IRootLab Tutorials

Classification with SVM

Julio Trevisan – juliotrevisan@gmail.com

25th/July/2013

  
This document is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](http://creativecommons.org/licenses/by-nc-sa/3.0/).

Loading the dataset 1

Pre-processing 2

Optimization of SVM parameters 5

Creation of required objects 5

Grid search 8

Visualization of results 10

Visualization of the optimization log 10

Classification confusion matrix for best parameters 12

# Introduction

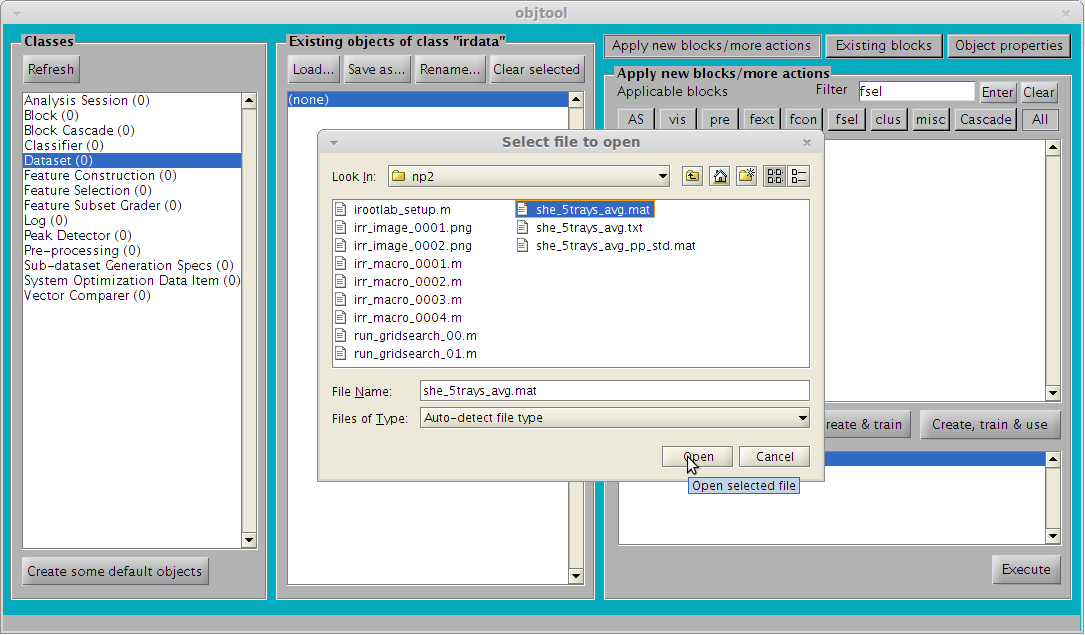
The SVM classifier has tuning parameters that need to be optimized before fitting data. This tutorial will take you through the steps of:

* finding these parameters
* visualizing information about the optimization progress
* getting a confusion matrix for the optimally tuned classifier

Recommended reading: [1]

# Loading the dataset

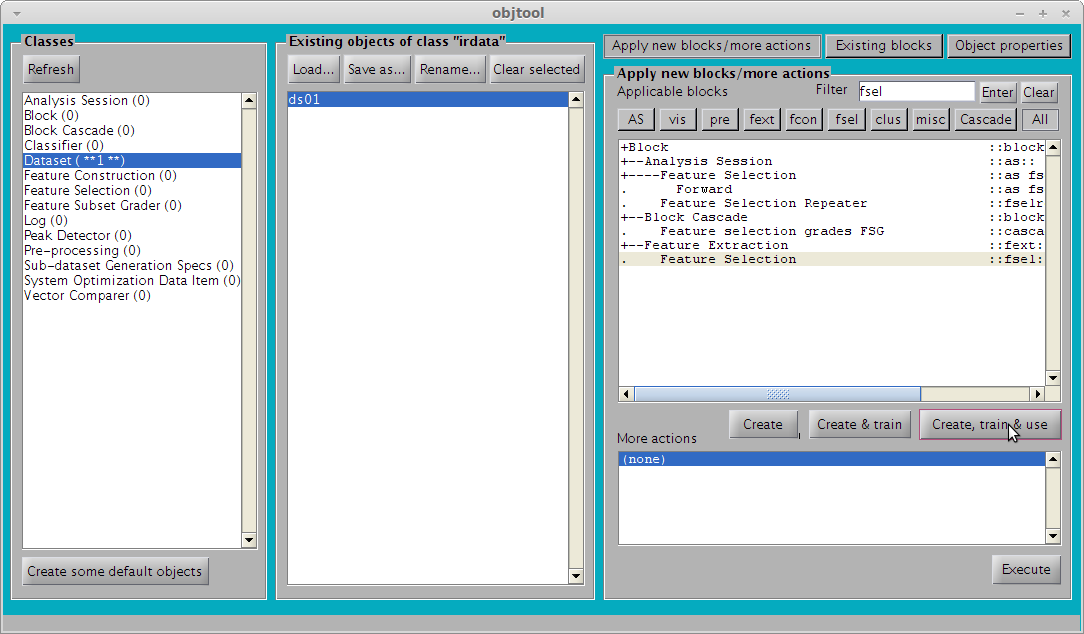
1. Start MATLAB and IRootLab as indicated in IRootLab manual (<http://irootlab.googlecode.com>).
2. In MATLAB command prompt, type “objtool”.
3. Click on “Load…” and select dataset.



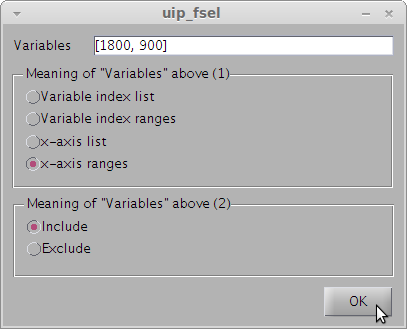
# Pre-processing

This tutorial will cut to the 1800 – 900 cm-1 region and apply 1st differentiation (Savitzki-Golay) followed by vector normalization (spectrum-wise), then normalization to the [0, 1] range (variable-wise).

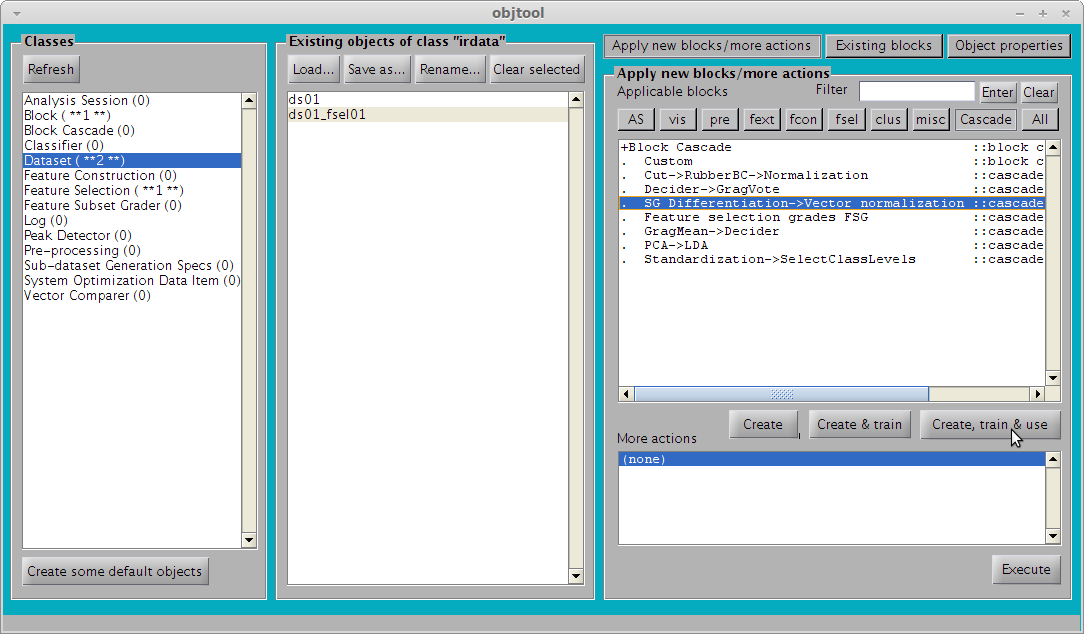
1. Locate and double-click “Feature Selection” in the right panel



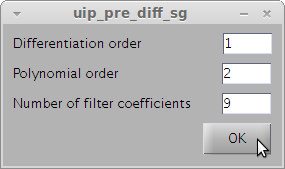
1. Click on “OK”.



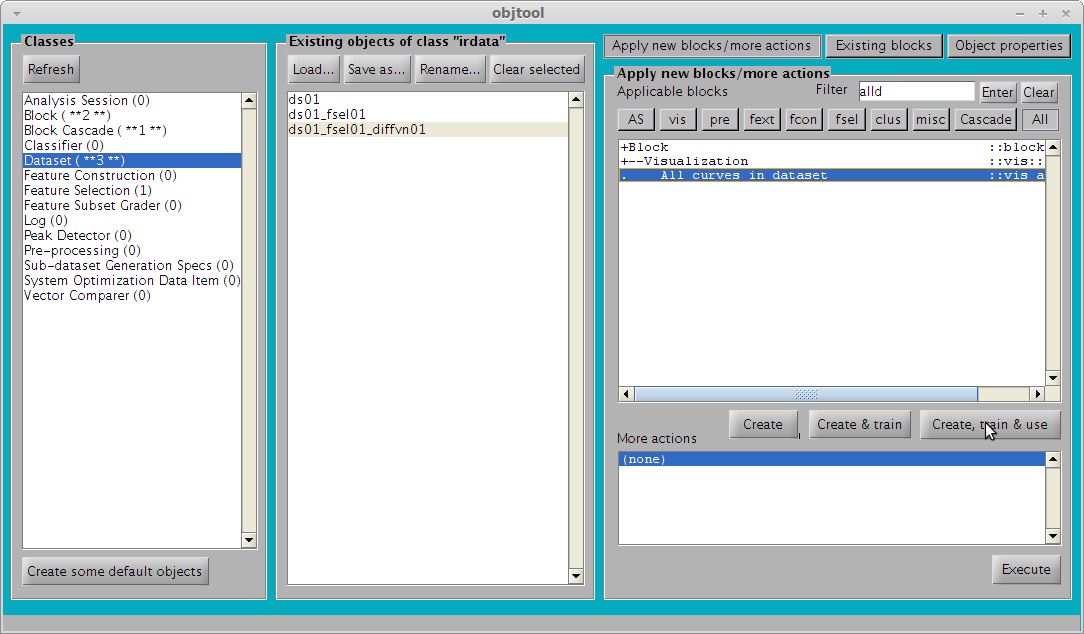
1. Select “ds01\_fsel01” in the middle panel.
2. Locate and double-click on “SG Differentiation->Vector normalization” in the right panel

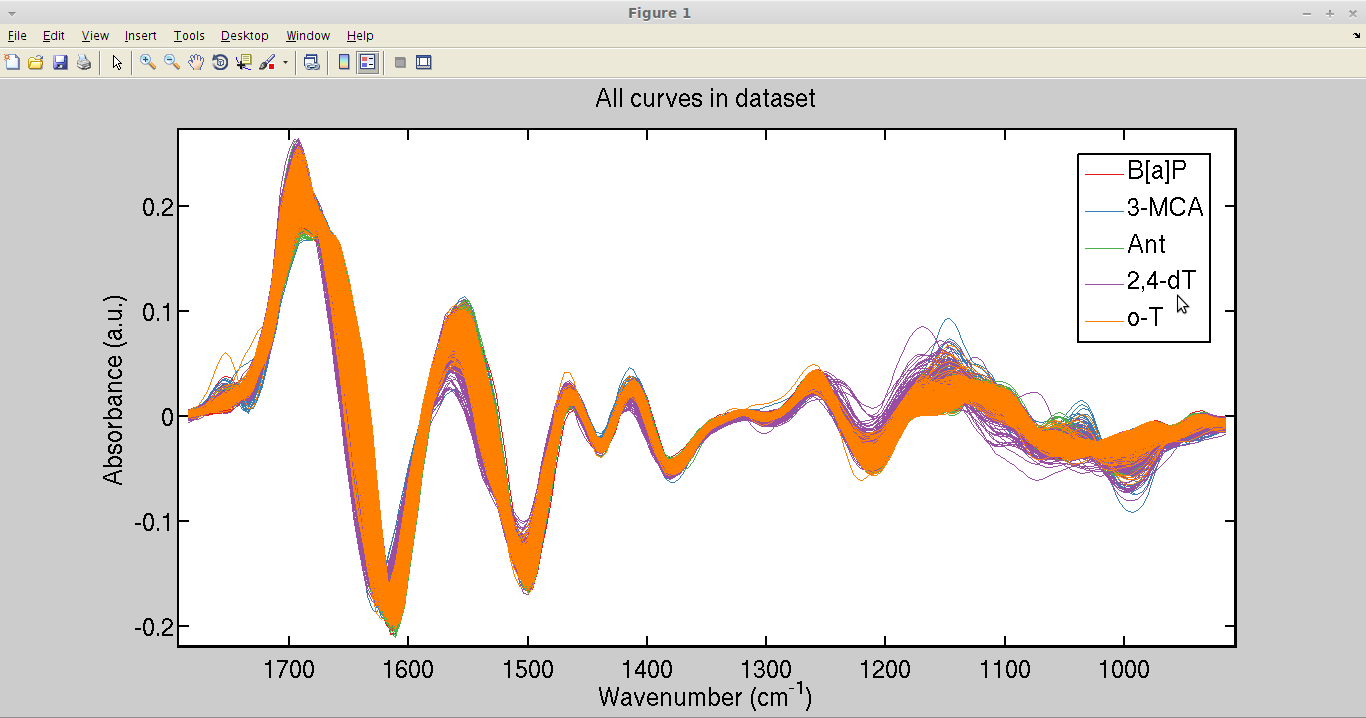


1. Click on “OK”

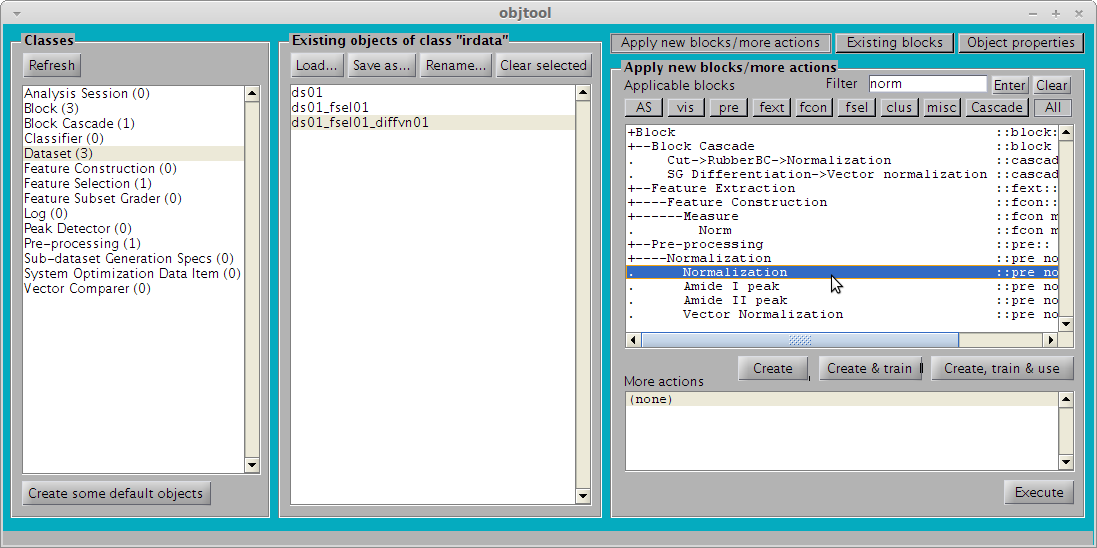


1. Click on “ds01\_fsel01\_diffvn01” in the middle panel
2. Locate and double-click “All curves in dataset”

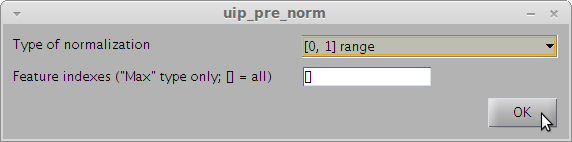




1. Locate and double-click on “Normalization” in the right panel.



1. Select “[0, 1] range” from the “Type of normalization” pop-up box
2. Click on “OK”



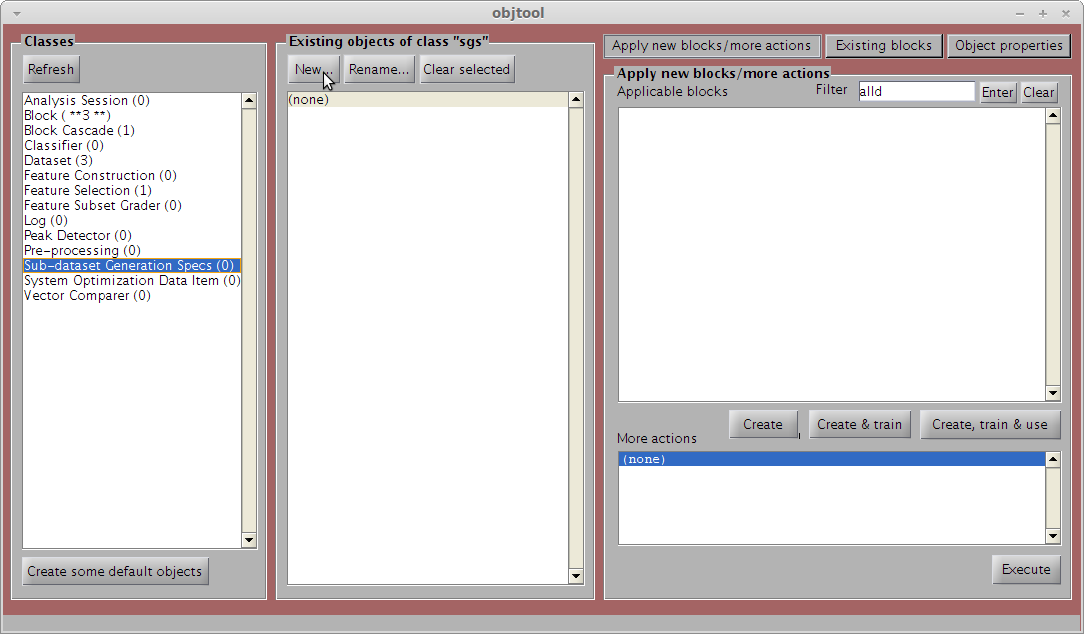
From now on, the procedure splits in two options. The first option is to work with SVM directly on the normalized data. The second option uses PCA as a variable reduction technique prior to SVM classification.

# SVM

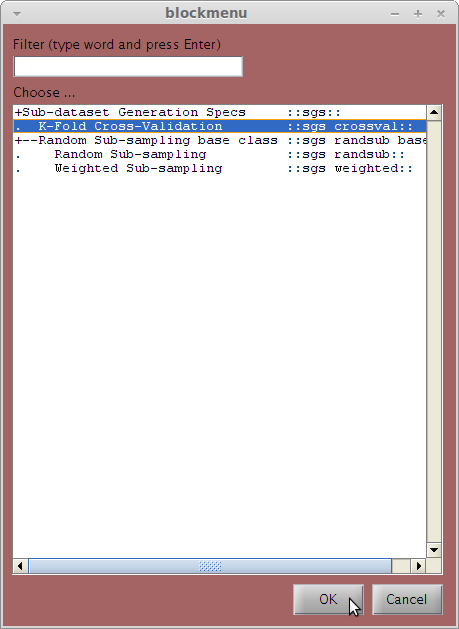
This tutorial utilizes the Gaussian kernel SVM, which implies that there are two parameters to tune: *c* and *gamma* (these parameters are referred to as *C* and *γ* in [1]). These parameters have to be tuned to the value that gives best classification. The optimization will use 5-fold cross-validation[2] to calculate the classification rates. The optimization technique is “grid search” as recommended[1].

## Creation of required objects

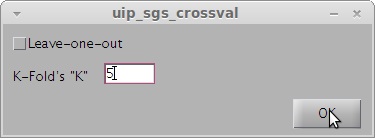
1. Click on “Sub-dataset Generation Specs” in left panel
2. Click on “New…” in middle panel



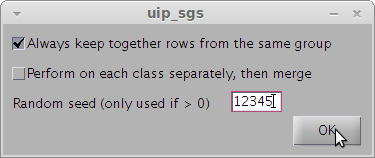
1. Locate and double-click on “K-fold Cross-Validation”



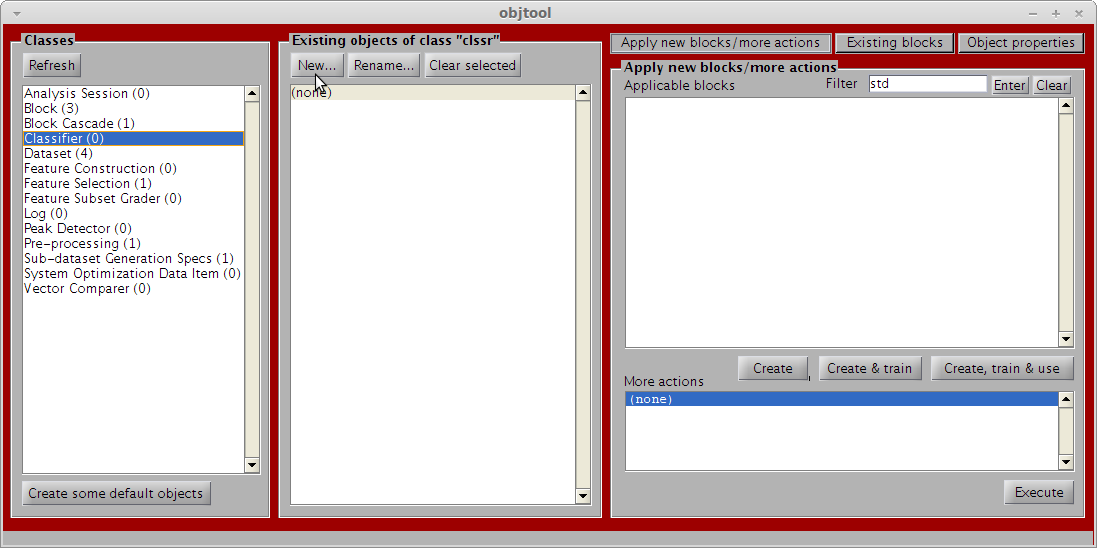
1. Enter “5” in the “K-Fold’s ‘K’” box



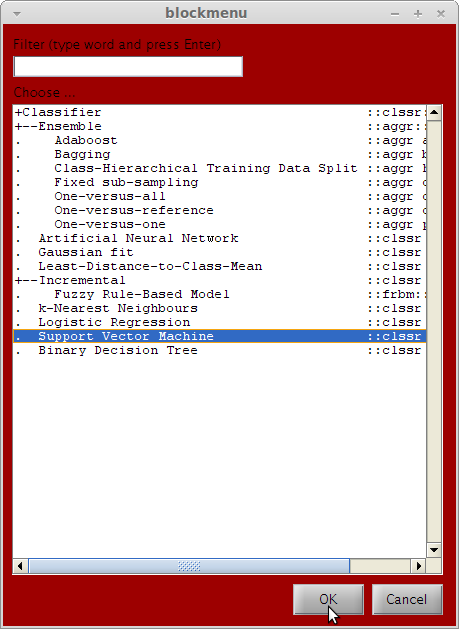
1. Optionally type any number (*e.g.*, 12345) in the “Random seed” box (recommended)
2. Click on “OK”



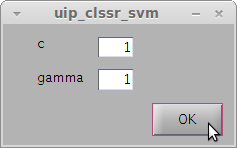
1. Click on “Classifier” in left panel
2. Click on “New…” in middle panel



1. Locate and double-click on “Support Vector Machine”

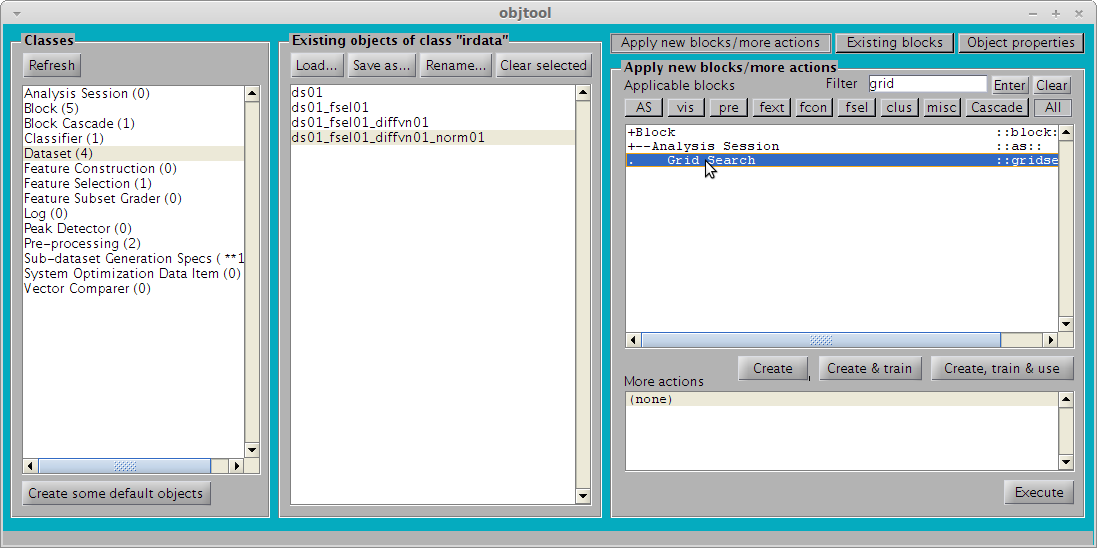


1. Click “OK” (the values in the boxes will not be used anyway)

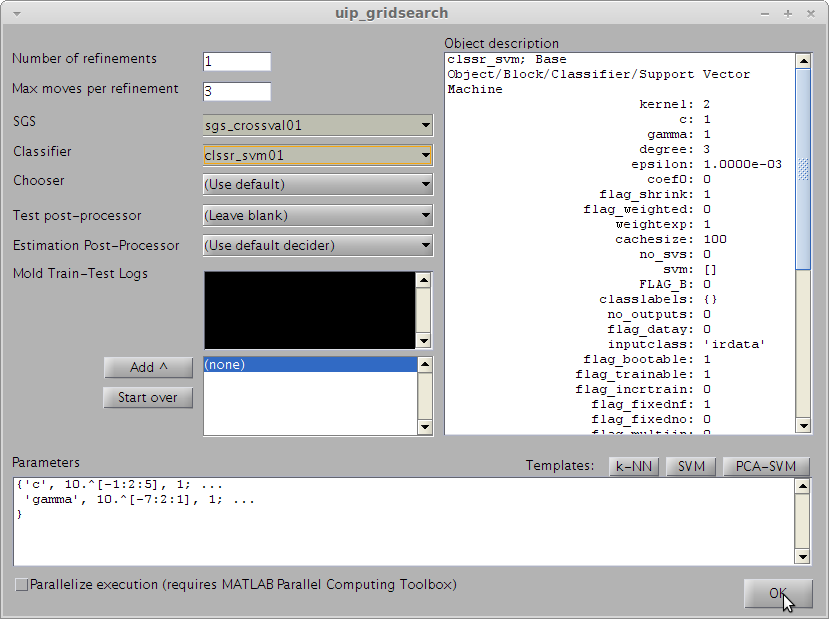


## Grid search

1. Click on “Dataset” in left panel
2. Click on dataset named “ds01\_fsel01\_diffvn01\_norm01” in middle panel
3. Locate and double-click “Grid Search” in right panel



1. In the “SGS” drop-down box, select “sgs\_crossval01”
2. In the “Classifier” drop-down bow, select “clssr\_svm01”
3. You may optionally change the search space of *c* and *gamma* or accept the default values.

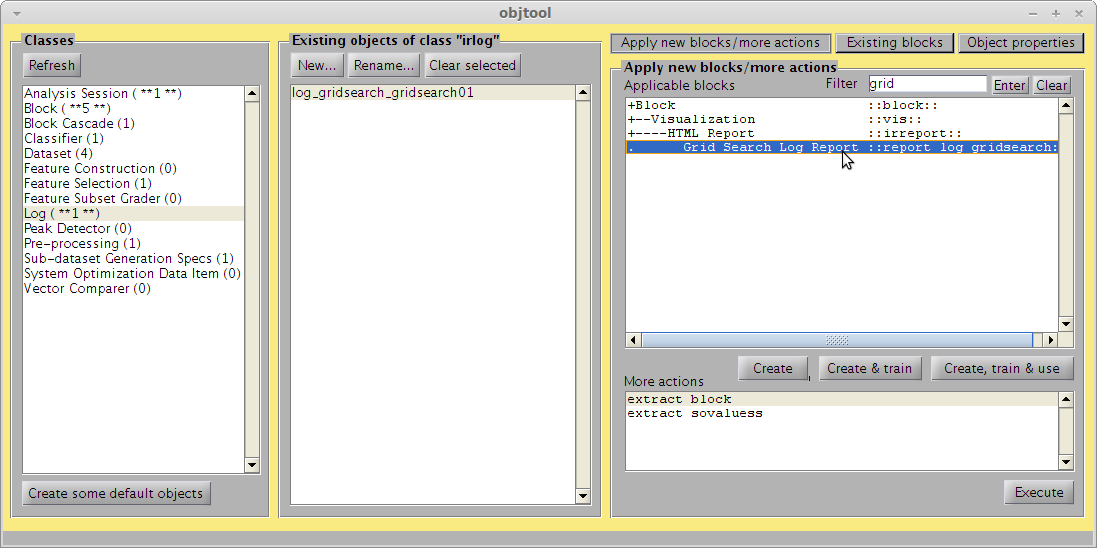


1. Click on “OK”. **Warning:** grid search is potentially time-consuming
2. Watch MATLAB command window for progress indicator

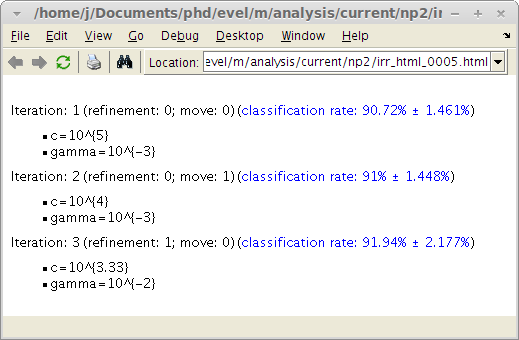
# Visualization of results

Visualization of iterations report

1. Click on “Log” in left panel
2. Select “log\_gridsearch\_gridsearch01” in middle panel
3. Double-click on “Grid Search Log Report” in right panel

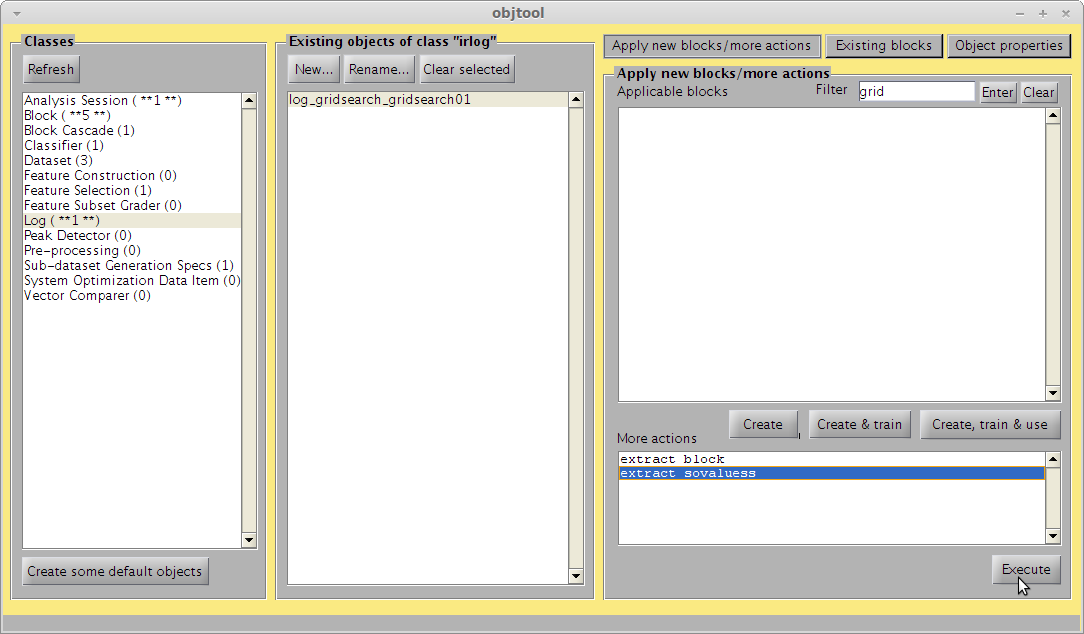


This will show the best classification rate found at each iteration, with respective parameters:

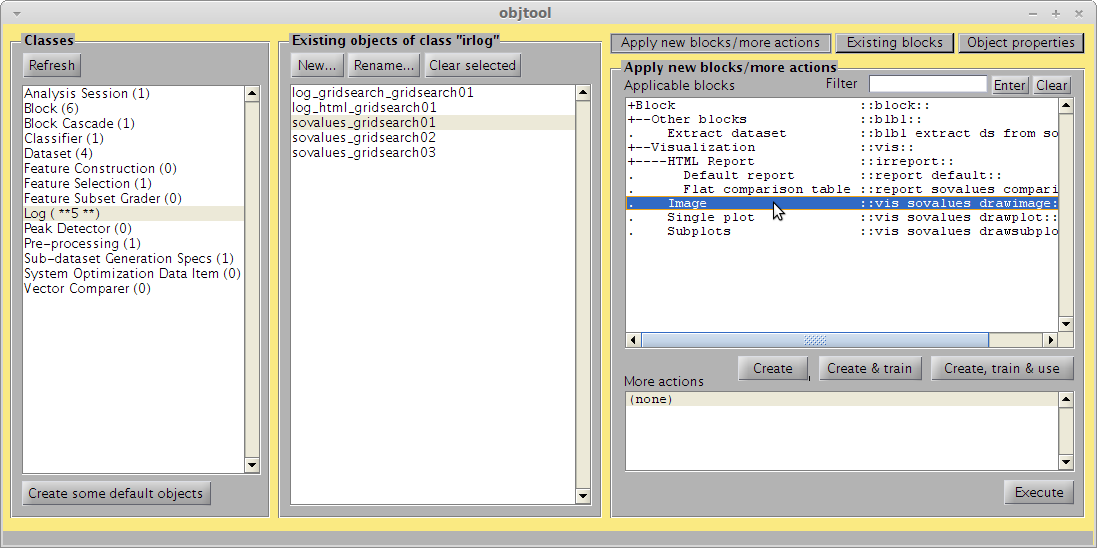


## Visualization of the optimization log

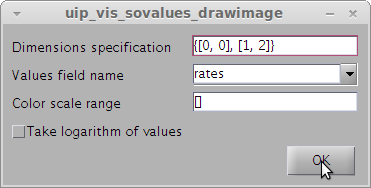
1. Click on “Log” in left panel
2. Select “log\_gridsearch\_gridsearch01” in middle panel
3. Double-click on “extract\_sovalues” in right panel

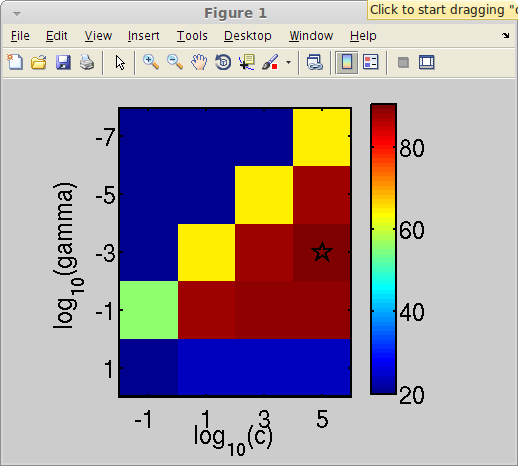


1. Click on “sovalues\_gridsearch01” in the middle panel
2. Locate and double-click on “Image” in the right panel

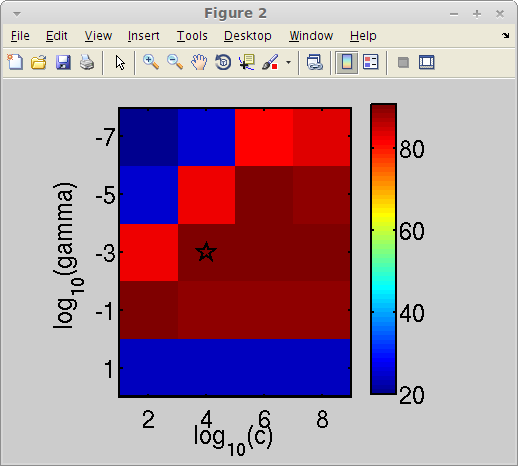
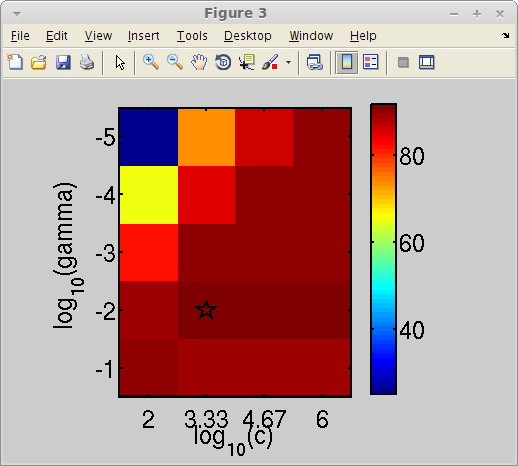


1. In the “Dimensions specification” box, change to “{[0, 0], [1, 2]}”
2. Click on “OK”



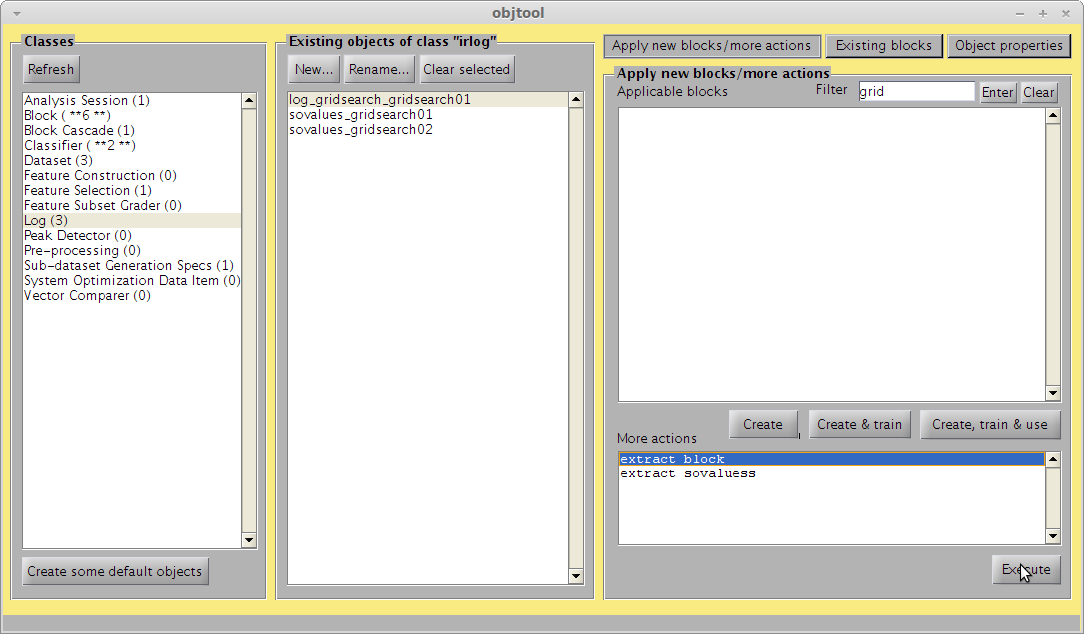


1. Repeat last 4 steps for both “sovalues\_gridsearch02” and “sovalues\_gridsearch02” objects in the middle panel

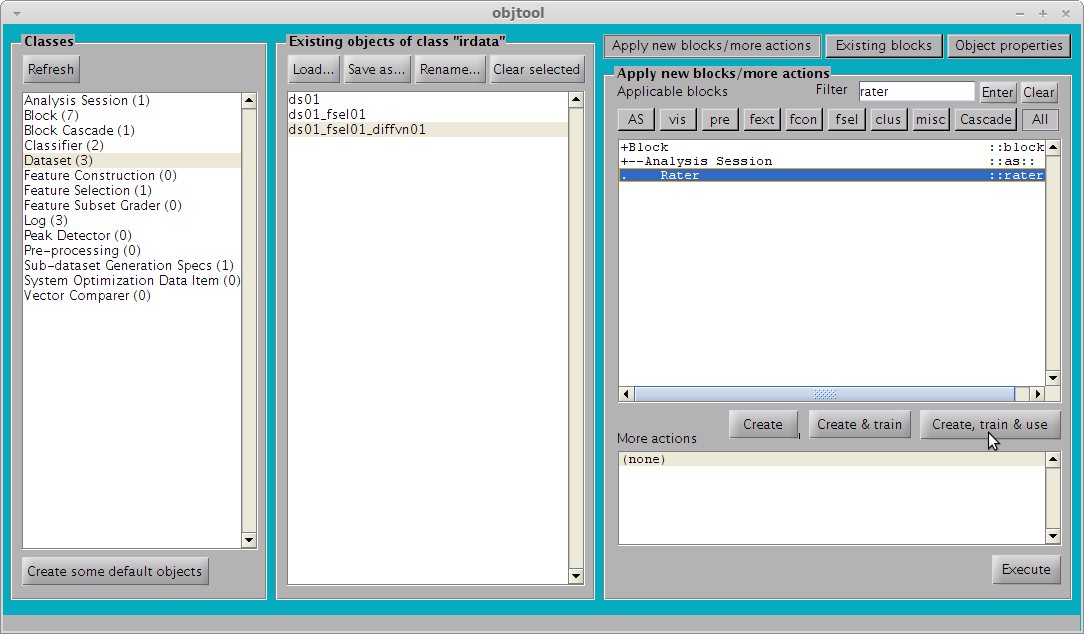
 

## Classification confusion matrix for best parameters

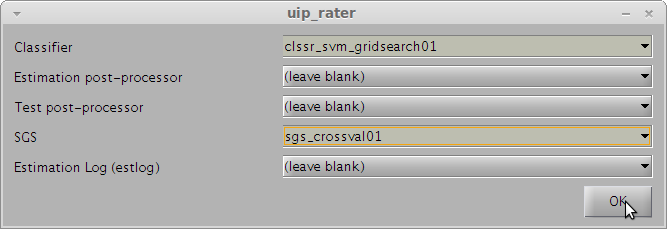
1. Click on “Log” in the left panel
2. Click on “log\_gridsearch\_gridsearch01” in the middle panel
3. Double-click on “extract\_block” in the right panel



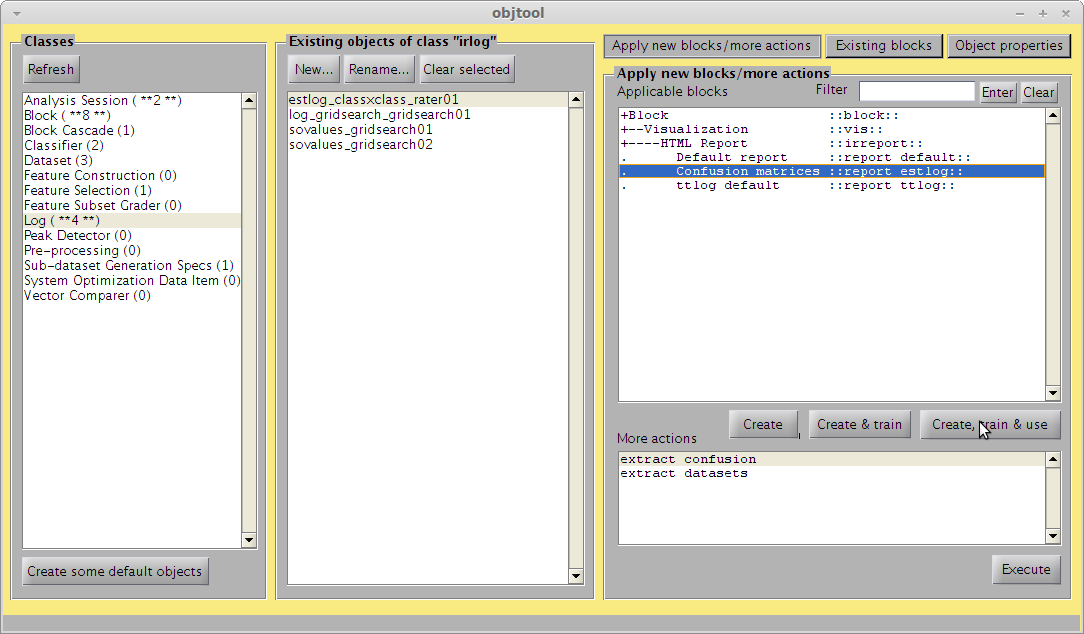
1. Click on “Dataset” in the left panel
2. Click on “ds01\_fsel01\_diffvn01\_norm01” in the middle panel
3. Double-click on “Rater” in the right panel



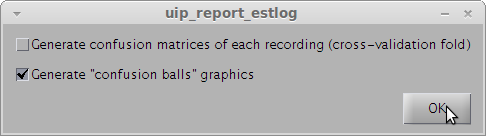
1. In the “Classifier” box, select “clssr\_svm\_gridsearch01” (this is the block that was created from the block extraction action above).
2. In the SGS box, select “sgs\_crossval01”. This will cause the cross-validated estimation to use the same dataset splits as the grid search optimization before.
3. Click on “OK”

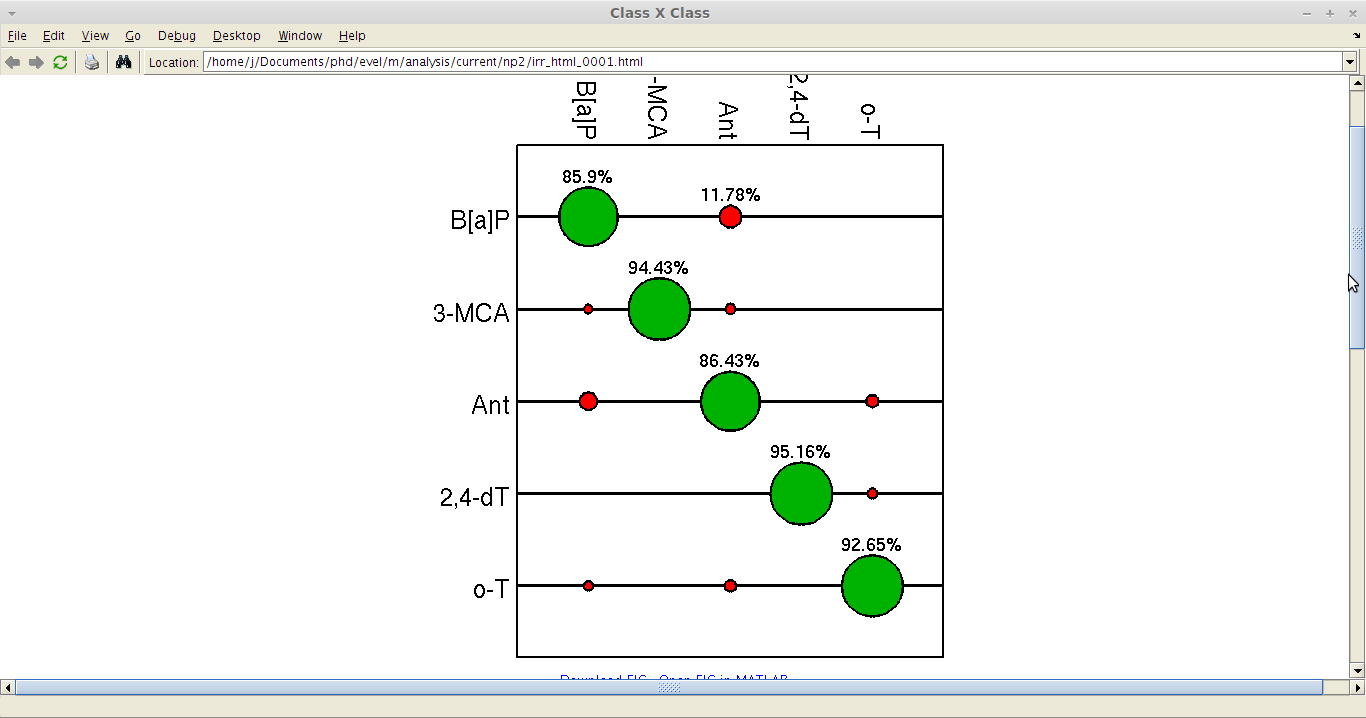


1. Click on “Log” in the left panel
2. Click on “estlog\_classxclass\_rater01” in the middle panel
3. Double-click on “Confusion matrices” in the right panel



1. Click on “OK”





# References

[1] C. Hsu, C. Chang, and C. Lin, “A Practical Guide to Support Vector Classification,” *Bioinformatics*, vol. 1, no. 1, pp. 1–16, 2010.

[2] T. Hastie, J. H. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2007.