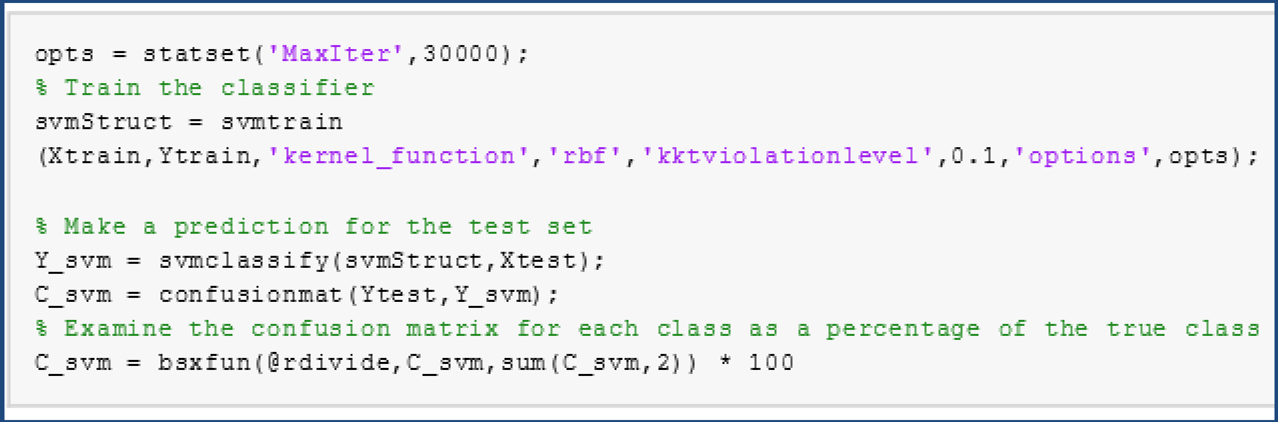
K	**K-Means Node**
	Kohonen Node

MATLAB

- "MATLAB has excellent built-in support for many data analysis routines," in particular, one of its most useful facilities is that of efficient exploratory data analysis which is a natural fit in the context of data mining.
- Pros of MATLAB:
  - **Portability:** All MATLAB users have the same range of basic functions at their disposal.
  - **Representing all data in the form of Matrices:** Allows varied algorithmic implementations which are crucial for data mining.

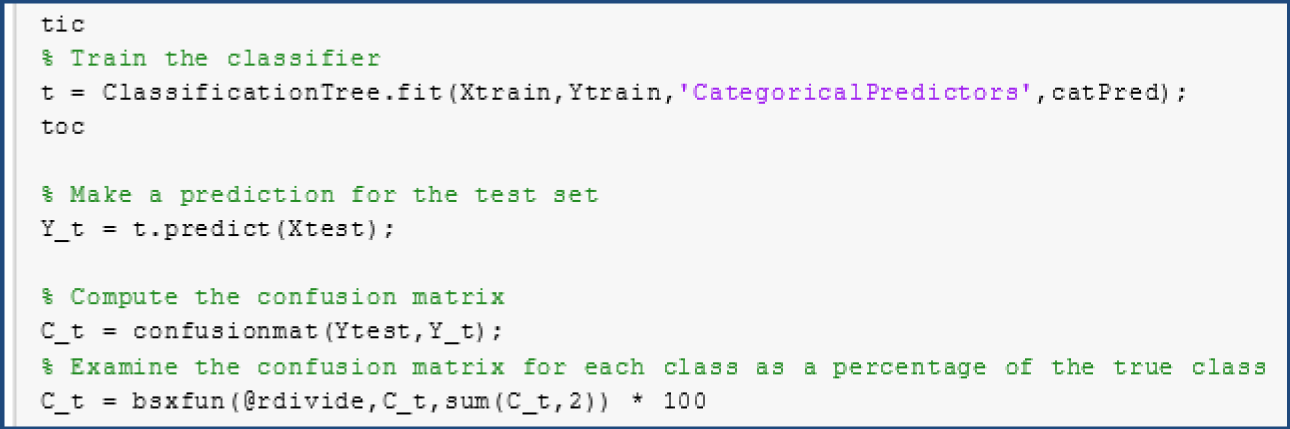
MATLAB (...contd)

- Support Vector Machines:



```
opts = statset('MaxIter',30000);
% Train the classifier
svmStruct = svmtrain
(Xtrain,Ytrain,'kernel_function','rbf','kktviolationlevel',0.1,'options',opts);

% Make a prediction for the test set
Y_svm = svmclassify(svmStruct,Xtest);
C_svm = confusionmat(Ytest,Y_svm);
% Examine the confusion matrix for each class as a percentage of the true class
C_svm = bsxfun(@rdivide,C_svm,sum(C_svm,2)) * 100
```

- Decision Trees:



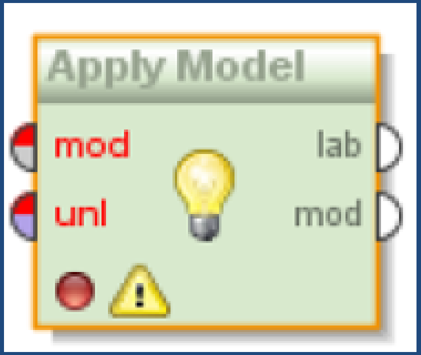
```
tic
% Train the classifier
t = ClassificationTree.fit(Xtrain,Ytrain,'CategoricalPredictors',catPred);
toc

% Make a prediction for the test set
Y_t = t.predict(Xtest);

% Compute the confusion matrix
C_t = confusionmat(Ytest,Y_t);
% Examine the confusion matrix for each class as a percentage of the true class
C_t = bsxfun(@rdivide,C_t,sum(C_t,2)) * 100
```

Open Source Tools

1. Rapid Miner
2. R
3. WEKA

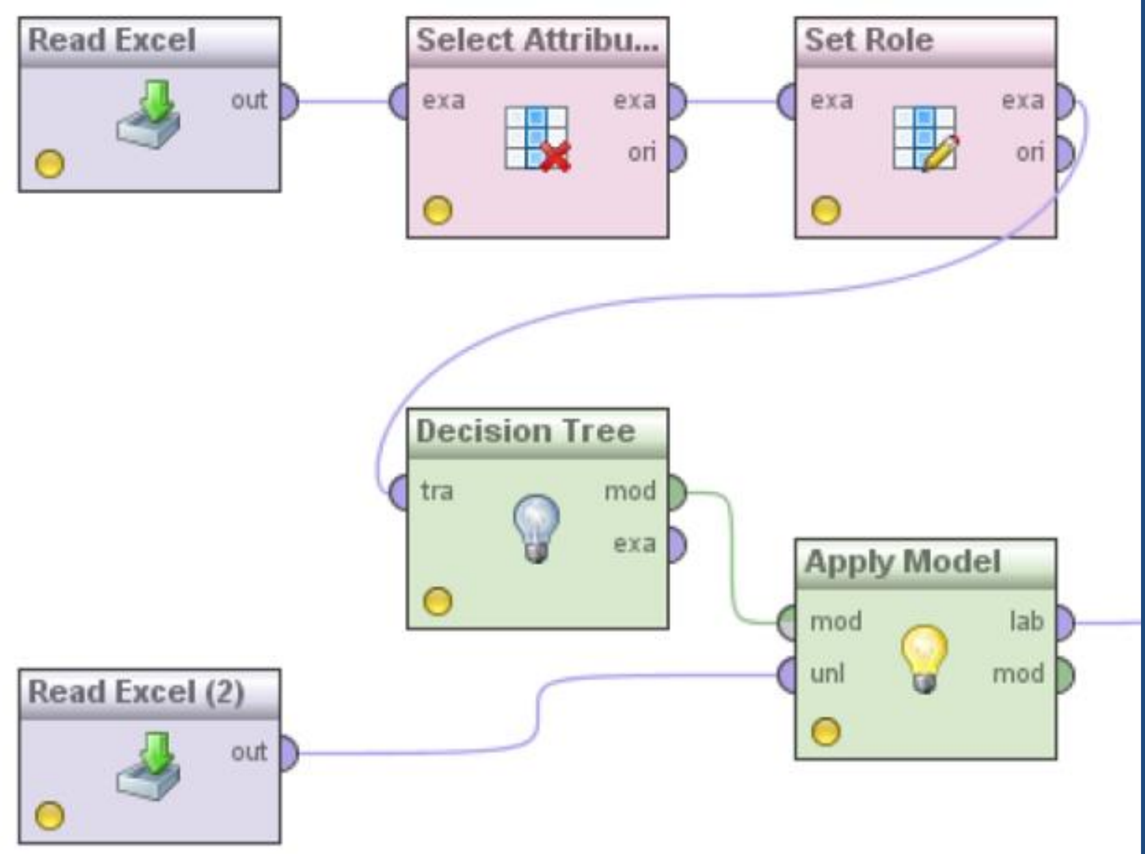
Rapid Miner

- Rapid Miner (Formerly Yale) is an environment for machine learning and data mining processes. World Wide Leading open source data mining solution.
- A modular operator concept allows the design of complex nested operator chains for a huge number of learning problems.



Rapid Miner (….contd)

If several operators are interconnected, then we speak of an analysis process or process for short. Such a succession of steps can for example load a data set, transform the data, compute a model and apply the model to another data set.



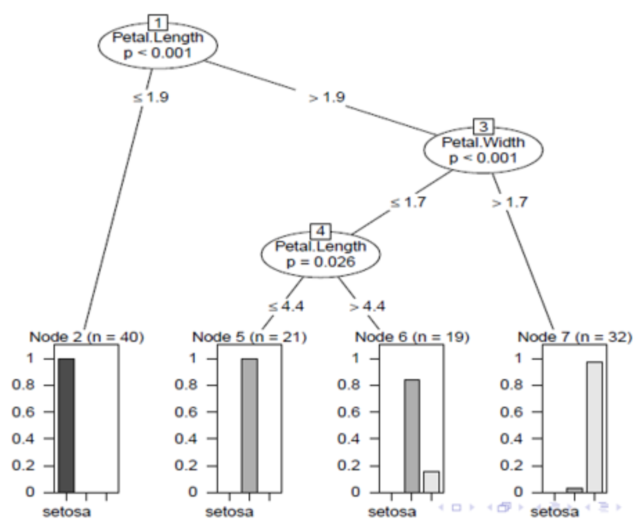
R

- R is a free software environment for statistical computing and graphics. [http://www.r-project.org].
- R can be easily extended with 4,728 packages available on CRAN(as of Sept 6, 2013).
- Data mining tasks (Classification, Clustering and Association Rules) can be easily performed using built in R commands.

R (...contd)

- Classification

Decision Trees	*rparty, party*
Random Forest	*randomForest, party*
SVM	*e1071, kernlab*
Neural Network	*nnet, neuralnet, RSNNS*



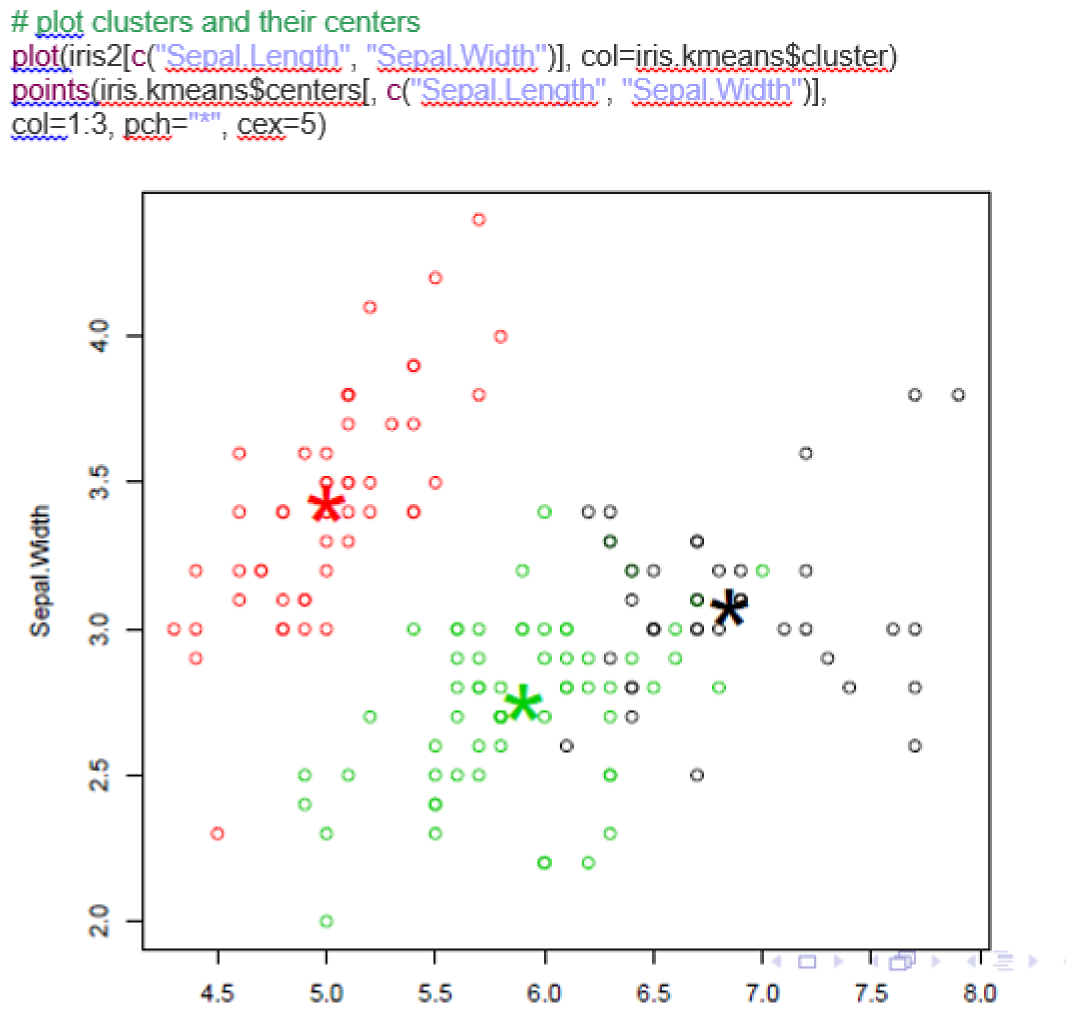
- # build a decision tree

```
library(party)
iris.formula <- Species ~ Sepal.Length + Sepal.Width +
Petal.Length + Petal.Width
iris.ctree <- ctree(iris.formula, data=iris.train)
```

R (....contd)

– Clustering

k-means	kmeans(), kmeansruns()
k-mediods	pam(), pamk()
Hierarchical clustering	hcust(), agnes(), diana()
BIRCH	Birch



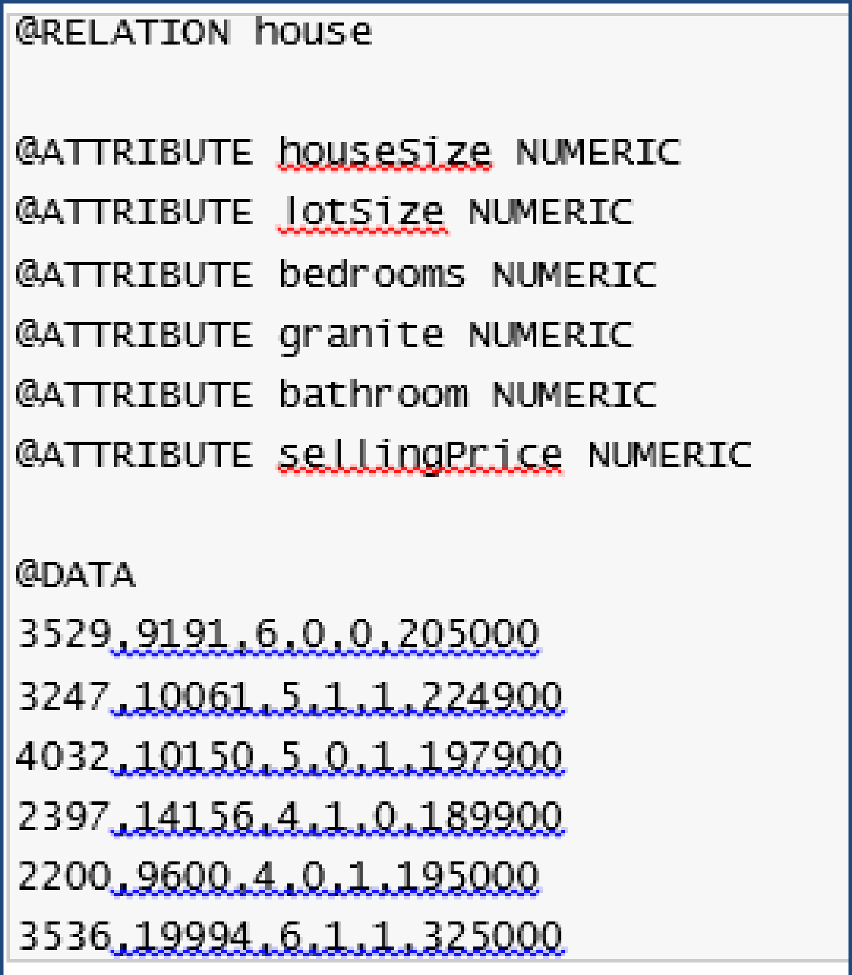
```
# plot clusters and their centers
plot(iris2[c("Sepal.Length", "Sepal.Width")], col=iris.kmeans$cluster)
points(iris.kmeans$centers[, c("Sepal.Length", "Sepal.Width")],
col=1:3, pch="*", cex=5)
```

WEKA

- WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997.
- The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves).
- It also has a general API, so you can embed WEKA, like any other library, in your own applications to such things as automated server-side data-mining tasks.

WEKA (….contd)

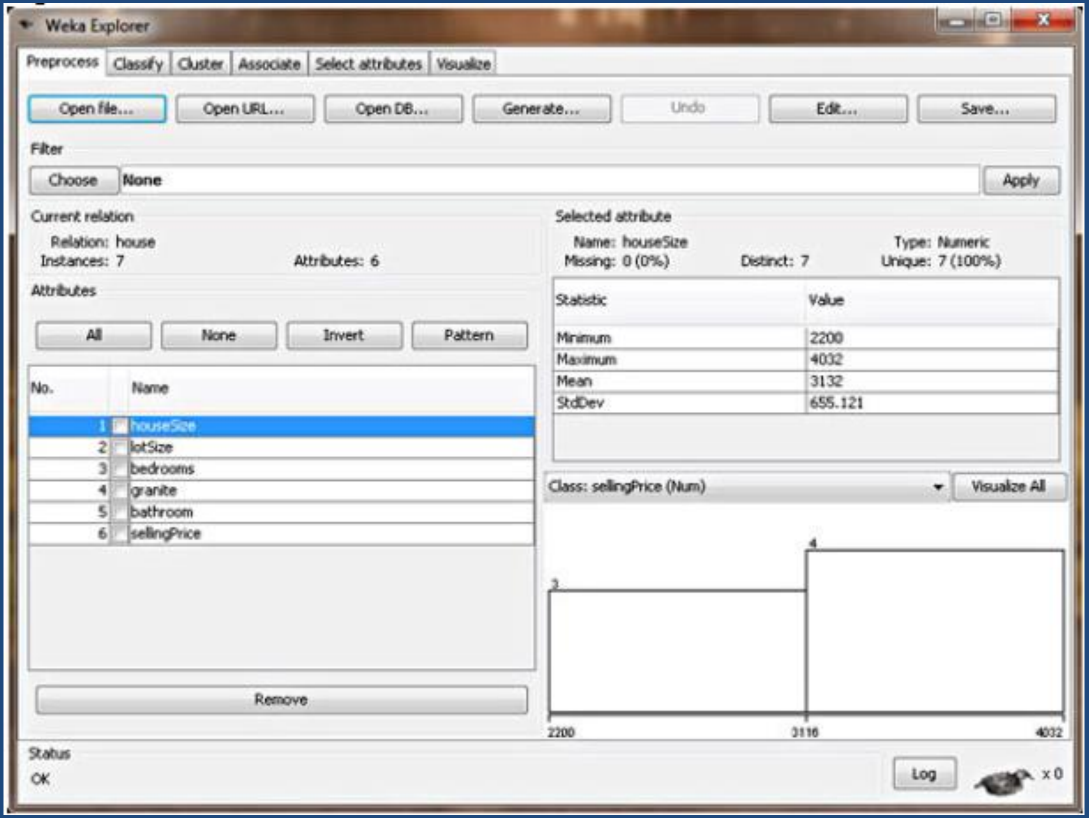
- Building a dataset:
  - To load data into WEKA, we have to put it into a format that will be understood. WEKA's preferred method for loading data is in the Attribute-Relation File Format (ARFF), where you can define the type of data being loaded, then supply the data itself.



```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
```

WEKA (….contd)

- Loading the dataset into WEKA:
  - Click Explorer -> Preprocess -> Open File
  - Now select the ARFF file you created.



Filter: None

Current relation

Relation: house

Instances: 7 Attributes: 6

Attributes

No.	Name
1	houseSize
2	lotSize
3	bedrooms
4	granite
5	bathroom
6	sellingPrice

Selected attribute

Name: houseSize Type: Numeric

Missing: 0 (0%) Distinct: 7 Unique: 7 (100%)

Statistic	Value
Minimum	2200
Maximum	4032
Mean	3132
StdDev	655.121

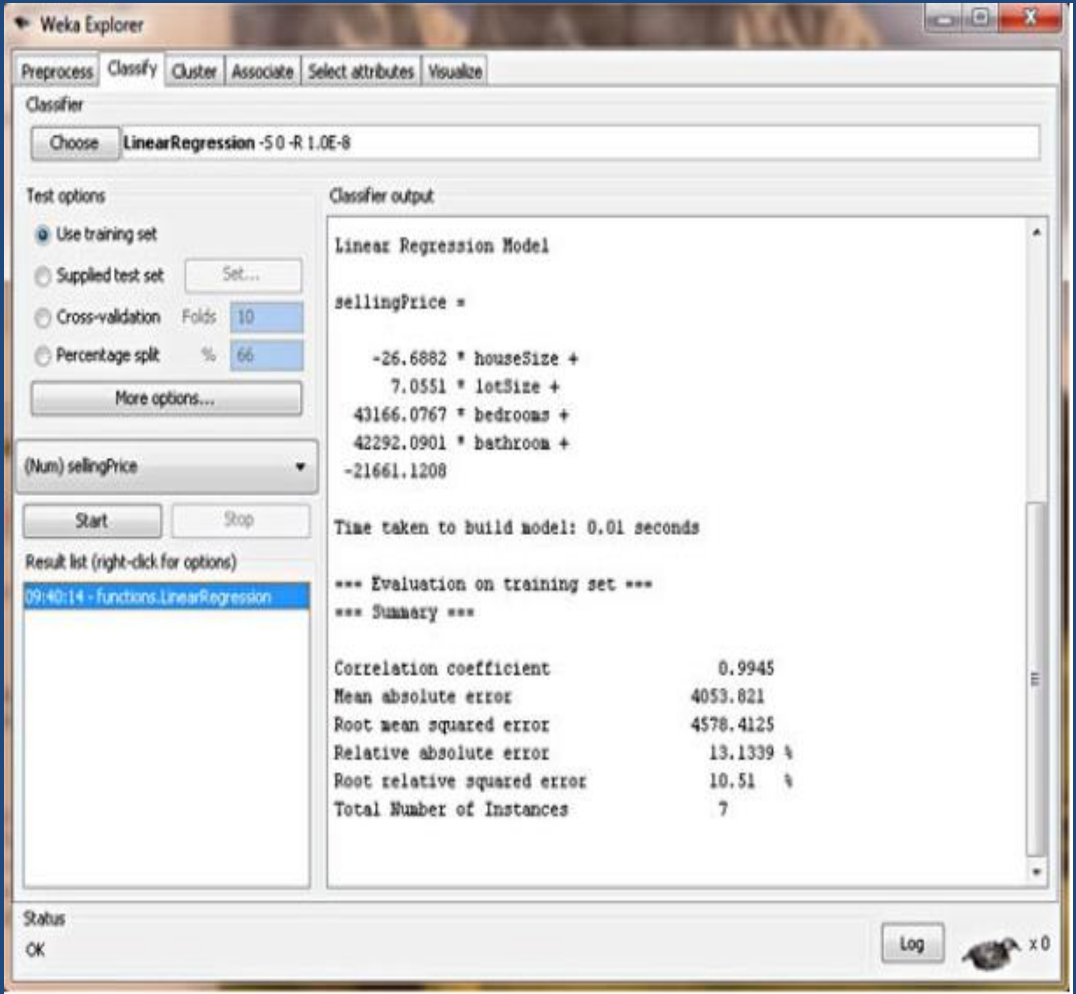
Class: sellingPrice (Num)

Status

OK

WEKA (...contd)

- Creating the model:
  - *1. Click the **Choose** button, then expand the **functions** branch.*
  - *2. Select the **LinearRegression** leaf.*
  - *3. Click Start*



Classifier: **LinearRegression** -S 0 -R 1.0E-8

Test options:
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Num) sellingPrice

Result list (right-click for options)
09:40:14 - functions.LinearRegression

Classifier output

```
Linear Regression Model

sellingPrice =

    -26.6882 * houseSize +
      7.0551 * lotSize +
  43166.0767 * bedrooms +
  42292.0901 * bathroom +
 -21661.1208

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient                  0.9945
Mean absolute error                   4053.821
Root mean squared error               4578.4125
Relative absolute error                 13.1339 %
Root relative squared error             10.51   %
Total Number of Instances                7
```

Status
OK

Conclusion

Name of the Tool	Data Mining Tasks	Visualization	Programming Needed?
1. SAS Enterprise Miner	Yes	Yes	No
2. IBM SPSS Modeler	Yes	Yes	No
3. MATLAB	Yes	No	Yes
4. RapidMiner	Yes	Yes	No
5. R	Yes	No	Yes
6. Weka	Yes	Yes	Yes

Acknowledgments

- PSC CUNY Grant
- Dr. Brian King
  - “Teaching Data Mining in the era of Big Data”
    - ASEE Annual Conference 2013, Atlanta, Georgia
- Wife - Nithya