JPL/Caltech Virtual Summer School in Big Data Analytics

Classification / Clustering exercises.

Software and Datasets

Download and install Orange (legacy version, try also the new version):
http://orange.biolab.si

Download iris.tab from:
http://www.astro.caltech.edu/~donalek/iris/

More information about Iris:
http://archive.ics.uci.edu/ml/datasets/Iris

Exercise 1: preliminary data analysis

With your language/software of choice:
- count the number of samples per class;
- produce scatterplots encoding the classes by color: what can you derive about linear/non linear separability?
- look for correlations.

Exercise 2: use Orange to build a kMeans

Load iris.tab
Setup a kMeans run with:
- number of clusters: optimized from 2 to 5;
- scoring: between cluster distance;
- distance measure Euclidian

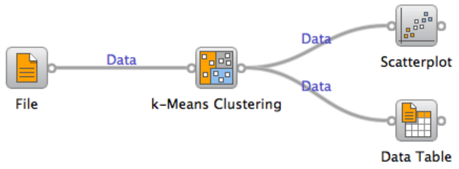
and visualize the results (data table, scatter plots).

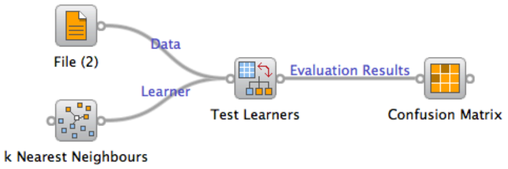
Then change scoring and distance measures and see how they affect the clustering.

Exercise 3: use Orange to build a kNN workflow

Load iris.tab
Run a kNN algorithm and analyze the results; plot a confusion matrix.
Change the cross-validation and sampling methods.

Hint: look at the screenshot!
Do the same exercise using your language/package of choice.





Exercise 4: use your own data for classification/clustering!

During the interactive session we can discuss about your data and look together on how to proceed and what's the best approach/model to use for classification or clustering.

What you need to do:

- prepare your data in a text or CSV format where each column is a parameter, each row a sample; keep class information as the last columns (better if classes are encoded with numbers).
- produce histograms, scatter-plots for parameters/classes;
- are there any missing data?
- etc.
- have plots and statistics online, include also a readme file that explain the dataset.