

B.Sc. in Computing Data Mining

Lab sheet #2 = getting started with Rapid Miner

rapid-i.com



Overview

Objective:

- ◆ Introduction to Rapid Miner and its interface

◆ Agenda:

- ◆ Introduction to rapid miner
- ◆ Rapid miner GUI's
- ◆ Inputting data sets
- ◆ MetaData
- ◆ Useful tips and features

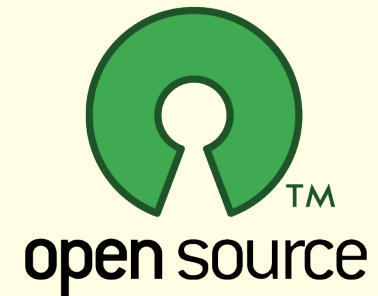
Exercises

There are a number of exercises through out the slides, and repeated again on the last slide.

Note down your answers to each question, as some of these will appear again in the moodle quiz at the end of the lab.

Background

- ◆ Rapid Miner is data mining tool developed by a company with close links to the University of Dortmund, Germany
- ◆ It won the European **Open Source Business Award** 2008 (announced Jan 31st 2008)
- ◆ **It is written in Java,**
 - ◆ Can run on any OS which has the JRE installed
 - ◆ Can be run from a GUI, or called from your own Java code.



Installing Rapidminer on your own machine

The latest version of Rapidminer, V6, is **NOT** free to use. Therefore we will be using their previous version, **V5.3.013**

It can be downloaded from [here](http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.3/) (
[http://sourceforge.net/projects/rapidminer/files/
1.%20RapidMiner/5.3/](http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.3/))

For windows: download rapidminer-5.3.013x64-install.exe and install (or the 32 bit version, depending on your machine). Defaults install it to **C:\program files**, and add it to the start>programs menu.

For mac: download rapidminer-5.3.013.zip. Unpack the zip file to your applications folder, and start rapidminer by running rapidminer/lib/rapidminer.jar<#>

Background

- ◆ **Rapid miner comes with over:**
 - ◆ Over 176 mining algorithms
 - ◆ Over 45 classes of data preparation functions.
 - ◆ Over 30 graphs for data visualisation,
 - ◆ and range of algorithms to evaluate attributes, allowing the user select the most predictive attributes in the dataset.
- ◆ Each function is available as an **OPERATOR**, (which is implemented as a Java class). A process is built by stringing operators together, with the output of one operator passing as input to the next. This is all done by drag and drop.

Installation

- ◆ The latest version can be downloaded from <http://rapid-i.com/content/view/26/84/>.
- ◆ Run the .exe to install.
- ◆ Defaults install it to C:\program files, and add it to the start>programs menu.

Starting Rapid Miner

- Windows:
 - You can start RapidMiner from the **start > all programs > rapidminer > rapidminer,**
 - or run **C:/program files/RapidMiner-5.2/RapidMiner.exe**
- Mac:
 - Start RapidMiner from **Applications > rapidminer-5.2 > lib > rapidminer.jar**



Repository

All Rapidminer files and datasets are stored in a folder called a **repository**.

When you first install rapidminer, you will be asked to create a repository.

Create a NEW folder for this, which you only use for Rapidminer work.

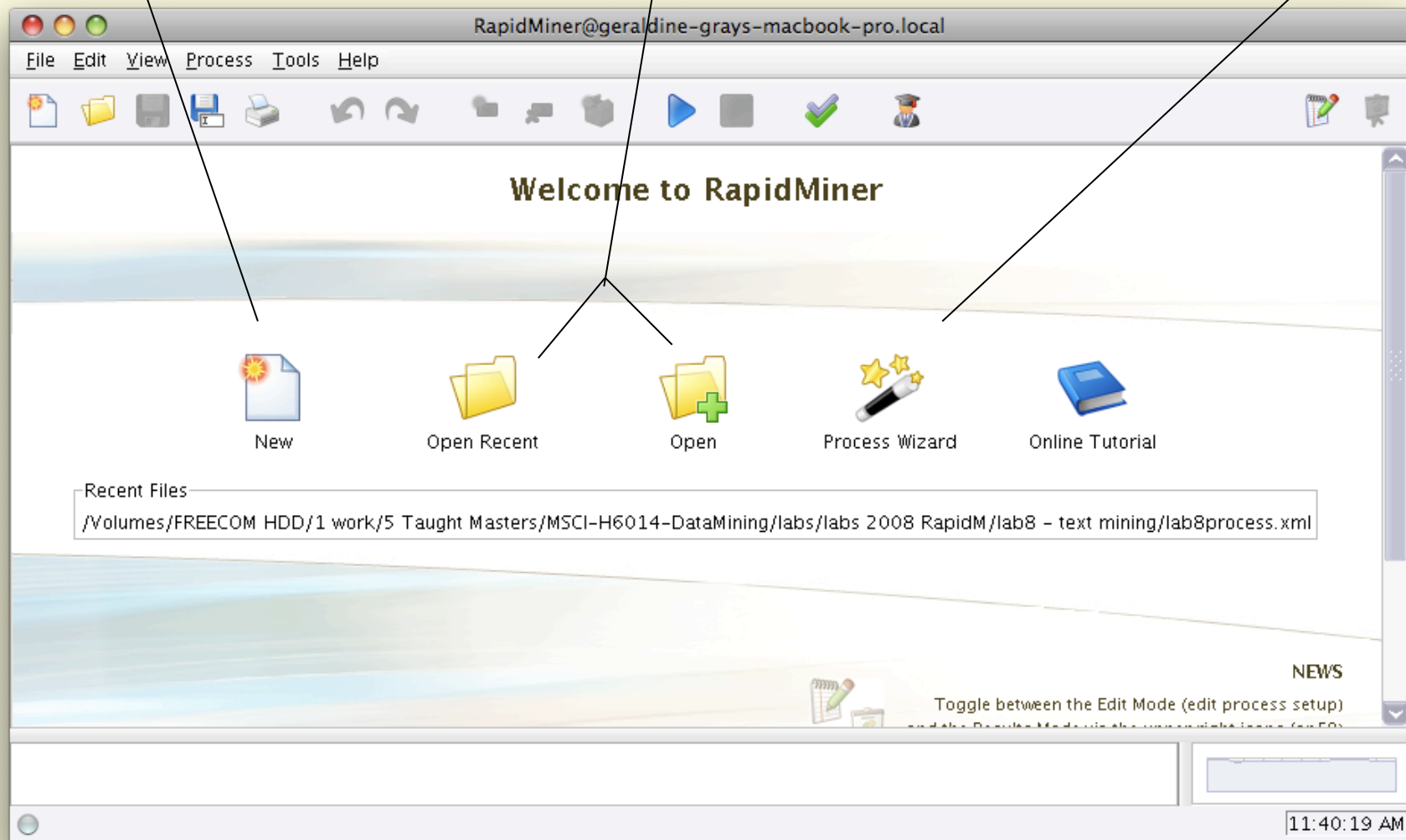
The folder can be local to the machine, on your X drive, or on a pen drive / external hard drive.

Starting Rapid Miner

Start a new process

Open an existing process

Customise a predefined process template



Select **New** process . .

G.Gray / L. Keyes

<#>

RAPID MINER GUI

Navigate repositories

Available operators and datasets

ments - RapidMiner@geraldine-grays-macbook-pro.local

The screenshot shows the Rapid Miner GUI interface. The top menu bar includes 'Tools', 'View', and 'Help'. Below the menu is a toolbar with icons for file operations, navigation, and execution. The main workspace is divided into several panes:

- Overview**: A small window showing a hierarchical view of the project.
- Repositories**: A list of data sources including 'Samples (none)', 'DB (null)', 'LocalRepository (work)', and 'RM-PhD-Repository (work)'.
- Operators**: A list of available operators for data processing.
- Main Process**: A central area showing a workflow diagram with three operators: 'Retrieve', 'Normalization', and 'PrincipalCom...'. The 'Retrieve' operator is connected to 'Normalization', which is connected to 'PrincipalCom...'. The 'Normalization' operator has parameters 'exa', 'ori', and 'pre'.
- Parameters**: A pane on the right showing settings for the 'Root (Process)' operator, including 'logverbosity' (set to 'init'), 'logfile', 'resultfile', 'random seed' (set to '2001'), 'send mail' (set to 'never'), and 'encoding' (set to 'SYSTEM').
- Problems**: A pane at the bottom showing 'No problems found'.
- Log**: A pane at the bottom showing a log of activities, including errors.
- System Monitor**: A pane at the bottom showing system resources like 'Max: 173 MB' and 'Total: 81 MB'.

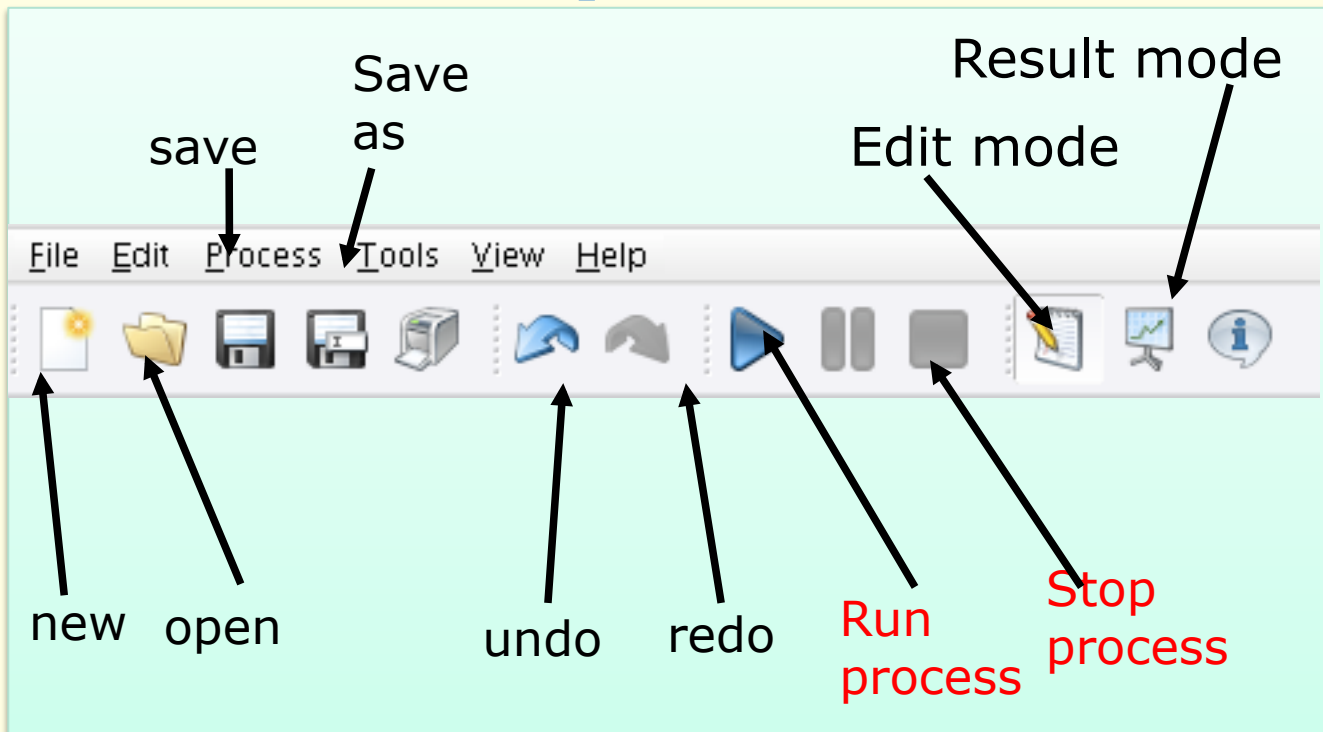
Log of activities, including errors

Stream / flow of operations.

Explanation of the current operator

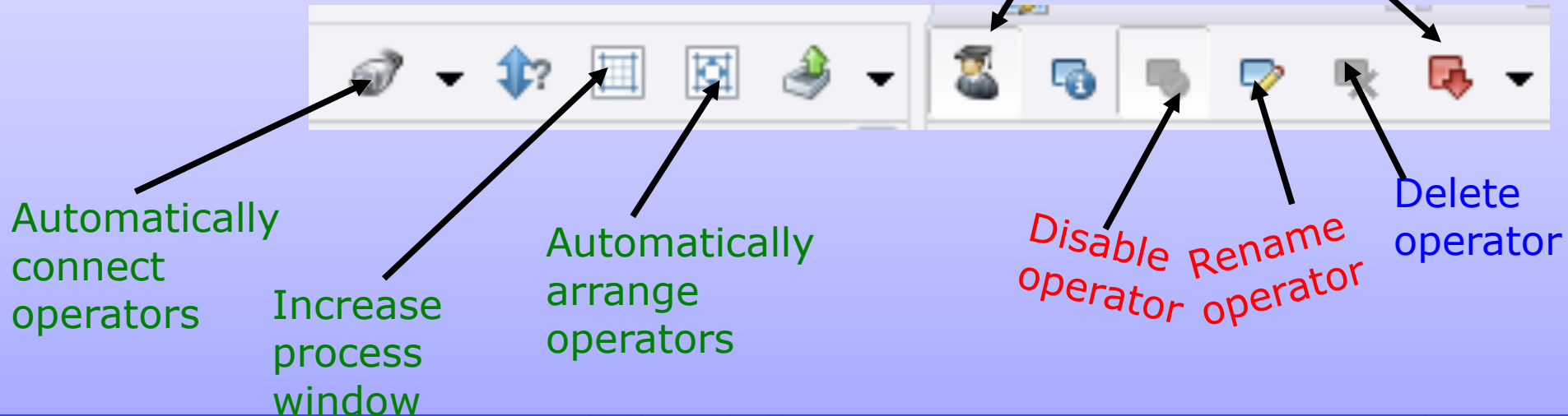
Parameter settings for selected operation

Rapid Miner toolbars



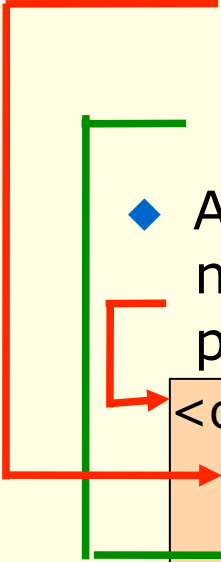
Expert mode
– to view
additional
parameters

Add
breakpoint



Rapid Miner & XML

- ◆ A process, which lists all operations to be executed on the data, is stored as an xml file. The XML file lists:
 - ◆ The name of each operator in the process, and the corresponding Java class that implements the operator
 - ◆ The input parameters to the operator
- ◆ A process starts with a root node of 'operator' with name="Root". All operations are embedded in this root process.



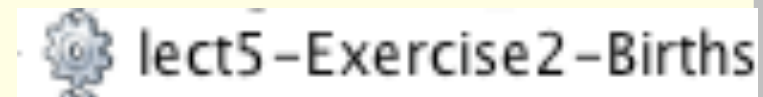
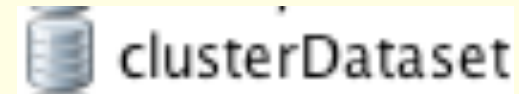
```
<operator name="Root" class="Process" expanded="yes">  
  <operator name="ArffExampleSource" class="ArffExampleSource">  
    <parameter key="data_file"  
      value="C:\rm_workspace\sample\data\iris.arff"/>  
    <parameter key="label_attribute" value="class"/>  
  </operator>  
</operator>
```

The diagram illustrates the mapping between the XML structure and the list items. A red line connects the first list item (the XML file) to the root element. A green line connects the second list item (operator name and class) to the inner operator element. A red line connects the third list item (input parameters) to the parameter elements.

Processes and Datasets

Your rapid miner repository (folder) will contain two types of objects:

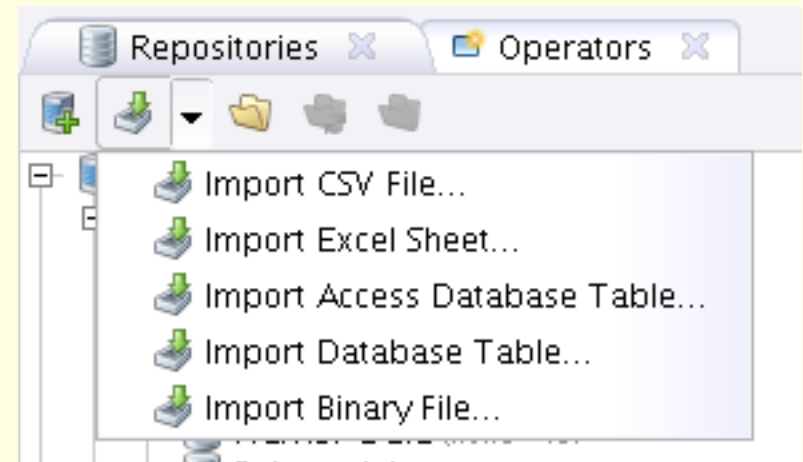
- **Datasets** – the actual data itself
 - The symbol is a cylinder
- **Processes** – a series of operators (written as a Java program) running on the dataset to analyse it. A Java program is
 - The symbol is a cog wheel
 - A process will read in a dataset, carry out various operators on it, and display the results. A process will NOT change the original dataset.



Reading in a dataset

There are **three** options to access a dataset:

1. You can use the **IMPORT/DATA** operator to read it into Rapidminer temporarily for a particular process. This is the least efficient option.
2. You can **import a dataset into a repository**, where it will be available to all processes via the **retrieve** operator. This is the most efficient method.
 - Rapidminer ships with a number of datasets already loaded in a repository called **SAMPLES**
3. If the dataset is already in the repository, you can simply retrieve it using the **Retrieve** operator.



Iris Dataset

We are first going to import the IRIS dataset – a simple dataset frequently references in text books and conference papers.

Can be obtained from the UCI Machine Learning Repository

<http://www.ics.uci.edu/~mlearn/MLRepository.html>

From the statistician Douglas Fisher

Three flower types (classes):

Setosa

Virginica

Versicolour

Four (non-class) attributes

Sepal width and length

Petal width and length



Virginica. Robert H. Mohlenbrock.
USDA NRCS. 1995

Importing a dataset into the repository

1. Download the file iris.txt from moodle or studentshare.
2. Start up RapidMiner, and **Start** a **new process**.
Select the repository where you want this process to be stored. Call the process **lab2/Iris-EDA** which will create a subfolder in your repository called lab2.
3. On the lefthand side repository window, **click** the down arrow beside the import icon, and select **Import CSV file**.
4. **Navigate** to where you have saved the Iris file (you'll need to display ALL files, not just .csv), select iris.txt and then **Click next**.



The next screen allows you to select the column separator for your file. **Select comma**. The data set should now **have five columns**. **Click next**.

Importing a dataset into the repository

The next window allows you to specify if all rows are data. In this case they are NOT, the first row is column names. **Click** on the first row under 'annotation'. **Select Name**. **Click next**.

Annotation	att1	att2	att3	att4	att5
-	sepallenght	sepalwidth	petallength	petalwidth	iris
-	4.9	3.0	1.4	0.2	Iris-setosa
Name	4.7	3.2	1.3	0.2	Iris-setosa
Comment	4.6	3.1	1.5	0.2	Iris-setosa
Unit	5.0	3.6	1.4	0.2	Iris-setosa
-	5.4	3.9	1.7	0.4	Iris-setosa
-	4.6	3.4	1.4	0.3	Iris-setosa
-	5.0	3.4	1.5	0.2	Iris-setosa

If there are rows you want Rapidminer to ignore, give them an annotation of 'comment'.

Importing a dataset into the repository

This next screen is **IMPORTANT** to get correct, as mistakes here can not be rectified easily without re-importing the dataset again.

The primary purpose of the window is to specify:

1. The datatype of each column
2. Which column is the label, i.e. the column you want mining algorithms to learn how to predict.

Should column be included? →

Column name →

Data type →

<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
sepal.length	sepal.width	petal.length	petal.width	iris
real	real	real	real	polyn...
attribut	attribut	attribut	attribut	label
4.900	3	1.400	0.200	Iris-setosa
4.700	3.200	1.300	0.200	Iris-setosa

Identify iris attribute. **CHANGE TO LABEL**

Actual data

Importing a dataset into the repository

The final step is to specify where in the repository the dataset should be stored. Select the local repository you created, and store it under data/iris

This will create a new folder called data (to be used for all datasets you import), and the iris dataset will be stored here.

Note: The steps to import a dataset temporarily for a particular process under the operators IMPORT/DATA are **very similar** to importing to a repository, and so will not be covered here.

1. Reading in a dataset

You have already started a new process called **Iris-EDA**. We will now use this to explore the iris dataset.

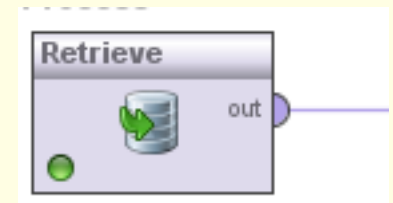
Datasets and Operators are added to a process by selecting them from the repository and operator list respectively on the left hand side window.

Operators are grouped by category: process control, utility etc.

Sample datasets, along with their meta data, are located in the repository

1. **Select** the **iris** dataset from the *yourRepository/data* and drag it onto the process window.

This adds a 'retrieve' operator to the process window.



1. Reading in the dataset

You will notice a semi-circle at the side of the operator with the word 'out'. There is a semicircle for each output parameter, which can be connected to the input of the next operator, or to 'res' at the side of the processor window. **Anything connected to 'res' will be displayed as a result when the process is run.**

2. **Connect** 'out' to 'res' by dragging the mouse from out to res.

Note: As you click on each operator in the process window, the parameter values for that operator are displayed on the left hand side.

You can run the process by selecting the 'play' button on the top menu bar.



1. Reading in a dataset

After running the process you ,may be asked if you want to view the results. Select Yes. This toggles you to the **result workspace** from where you can view the results of the process to date. You can toggle back to the design view at any time:



There are three views in the results window: 1. **meta data view** (attribute names and datatypes); 2. **data view** (the data itself); 3. **plot view** (graphs of the data)

<input checked="" type="radio"/> Meta Data View <input type="radio"/> Data View <input type="radio"/> Plot View <input type="radio"/> Annotations					
ExampleSet (146 examples, 1 special attribute, 4 regular attributes)					
Role	Name	Type	Statistics	Range	
label	iris	polynomial	mode = Iris-setosa (49), least	Iris-setosa (49), Iris-v	0
regular	sepalenght	real	avg = 5.849 +/- 0.836	[4.300 ; 7.900]	0
regular	sepalwidth	real	avg = 3.056 +/- 0.432	[2.000 ; 4.400]	0
regular	petallength	real	avg = 3.755 +/- 1.768	[1.000 ; 6.900]	0
regular	petalwidth	real	avg = 1.195 +/- 0.761	[0.100 ; 2.500]	0

Dataset Views

1) Meta Data view

As you can see, this data set has five attributes. The class is the type of iris; the remaining four attributes measure sepal and petal dimensions for instances of each type of iris.

The Meta data view also displays **statistics** on each attribute.

Note: averages are given for numeric attributes, mode is given for non numeric attributes.

Unknown refers to the number of **missing values (blanks)** for that attribute.

<input checked="" type="radio"/> Meta Data View <input type="radio"/> Data View <input type="radio"/> Plot View <input type="radio"/> Annotations					
ExampleSet (146 examples, 1 special attribute, 4 regular attributes)					
Role	Name	Type	Statistics	Range	
label	iris	polynomial	mode = Iris-setosa (49), least	Iris-setosa (49), Iris-v	0
regular	sepallength	real	avg = 5.849 +/- 0.836	[4.300 ; 7.900]	0
regular	sepalwidth	real	avg = 3.056 +/- 0.432	[2.000 ; 4.400]	0
regular	petallength	real	avg = 3.755 +/- 1.768	[1.000 ; 6.900]	0
regular	petalwidth	real	avg = 1.195 +/- 0.761	[0.100 ; 2.500]	0

Dataset views

Exercise 1: From an initial glance at the dataset statistics make an assessment of the following:

1. How would you describe the information content of the attributes – good or bad. Explain your answer.
2. Do all numeric attributes have similar ranges?
3. Do any min or max values look unusually large or small indicating errors or outliers?
4. Are average values roughly midway indicating data is not skewed?

Dataset Views

2) Data View shows you the **data** in the file, and also allows you to check the **quality** of the dataset.

Exercise 2:

Do any rows have missing values?

ExampleSet (146 examples, 1 special attribute, 4 regular attributes)

Row No.	iris	sepalheight	sepalwidth	petalheight	petalwidth
1	Iris-setosa	4.900	3	1.400	0.200
2	Iris-setosa	4.700	3.200	1.300	0.200
3	Iris-setosa	4.600	3.100	1.500	0.200
4	Iris-setosa	5	3.600	1.400	0.200

View Filter (150 / 150):

all

- all
- no_missing_attributes
- missing_attributes
- no_missing_labels
- missing_labels

Dataset Views

3) Data plot view.

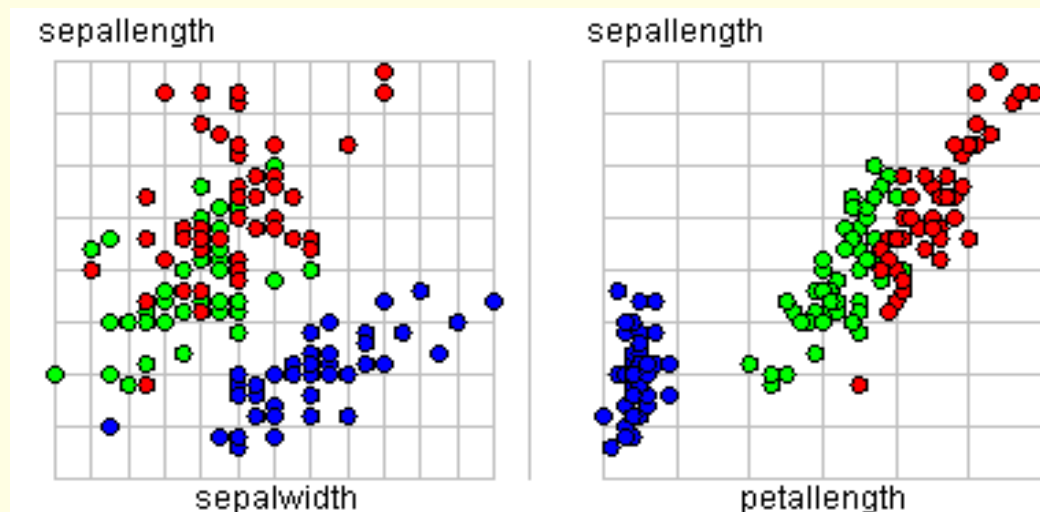
The plot view provides a range of graphs from which to visualise and explore the data set.

Select scatter matrix. Set plot to class.

The right hand plane shows each attribute plotted against each other attribute, overlaid in colour by class.

Exercise 3:

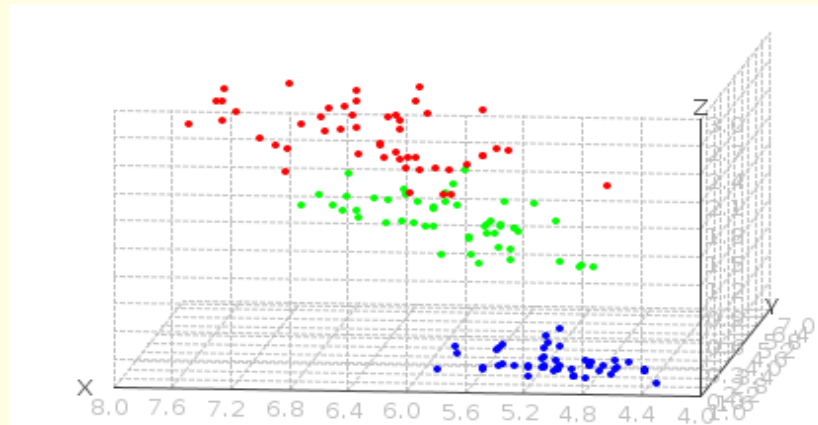
Which attributes are most closely correlated (i.e. as the value of one attribute increases, so does the value of the other)?



Dataset views

Select a 3D colour graph.

Set the three **axes** to: petal length; petal width & sepal length. Set **Colour** – class. Use the mouse to drag the 3-D image so that you can view it from different perspective.



Exercise 4: Looking at the graph, how accurate do you think a classifier might be (are there clear distinctions between the three classes, or is there overlap between classes?)

There are over 30 graphs to choose from in all.

Dataset views

Select a Quartile Color Matrix.

Leave colour at none.

Exercise 5:

1. Which attributes seems most normally distributed?
2. Which attributes is most skewed?



Exercise 6: Using the same steps as per the IRIS dataset, use the time that is left to complete an Exploratory Data Analysis of the churnWithMissing.csv data set available on moodle and studentshare. This dataset is to allow a mining algorithm learn how to predict if customers of a phone company are likely to churn (move to another provider).

- Are there any missing values?
- Are there enough rows? (are there at least 20 times as many rows as columns)
- Using statistics, quartiles/box plots or histograms, are there any outliers?
- Which numeric columns are NOT normally distributed?