ADAMS

Advanced **D**ata mining **A**nd **M**achine learning **S**ystem

Module: adams-visualstats

Peter Reutemann

June 10, 2013

Contents

List of Figures

Chapter 1

Introduction

Visualizing data is very important to understand the data that you are dealing with. Even more important it is to visualize the statistics and evaluations that you generated in order to better understand your models, whether they are working properly and not just plain useless. The *visualstats* module contains some useful visualizations for statistics which are discussed in the next chapter.

Chapter 2

Flow

2.1 Actors

The following sinks are available:

- *BoxPlot* – displays box plots [2] of the attributes of a dataset.
- *FourInOneDisplay* – Plots the residuals[3] in various ways: probability plot, fit, histogram [5], order.
- *Histogram* – plots a histogram[5] of numeric data.
- *MatrixPlot* – scatter plots of attributes, all vs all.
- *ProbabilityPlotDisplay* – plots the probabilities of predictions, with option regression line, e.g., a normal probability plot [6].
- *ScatterDisplay* – plots one attribute vs another [4].
- *ZScoreDisplay* – displays the z-scores of predictions.

2.2 Examples

The following sections demonstrate how to use the previously introduced actors.

2.2.1 Box plot

The *BoxPlot*[1] sink takes a WEKA dataset (`weka.core.Instances`) object as input. You can specify the range of attributes to display.

The plot, as it shows the range for an attribute, makes only sense to use for numeric attributes. Hence it is easiest to use a filter to remove non-numeric attributes, e.g., *RemoveType* with *numeric* attributes as the type to remove and the *invertSelection* flag enabled[2].

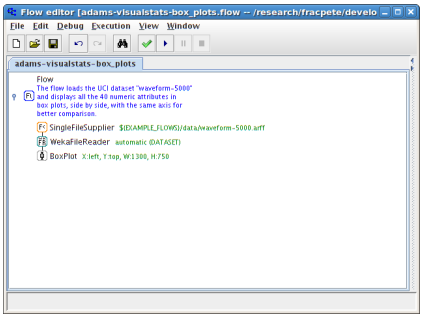

Figure 2.1: Box plot flow.

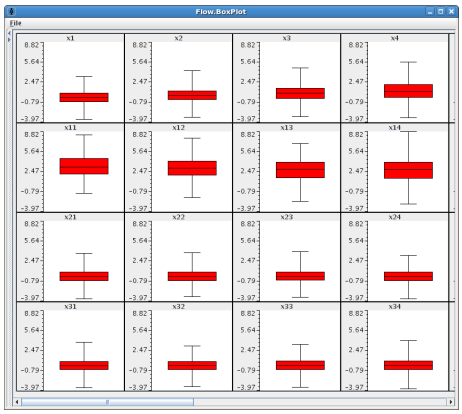

Figure 2.2: Box plot output.

[1] adams-visualstats-box_plots.flow
[2] adams-visualstats-box_plots.flow

2.2.2 Histogram

With the *Histogram* sink, you can quickly plot an attribute of a WEKA dataset or any double array. Figures 2.3 and 2.4 show a flow[3] that generates a histogram for the *sepallength* attribute of the UCI dataset *iris*.

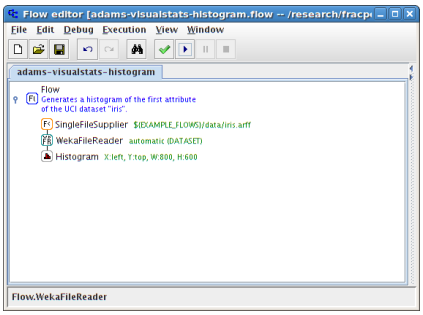

Figure 2.3: Histogram flow.

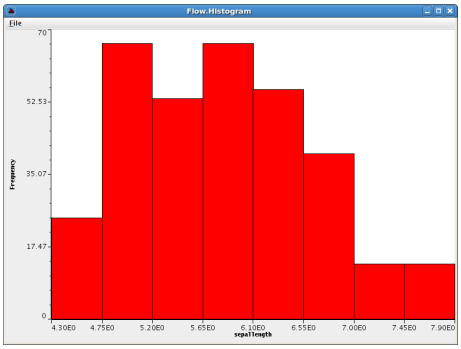

Figure 2.4: Histogram output.

[3] adams-visualstats-histogram.flow

2.2.3 4-in-1

The 4-in-1 plot is a quick solution to plotting the residuals from a regression analysis[7] in a single plot. The four displays are:

- *normal probability plot* – if the data points are along the diagonal, the data seems to be normal distributed. An s-shape suggests a uniform distribution.
- *histogram* – a bell-shaped histogram indicates a normal distribution.
- *versus fit* – plots the residuals against their predicted value, with a random distribution around the mean indicating a good fit of the model.
- *versus order* – shows the residuals how the were generated by the model, one after the other. A *trumpet* shape here indicates a non-constant variance.

Figure 2.5 shows a flow for generating a 4-in-1 plot based on a cross-validation run of LinearRegression on the UCI dataset *bodyfat*. In Figure 2.6 you can see the generated plot.

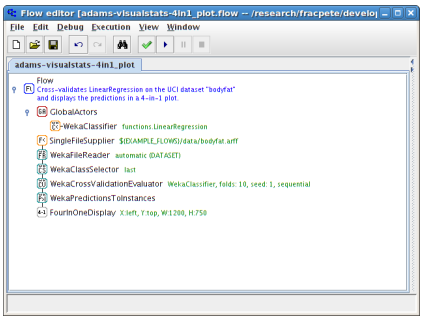

Figure 2.5: Flow using 4-in-1 display.

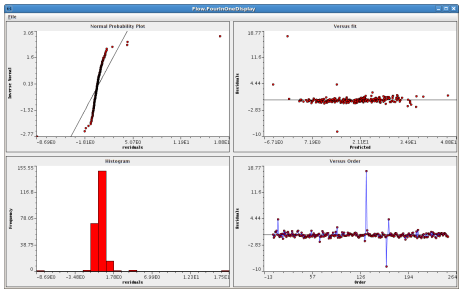

Figure 2.6: 4-in-1 display output.

ADAMS comes with some example flows using different datasets.[4] [5] [6]

[4] adams-visualstats-4in1_plot.flow
[5] adams-visualstats-4in1_plot2-slug_original.flow
[6] adams-visualstats-4in1_plot2-slug_ln.flow

2.2.4 Matrix plot

The *MatrixPlot* sink simply plots all attributes versus each other. This, in combination with overlays such as LOWESS (locally weighted scatterplot smoothing), allows you to determine whether there is a relationship (or collinearity) between attributes. Figures 2.7 shows a flow[7] that displays the UCI dataset *iris* as a matrix plot (2.8). The almost straight line of the LOWESS overlay shows, that there is a linear relationship between *petallength* and *petalwidth*. The nominal class also has a good separation for the *petallength* and *petalwidth* attributes (see bottom row).

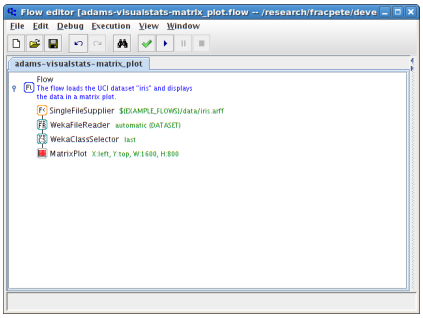

Figure 2.7: Matrix plot flow.

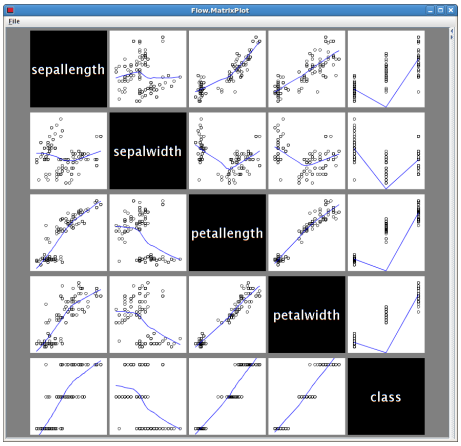

Figure 2.8: Matrix plot output.

[7] adams-visualstats-matrix_plot.flow

2.2.5 Probability plot

The probability plot is used for visualizing predictions results, actual vs predicted. The plot allows you to choose from the following distributions:

- exponential
- gamma
- logistic
- *log*(logistic)
- normal *(default)* [6]
- *log*(normal)

Depending on the distribution, it is also possible to display a regression line, i.e., the optimal line for predicted vs actual.

The flow[8] depicted in Figure 2.9 uses the *log*-transformed *slug* data (see [8]) to cross-validate LinearRegression and then displays the predictions using the *normal* distribution (see Figure 2.10).

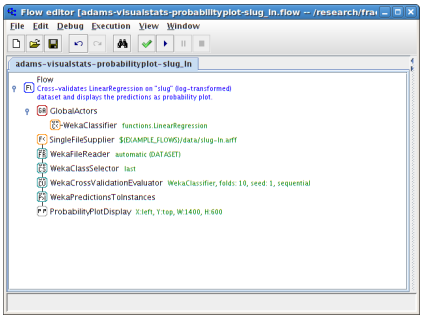

Figure 2.9: Probability plot flow.

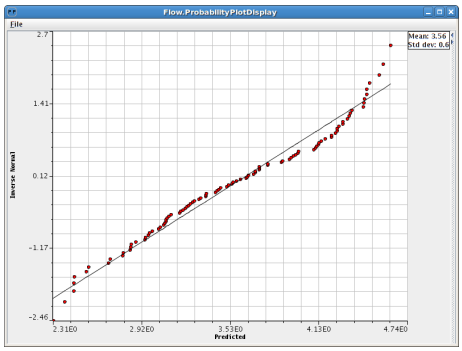

Figure 2.10: Probability plot output.

[8] adams-visualstats-probabilityplot-slug_ln.flow

2.2.6 Scatter display

The scatter display or scatter plot allows you to plot two attributes from a dataset against each, one on the x-axis, the other on the y-axis. You can add overlays to the graph, like a simple diagonal (for regression predictions this is the optimum when plotting actual vs predicted) or the LOWESS (locally weighted scatterplot smoothing). The flow[9] shown in Figure 2.11 generates the output in Figure 2.12.

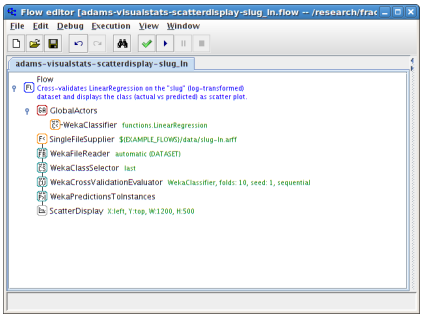

Figure 2.11: Scatter display flow.

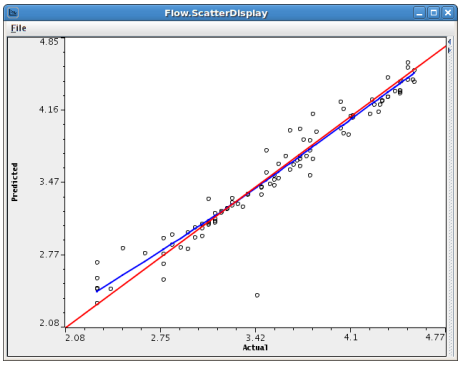

Figure 2.12: Scatter display output.

[9] adams-visualstats-scatterdisplay-slug_ln.flow

2.2.7 Z-Score display

The z-score (or standard score) shows how many standard deviations an observation is off the mean. By default, the display shows markers for the mean, at $\pm 2\sigma$ (covers roughly 95% of predictions) and $\pm 3\sigma$ (covers almost all predictions). Figure 2.13 shows a flow[10] for displaying z-scores and classifier errors. The z-score display is shown in 2.14 and the classifier errors in 2.15.

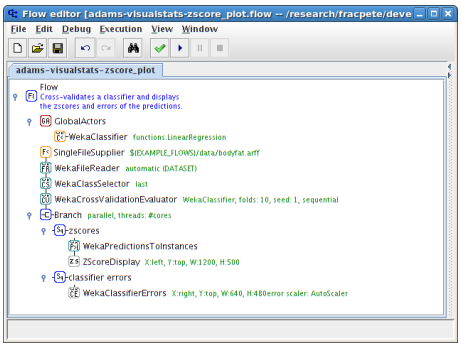

Figure 2.13: Flow for generating z-scores.

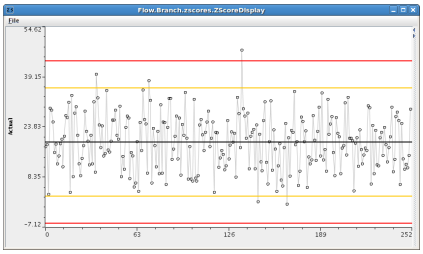

Figure 2.14: Z-Score output.

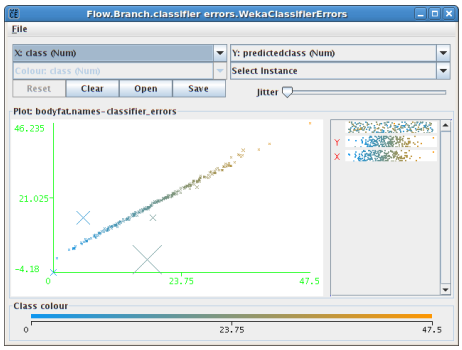

Figure 2.15: Classifier errors output.

[10] adams-visualstats-zscore_plot.flow

Bibliography

[1] *ADAMS* – Advanced Data mining and Machine learning System
`https://adams.cms.waikato.ac.nz/`

[2] *Box plot* – Graphically depicting groups of numerical data
`http://en.wikipedia.org/wiki/Box_plot`

[3] *4-in-1 plot* – Four different plots of residuals
`http://stattrek.com/regression/residual-analysis.aspx`

[4] *Scatter plot* – Plotting two attributes against each other
`http://en.wikipedia.org/wiki/Scatter_plot`

[5] *Histogram* – Visual impression of the distribution of data
`http://en.wikipedia.org/wiki/Histogram`

[6] *Normal probability plot* – Graphical technique for normality testing
`http://en.wikipedia.org/wiki/Normal_probability_plot`

[7] *Regression analysis* – statistical technique for estimating the relationships among variables
`http://en.wikipedia.org/wiki/Regression_analysis`

[8] Barker, G, and McGhie, R (1984). The Biology of Introduced Slugs (Pulmonata) in New Zealand: *Introduction and Notes on Limax Maximus*, NZ Entomologist 8, pp 106-111.