

Birla Institute of Technology and Science, Pilani

CS C415/IS C415 Data Mining

Lab 1

Introduction and Data Preprocessing

Objectives

- Getting started with IBM SPSS Modeler
- Data Preprocessing

Getting started with IBM SPSS Modeler

The IBM SPSS Modeler is a data mining, modeling and reporting tool. It provides a nice GUI to carry out all the data mining tasks in form of Nodes and Stream Flows. Nodes are the icons or shapes that represent individual operations on the data. The nodes are linked together in a stream to represent the flow of data through each operation i.e. A set of actions (reading in, preprocessing, classification/association rule mining/clustering, reporting, etc) on some input data is called a stream.

Modeler Interface -

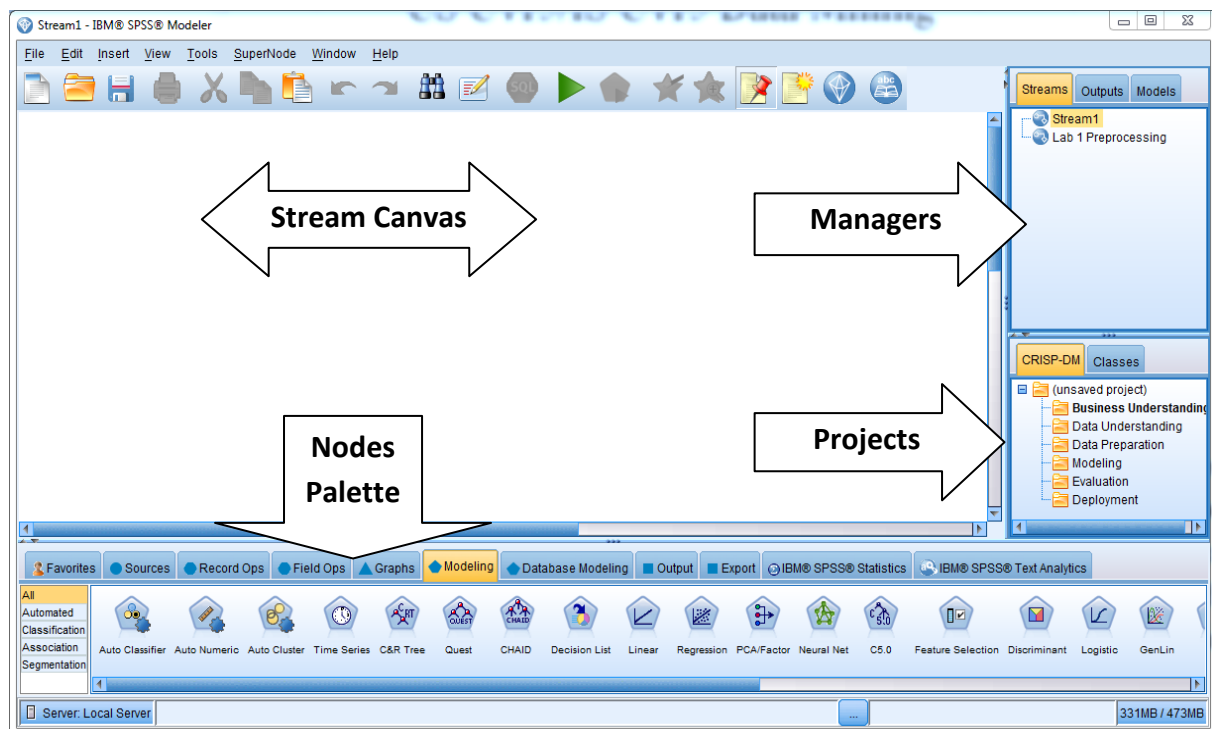


Fig 1 - SPSS Modeler Interface with main components

Modeling -

A model is a set of rules, formulas, or equations that can be used to predict an outcome based on a set of input fields or variables. For example, a financial institution might use a

model to predict whether loan applicants are likely to be good or bad risks, based on information that is already known about past applicants.

To build a stream that will create a model, we need at least three elements:

- A source node that reads in data from some external source.
- A modeling node(classification, association, clustering, etc) that generates a model nugget when the stream is run.
- [Optional] An output node if we want results in tabular or graphical form.

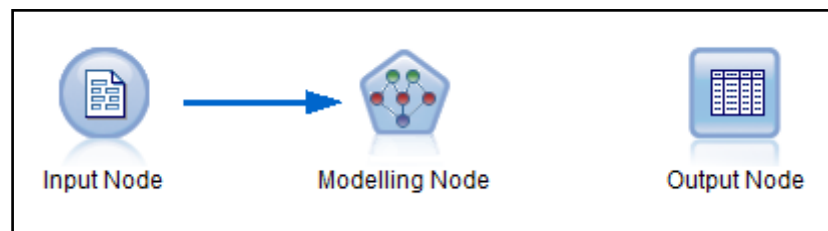







Fig 2 - An abstract stream



Source Nodes






Some important data source nodes are-

Symbol	Node type	Imports data from
	Database Node	MS SQL Server, DB2, Oracle (using ODBC)
	Variable File Node	Delimited text data (*.csv files)
	Excel Node	Microsoft Excel
	XML Node	XML files
	User Input Node	Generate synthetic data

Record Operation Nodes







Record operations nodes are used in data understanding and data preparation. Some important record operation nodes are-

Symbol	Node type	Function
	Select Node	Selects or discards a subset of records from the data stream based on a specific condition
		<i>Display all records having an attribute value above a threshold</i>
	Sample Node	Selects a subset of records using a sample type like random or fixed method
		<i>Display every fifth record</i>

	Aggregate Node	Replaces a sequence of input records with summarized output records
		<i>Find class-wise mean and standard deviation of all records</i>
	Sort Node	Sorts record into ascending or descending order based on values of one or more fields
	Merge Node	Takes multiple input records from different sources and creates a single output record containing some or all of input fields
	Append Node	Concatenates sets of records
	Distinct Node	Removes duplicate records


Field Operation Nodes




These nodes are used to select, clean, or construct data in preparation for analysis. Some important field operation nodes are-

Symbol	Node type	Function
	Type Node	Specifies field metadata and properties. Eg - measurement level (continuous, nominal, ordinal, or flag) for each field can be specified, options for handling missing values and system nulls can be set.
	Filter Node	Filters(discards) fields, rename fields, and maps fields from one source node to another
	Derive Node	Modifies data values or creates new fields from one or more existing fields
		<i>Create a new field as the multiplication of two continuous fields</i>
	Filler Node	Replaces field values and changes storage (replace all blank values with a specific value)
	Binning Node	Creates new nominal fields based on the values of one or more existing continuous fields.
	Partition Node	Generates a partition field, which splits the data into separate subsets for the training, testing, and validation stages of model building.

Output Nodes




Output nodes provide the means to obtain information about data and models. Some important output nodes are-

Symbol	Node type	Function
	Table Node	Displays the data in table format, which can also be written to a file

	Matrix Node	Creates a table that shows relationships between fields
	Analysis Node	Performs various comparisons between predicted values and actual values for one or more model nuggets
	Data Audit Node	Provides a comprehensive first look at the data, including summary statistics, histograms and distribution for each field, as well as information on outliers, missing values, and extremes.




Graph Nodes

These nodes are used for visualizing the data in a mathematical form. Some important graph nodes are-

Symbol	Node type	Function
	Graphboard Node	Offers many different types of graphs in one single node.
	Plot Node	Shows the relationship between numeric fields.
	Web Node	Illustrates the strength of the relationship between values of two or more symbolic (categorical) fields.

Export Nodes

These nodes provide a mechanism for exporting data in various formats to interface with other software tools. Some important export nodes are-

Symbol	Node type	Function
	Database Export Node	Writes data to an ODBC-compliant relational data source.
	Flat File Export Node	Outputs data to a delimited text file.
	Excel Export Node	Outputs data in Microsoft Excel Format (*.xls)

Data Preprocessing

Input Files-

Data file => iris_data.csv
Metadata file => iris_metadata.txt

Reading Data

1. Read the given metadata file and find out about the number of fields and their types (i.e. continuous, nominal, etc), also have a look at input data file. Notice, field names are not mentioned.

2. Drag-n-drop a *Variable File Node* (from Sources tab in Nodes palette) on to *Modeler Canvas*. Double-click on *Variable File Node* and go to "File" tab. Browse and select the given input data file.
3. Uncheck "Read field names from file", and check "Specify number of fields" and set the field count value according to information given in metadata file.
4. Go to "Data" tab and make sure that you are satisfied with the "Storage" type of each field. Otherwise, check "Override" and change the storage type.
5. Go to "Filter" tab and change the output field names according to the names mentioned in metadata file.
6. Go to "Types" and make sure "Measurement" of each field is correct (i.e. continuous, nominal, etc)
7. [Optional] To know about range of values that each field have in this data, click on "Values" cell in front of the field, select "<Read>" and click on "Read Values" button.
8. [Optional] To mark a field as Classifier field (on which classification has to be performed), mark it as "Target" in "Role" cell.
9. Click "Apply" and "OK". Now, a source node with your input file name must appear on the modeler canvas.
10. [Optional] To view the output of this source node, add a "Table node" , "Connect" it to source node and "Run" the stream.

Running a stream

- To run whole stream, use the "Run Stream" button in toolbar
- To run down-stream from a particular node, right click on it and select "Run From Here"

Connecting Node N_1 to N_2

- Right-click on N_1 , select "Connect" and then select N_2 , or
- Select N_1 , press "F2" and then select N_2 .

Analyzing Input

1. To understand the input data, use "Data Audit Node" from "Output" tab in node palette and connect it to the source node.
2. Double-click on data audit node and go to "Settings" tab.
3. Either select "Default" for auditing all fields or select "Use custom fields" and select required fields.
4. Check all options in "Display".
5. Go to "Quality" tab and select "Outlier and Extreme Detection method". Specify the values for outliers and extremes.
6. Click "Apply" and "OK". Now, a data audit node will appear on the modeler canvas. "Run" the stream. (Ctrl+E)
7. Observe output ("Audit" and "Quality" tabs) carefully and try to determine necessity of preprocessing.

From "Audit" tab, determine answers of following questions.

- a. Is cardinality of nominal attributes correct?
 - Eg- cardinality of Boolean attributes should be 2
 - See under column labeled as "Unique"
- b. Are Min, Max, Std. deviation, variance, etc. are under permissible limits?
- c. Does input data needs Sampling (Is it too large)?
- d. How values are distributed for each attribute?
 - See the graphs

From "Quality" tab, determine answers of following questions.

- a. Are there any "Outliers" and/or "Extreme" values?
- b. Are there any "Missing/Null" values?

Handling Missing Values

Remove records having a NULL value

1. Select a "Select Node" from "Record Ops" tab in nodes palette.
2. Connect source node to the select node.
3. Double click on select node and go to "Settings" tab.
4. In order to discard records with missing values select "Discard" and write the expression which selects such records. Expressions can be easily written using Expression Builder Tool.
 - a. Click on "Expression Builder" button. Left hand side lists all functions with their return values and right hand side lists fields in the input dataset. Connectors specified in middle are used to connect two expressions.
 - b. To select all records with missing Field1 value, select "@NULL" function from function list and insert it into expression. Select Field1 from field list and insert it as argument of above mentioned function. "@NULL" is used for numeric values only, for string values use "STRING1 matches STRING2" and specify STRING2 as "".
 - c. To select all records with any one of the missing field values, write expressions as explained above, joined by "OR" operation.
 - d. Verify the expression by clicking "Check". Remove syntax error, if any. Click "OK".
5. Click "Apply" and "OK". Now, a select node will appear on the modeler canvas.
6. [Optional] To view the output of this select node, add a "Table node", "Connect" select node to the table node and "Run" the stream from the select node.

Replace Numeric attributes by mean value


1. To obtain the mean of each field, connect a "Set Globals" node to the stream. Select required fields and operations (Mean, Sum, Min, Max and SDev). Click Run.
1. Select a "Filler Node" from "Field Ops" tab in nodes palette.
2. Connect source node to this filler node.
3. Double click on filler node and go to "Settings" tab.




4. Select all numeric fields in *"Fill in fields"*.
5. Select *"Replace"* condition as *"Blank and null values"*.
6. In *"Replace with"* box, open the *Expression Builder* window. Select *Globals* in right-hand side panel and select Global Mean of appropriate fields. In case of multiple fields selection use @FIELD keyword. Write following expression in replace with box -

$$@GLOBAL_MEAN(@FIELD)$$
7. Click *"Apply"* and *"OK"*. Now, a filler node will be appear on the modeler canvas.

Remove Nominal attributes having null value


1. In order to remove nodes with missing nominal values, select a *"Select Node"* from *"Record Ops"* tab in nodes palette.
2. Connect source node to this select node. Go to *"Settings"* tab.
3. Select *"Discard"* and write the expression which selects nominal attributes with null values using *"STRING1 matches STRING2"* and specify STRING2 as *""*. 
4. Click *"Apply"* and *"OK"*. Now, a select node will appear on the modeler canvas.
5. [Optional] To view the output of this select node, add a *"Table node"* ,*"Connect"* this node to the table node and *"Run"* the stream from the select node.

Discretization (Binning)

1. Select a *"Binning Node"* from *"Field Ops"* tab in nodes palette and connect select node of previous step to this node .Go to *"Settings"* tab.
2. Select all the numeric fields with *"Continuous"* values as *"Bin Fields"*.
3. As per the requirement, you may choose among many Binning methods like *"Fixed-width"* or *"Tiles"* or *"Ranks"* or *"Mean/Std Deviation based"*. For now, select *"Fixed-width"* method. Read *Help* to know what these term mean. 
4. Select *"Number of bins"* as appropriate.
5. [Optional] Go to *"Bin Values"* tab, click on *"Read Values"* button. You can see the lower and upper range of each of the bins for each of the binned field.
6. Click *"Apply"* and *"OK"*. Now, a binning node will appear on the modeler canvas.
7. New Binned values are appended as new fields in the dataset. To remove old fields, select a *"Filter Node"* from *"Field Ops"* tab in nodes palette.
8. Connect binning node to this filter node and double click on it.
9. Click on *"Filter"* column of old fields to filter them out. Click *"Apply"* and *"OK"*.
10. [Optional] To view the output of this node, add a *"Table node"* ,*"Connect"* this node to the table node and *"Run"* the stream from this node.

Sampling

1. Select a *"Sample Node"* from *"Record Ops"* tab in nodes palette. Connect select node of Step 1 to this sample node. Double click on sample node and go to *"Settings"* tab.
2. As per the requirement, you may choose among the *"Simple"* or *"Complex"* sample method. For now, select *"Simple"* method.

3. You can select the sampling criterion as either "First n" or "1-in-n" or "Random %". Try to understand difference between all the operations. 
4. Click "Apply" and "OK". Now, a sample node will appear on the modeler canvas.
5. [Optional] To view the output of this sample node, add a "Table node" ,"Connect" this node to the table node and "Run" the stream.

Normalization

4. Select a "Filler Node" from "Field Ops" tab in nodes palette. Connect select node of Step 1 to this filler node. Double click on sample node and go to "Settings" tab. Select all numeric fields in "Fill in fields". Select "Replace" condition as "Always".
5. In "Replace with" box, write your normalization expression. For example, for 0-1 normalization you should write

$$(@FIELD - @GLOBAL_MIN(@FIELD))/(@GLOBAL_MAX(@FIELD)-@GLOBAL_MIN(@FIELD))$$
6. Click "Apply" and "OK". Now, a filler node will appear on the modeler canvas.
7. [Optional] To view the output of this filler node, add a "Table node" ,"Connect" this node to the table node and "Run" the stream.

Correlation Determination

1. Select a "Statistics Node" from "Output" tab in nodes palette.
2. Connect select node of Step 1 to this node.
3. Double click on the node and go to "Settings" tab.
4. Under "Examine" box, select all numeric fields. You may check all statistics parameter under "Statistics" box.
5. Under "Correlate" box, again select all numeric fields.
6. You may adjust correlation strength parameters for weak, medium and strong correlation. Click "OK".
7. Click "Apply" and "OK". Now, a statistics node will appear on the modeler canvas.
8. Run the stream from this node and observe the correlations between different fields. It will help you in selecting features to be consider for further processing.

Assignment -

1. Use "Auto Data Preparation" node and configure it to perform all types of preprocessing techniques explained above.
2. Replace outlier values with null values using "Data Audit" Node.
3. Replace missing values by the mean of the values of records having same class value.
4. Replace missing values by majority voting of the values of records having same class value.
5. Perform Stratified Sampling and observe the difference in the output.
6. Perform binning using Tiles, Ranks and Mean based methods.
7. Perform z-score normalization without using Auto Data Preparation Node.

