BIO705 – Machine Learning for non-Programmers

As this module draws to a close, we want to undertake a short but fair assessment to test your skills analysing a data set and working with KNIME. This exercise will build on your previous experience, developing your analysis of the Iris data further, and will introduce a new data set.

Submission

From starting the assessment, you will have 10 days to submit your work. You should submit two documents: (a) a KNIME worksheet that implements as many of the tasks as possible, (b) a Word document containing appropriate text to answer the questions. Email your submissions to me directly (s.watterson@ulster.ac.uk)

The total mark is out of 100.

Activities

Create a new KNIME worksheet. Use a CSV reader to open the Iris.csv file.

The iris data set comprises 50 observations for each of 3 species of Iris. We will start by comparing pairs of species.

Using row filter nodes create 3 tables (and remember to use the domain calculator node to establish the correct meta data for each table): one for each of (a) Iris-versicolor vs Iris-setosa, (b) Iris-versicolor vs Iris-virginica, (c) Iris-setosa vs Iris-virginica. [10 marks]

For each of the 3 pairs (a), (b) and (c),

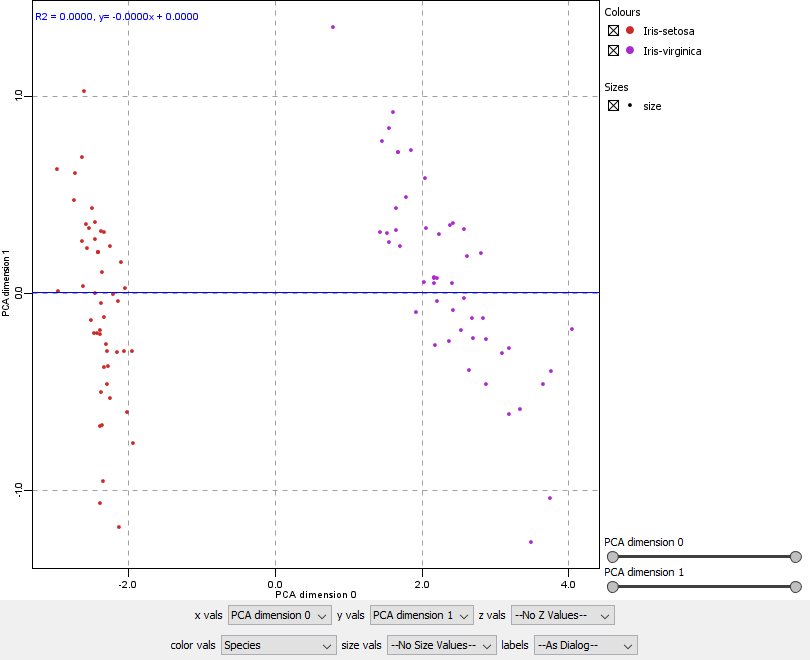
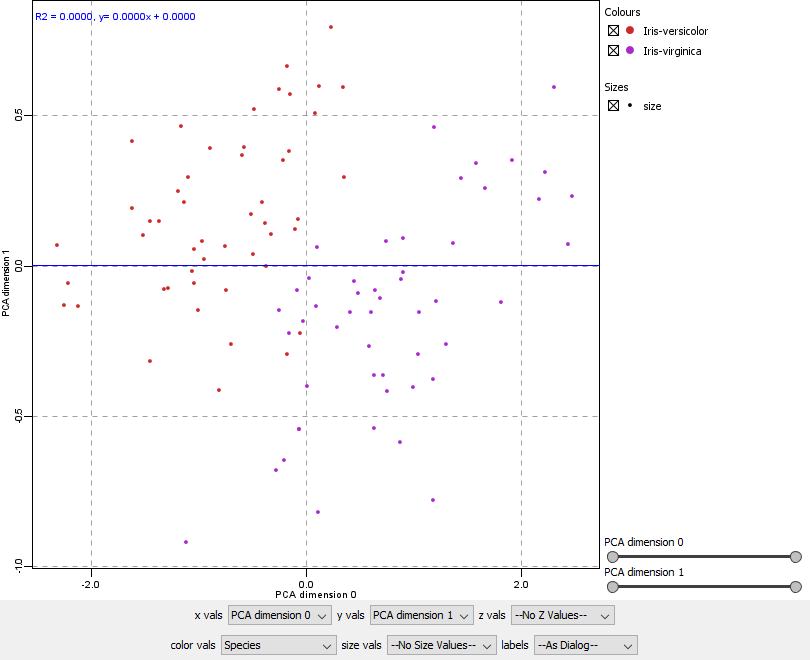
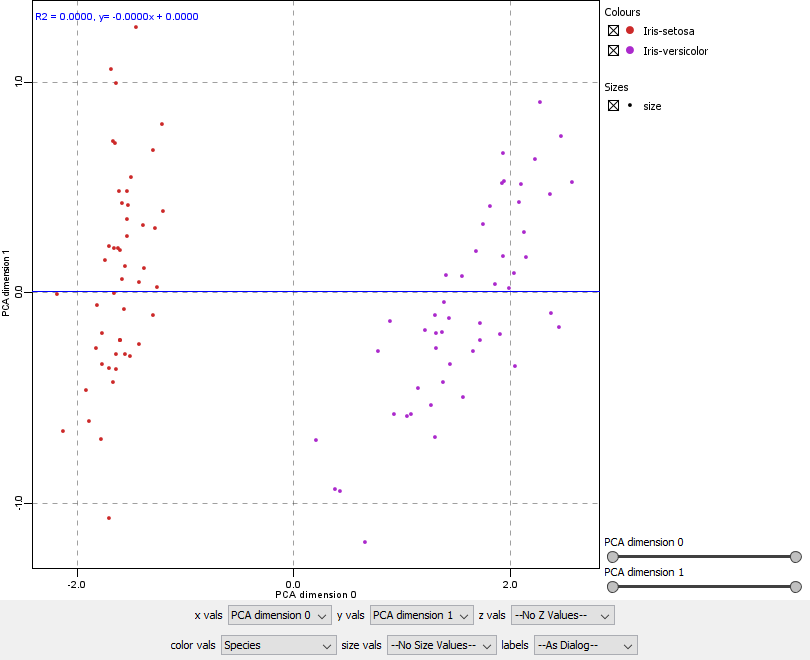
(i) calculate the k-means clustering for 2 clusters [10 marks]

(ii) calculate the overall silhouette coefficient [10 marks]

(iii) use the SVM Learner, SVM predictor and ROC Curve (local) nodes to calculate an AUC for a SVM classifier [Hint: in the Predictor nodes remember to tick the box, Append columns with normalised class distribution so that the ROC node has probabilities to use.] [10 marks],

(iv) calculate a 2D PCA using the PCA node, plot the data points using first two principle components using the 2D/3D Scatterplot node and colour code the nodes by species (this is a setting in the view generated by the node) [10 marks].

Your scatterplots should look like this:-

Questions

1. Compare the silhouette scores and AUCs. Can you explain why higher AUCs are observed for higher silhouette scores and lower AUCs for lower silhouette scores? [5 marks]

We are now going to switch datasets. Please download the following file:-

<https://stevenwatterson.github.io/MLShortCourse/Titanic.csv>

This contains data on the passengers aboard RMS Titanic and we will use this data to build classifiers predicting who was likely to survive. Whilst most of the data is complete, the Age and Cabin columns contain much missing data.

Import the data to your worksheet using a CSV reader node and, using a column filter, remove the Age and Cabin columns from the data set. [5 marks]

Build a (i) logistic regression classifier, SVM classifier and (ii) random forest classifier predicting survival and calculate the AUC of each. [10 marks]

Next, duplicate your logistic regression, SVM and random forest classifiers and introduce 10 fold cross validation (using the x-partitioner and x-aggregator nodes), remembering to train (using the Learner nodes) and test (using the Predictor nodes) the classifiers on separate parts of the data set. Calculate the AUC for each of the cross validated learners. [20 marks]

Questions

2. Before the introduction of cross validation, how do the AUC of the three classifiers compare? [5 marks]

3. How do the AUCs of the classifiers compare before and after cross validation? [5 marks]